6th LEARNER CORPUS RESEARCH CONFERENCE

Padua, 22-24 September 2022

# Book of Abstracts

## Table of Contents

### PLENARY TALKS



### FULL PAPERS

## WORK-IN-PROGRESS REPORTS

## POSTERS

## SOFTWARE DEMOS


Sponsor of the Benjamins Poster Prize

## A marriage of two minds? Learner translation corpora in learner corpus research

Silvia Bernardini
Università di Bologna
silvia.bernardini@unibo.it

Learner corpus research (LCR) is understood, quite naturally, as the adoption of corpus linguistics techniques in the study of language learning and acquisition, in other words for describing and modelling non-native or second language varieties through the investigation of learners' production. The field of corpus-based translation studies, which has come to the fore and developed in parallel to LCR, also aims to conceptualize and investigate what is purported to be a separate language variety, namely translated language. The two fields thus have several points of contact, that have recently led to a partial alignment of interests and priorities. As a result, several learner translation corpora (LTCs) have seen the light, to which well-established practices from LCR (such as error annotation) are also applied.

In this talk, I will first of all discuss the ways in which LTCs can be of interest to the field of LCR and language pedagogy at large. First, pedagogic translation has consistently been practiced in the language classroom, and has even been rehabilitated in recent years. Second, translation data provide direct first language equivalents for produced second/target language segments, which may complement datasets resulting from freer production tasks. Third, and more importantly, bringing the two research frameworks together allows one to pursue the fascinating, and very ambitious, goal of understanding similarities and differences between L2 and translated production seen as instances of constrained communication in language contact situations.

Somewhat provocatively, I will also point out several important methodological and theoretical issues. Indeed, notwithstanding the potential advantages of this alignment, the nature of the data is such that one may wonder whether it is legitimate to include current LTCs fully among learner corpora, or even to consider them corpora at all. To illustrate my point, I will refer to an exploratory attempt at combining translation and essay writing data in the exploration of English topic-neutral, high-frequency collocations. The datasets used are not closely comparable in terms of topic and register, since they contain texts assembled from previous coursework and examinations: a suboptimal, yet rather common condition applying to non-experimental settings in which translated and non-native language varieties are compared.

Rather than provide answers to such complex questions, I hope to stimulate discussion on how we can make sense of learner corpora and LTCs, and of complex datasets representing multiple instances of learner production in general, while remaining true to the methodological and theoretical assumptions informing corpus linguistics.

## Explorations of variability: Evidence from L1 and L2 corpora of German

Anke Lüdeling
Humboldt-Universität zu Berlin
anke.luedeling@hu-berlin.de

In recent years, corpus linguistics has learned a lot about variability between corpora and texts. We have seen research on external factors such as task effects (Crowther et al. 2015, Gablasova et al. 2017, Schnur & Rubio 2021, Weiss 2017) and setting, as well as speaker factors such as age, socio-economic properties, aptitude, or motivation (Birdsong 2018, Dörnyei & Ryan 2015, Granena 2013, Larsen-Freeman 2018, among many others).

Building on this research, this paper will dig even further into different aspects of variability and discuss a number of theoretical and methodological implications. Using two German corpora that are maximally controlled for external factors with matching L1 and L2 subcorpora (Kobalt, Zinsmeister *et al.* 2012 and Falko, Reznicek et al. 2012) and deeply annotated for syntactic categories as well as morphological subclasses (see Lukassek et al. 2022, Shadrova 2021), this paper will investigate intra- and inter-speaker variation with respect to morphology and syntax. The results of recent research I have conducted with colleagues indicate a surprisingly high degree of variance between L1 speakers in the distribution of word formation subclasses (cf. Shadrova et al. 2021), which challenges the construct of native speaker homogeneity beyond stable, stratified, and situational variation and raises methodological questions for comparative L1-L2 studies as well as to the role of frequency in the entrenchment.

I will further present evidence for intra-individual differences, looking into procedural factors, such as priming and self-priming and within-text register fluctuation, highlighting the necessity of accounting for text dynamics and aspects of (the acquisition of) register knowledge. At the same time, the distribution of syntactic categories such as dependencies and parts of speech is much less variable across speakers and even between corpora, suggesting categorical differences between syntax and the lexicon in production and challenging the notion of "constructions all the way down" (Goldberg 2006, 18).

Based on these explorations of variability between and within speakers, as well as between linguistic layers, I will discuss the potential of small, deeply annotated, and well-understood corpora, which are ideally suited to accommodate the needs of careful linguistic analysis in a complex space of interactions and path-dependencies.

**References**

Birdsong, D. (2018). Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology* 9(81), https://doi.org/10.3389/fpsyg.2018.00081.

Crowther, D., Trofimovich, P., Isaacs, T. & Saito, K. (2015). Does a Speaking Task Affect Second Language Comprehensibility? *The Modern Language Journal* 99(1), 80-95.

Dörnyei, Z. & Ryan, S. (2015). *The Psychology of the Language Learner Revisited* Routledge.

Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. (2017). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics* 38(5), 613-637.

Goldberg, Adele E. (2016). *Constructions at Work: The Nature of Generalizations in Language*. Oxford University Press.

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning* 63(4), 665-703.

Larsen-Freeman, D. (2018). Looking ahead: Future directions in, and future research into, second language acquisition. *Foreign Language Annals* 51(1), 55-72.

Lukassek, J., Akbari, R. & Lüdeling, A. (2022). Guidelines for the morphological annotation of nouns in Falko. To appear in *REALIS White Paper Series*, https://sfb1412.hu-berlin.de/.

Reznicek, M., Walter, M. Kari Schmidt, Lüdeling, A., Hirschmann, H., Krummes, C. & Andreas, T. (2012). Das Falko-Handbuch: Korpusaufbau und Annotationen. Unpublished, https://hu-berlin.de/falko.

Schnur, E. & Rubio, E. (2021). Lexical complexity, writing proficiency, and task effects in Spanish Dual Language Immersion. *Language Learning and Technology* 25(1), 53-72. https://hdl.handle.net/10125/73425.

Shadrova, A. (2021). *Kobalt: Extension Corpus and Annotation Guidelines for Verb Classification and Dependency Adjustments.* Zenodo. https://doi.org/10.5281/zenodo.5730224.

Shadrova, A., Linscheid, P., Lukassek, J., Lüdeling, A. & Schneider, S. (2021). A challenge for contrastive L1/L2 corpus studies: Large inter-and intra-individual variation across morphological, but not global syntactic categories in task-based corpus data of a homogeneous L1 German group. *Frontiers in Pschology* 12, https://doi.org/10.3389/fpsyg.2021.716485.

Weiss, Z. (2017). *Using measures of linguistic complexity to assess German L2 proficiency in learner corpora under consideration of task-effects.* Unpublished MA thesis, http://www.sfs.uni-tuebingen.de/~zweiss/.

Zinsmeister, H., Reznicek, M., Ricard-Brede, J., Rosén, C., & Skiba, D. (2012). Das Wissenschaftliche Netzwerk „Kobalt-DaF“ Korpusbasierte Analyse von Lernertexten für Deutsch als Fremdsprache. *Zeitschrift für germanistische Linguistik* 40(3), 457-458.

## Learner corpus research: Some problems, some questions, and some possible answers

Hilary Nesi
Coventry University
hilary.nesi@coventry.ac.uk

This talk will explore issues associated with the notion of the learner corpus, illustrated with references to my own experience as a language teacher, language learner, researcher and corpus designer.

First of all, it will ask how we should define the 'learners' who produce learner corpus content. There seem to be three basic ways of deciding this - by self-identification, by mother-tongue status, or according to their presence in a language learning class. The first two definitions are a bit problematic, as in some respects we can all self-identify as learners, and most people in the world have mixed proficiencies in more than one language. Many people report that they are happier using one language at home and in their own cultural contexts, and another language when communicating their academic or professional expertise, especially if they have acquired their expertise in the medium of this other language. The 'native speaker' designation is increasingly rejected by educationalists and journal editors because it implies superior communication skills in the mother tongue, something we know is by no means guaranteed. On the other hand, if learners are only learners when performing in the language learning class, the only texts that can be included in a learner corpus are those produced for the purposes of language learning or assessment. There is a danger that such texts will be coloured by the demands of the language learning syllabus, with certain linguistic features included only for the purposes of display.

In light of this, we also have to decide on appropriate methods of learner corpus analysis. Most approaches, beyond identifying typical structural errors, require some comparison with texts produced by 'non-learners' – probably people who use the language as their mother tongue, people considered expert speakers or writers, or both. Corpus compilers know that it can be rather difficult to identify who is and who is not a native speaker, especially in studies involving large numbers of texts, perhaps produced by multiple authors. Moreover, any comparison between texts produced in different situational contexts automatically introduce extra variables that have nothing to do with language learning status: whether we compare texts produced by experts with those produced by novices; local texts with those produced for international audiences; or texts produced in the language classroom with those produced for any genuine academic, professional or social purpose.

The best thing seems to be to address these problems full on, acknowledging the inherent difficulties in learner corpus research and making allowances for them when designing corpora and drawing our conclusions. I hope the talk will be thought-provoking, and give rise to some lively discussion.

## Splitting and joining sentences in Italian-Russian inverse translation

Ksenia Balakina
University of Bologna
ksenia.balakina2@unibo.it

This study aims to investigate the shifts in sentence boundaries that occur in Italian-Russian learner translations. In particular, the study deals with the process of inverse translation that remains underrepresented in translation research (see, e.g., Ferreira, Schwieter 2017) and examines data from a parallel learner corpus (according to M. A. Lefer (2021), a scarcely represented corpus variety).

The focus of this study is what was referred to by Newmark as a "natural unit of translation" – the sentence (Newmark 1988, 65). As he pointed out, sentence boundaries are not normally rearranged unless there are good reasons to do that, which is not always the case when examining learner translations.

Within the framework of translation studies, the processes of sentence splitting and joining can be viewed as a manifestation of translation universals, such as simplification and explication (Baker, 1996: 179-183). On the syntactic level, the tendency to simplification thus leads to breaking up long sentences in translation, whereas the explication tendency is expressed by greater explicitness of the syntactic relationship between joined sentences.

Since it is the inverse translation that is under examination in this study, it is important to mention as well the second-language-acquisition perspective, from which learners that haven't mastered a language at a sufficient level tend to simplify their production in the target language (Ellis 2008: 80–82).

Few publications have so far reported on the dynamics of splitting and joining sentences in translation: most of them investigated the phenomenon in professional published translations (Fabricius-Hansen 1999, Bisiada 2013, Nádvorníková 2017, Frankenberg-Garcia 2019), whereas the publications addressing learner translations (eg. Kunilovskaja, Morgoun 2013) analyze direct translations (into L1).

The purpose of this study is thus to answer the following research questions:

- To which extent are sentence boundaries changed in learner inverse translations?
- Do learners split sentences more often than join them when translating texts into a non-native language?
- Are there differences in sentence splitting/joining when comparing translations performed by students with different levels of target language proficiency and translation experience?
- What are the syntactic structures that most often "trigger" sentence splitting?

The study is based on a parallel learner corpus of inverse translations from Italian into Russian. The corpus represents a collection of translations produced by Italian-speaking undergraduate students attending translation courses and studying Russian as a foreign language at university. The learner translations into Russian count almost 240 thousand words and include two balanced subsets of translations performed by two different groups of learner translators:

- Undergraduate second-year students (attending intermediate Russian language courses) with no experience in translation into Russian;
- Third-year students (attending advanced Russian language courses) with one year of experience in Italian-Russian translation.

The automatic processing (including the sentence segmentation) was performed by the Sketch Engine online text analysis tool. All the instances of sentence splitting and joining were extracted from the corpus to be subsequently manually annotated by categories describing the syntactic structures that were deleted or added in translation when splitting and joining sentences respectively.

The quantitative analysis produced the following general results:

- The overall level of sentence boundaries preservation is very high with the total amount of split/joined translation examples counting less than 5% of translated sentences;
- Sentence splitting is more frequent in learner translations than sentence joining which is in line with the conclusions made in the studies based on the professional translation (e.g., Bisiada 2013, Frankenberg-Garcia 2019);
- Third-year students tend to split sentences more frequently than second-year students.

A more detailed analysis was performed using the sample of split sentences. The annotation allowed us to identify the syntactic relationships that are most often split and rearranged as border structures of two sentences

in the target text. Most examples of splitting relate to coordinate (source) clauses; other frequently split structures include subordinate clauses, verbal constructions, appositives, and lists. The analysis was also useful to detect the effects produced by sentence splitting on the quality of learner translations: these effects range from improved readability of long, complicated source sentences to thematic progression issues and cohesion errors.

When interpreting and discussing the results in the final part of the study various aspects are taken into consideration, such as target language proficiency level and translation experience, on the one hand, and the (increasing) levels of complexity of the source texts, on the other; difficulties related to the target language acquisition as well as the approaches to inverse translation teaching.

**References**


Baker, M. (1996). Corpus-based Translation Studies: the challenges that lie ahead. In H. Sommers (ed.) *Terminology, LSP and Translation Studies in Language Engineering in Honour of J.C. Sager*. Amsterdam: John Benjamins.

Bisiada, M. (2013). *From Hypotaxis to Parataxis: An Investigation of English–German Syntactic Convergence in Translation.* PhD thesis. University of Manchester.

Ellis, R. (2008). *The Study of Second Language Acquisition.* Second Edition. Oxford: Oxford University Press.

Fabricius-Hansen, C. (1999). Information packaging and translation: aspects of translational sentence splitting (German–English/Norwegian). In M. Doherty (ed.) *Sprachspezifissche Aspekte der Informationsverteilung [Language-specific aspects of information distribution]*. Berlin: Akademie Verlag.

Frankenberg-Garcia, A. (2019). A corpus study of splitting and joining sentences in translation. *Corpora*, *14*(1), 1-30.

Kunilovskaya, M. A., & Morgoun, N. L. (2013). Gains and pitfalls of sentence-splitting in translation. *Vestnik Permskogo nacional'nogo issledovatel'skogo politehničeskogo universiteta. Problemy jazykoznanija i pedagogiki*, (8).

Lefer, M.-A. (2021). *Breaking new ground in contrastive and translation studies: Learner translation corpora to the fore.* Using Corpora in Contrastive and Translation Studies (Bertinoro, Italy, 09/09/2021 - 11/09/2021).

Nádvorníková, O. (2017). Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus. *Language Use and Linguistic Structure*, 445-461.

Newmark P. (1988). *A Textbook of Translation.* Hertfordshire: Prentice Hall.

# Linguistic style in a second language: Exploring cross-task individual differences in complexity in a large-scale corpus

Elizabeth Bear, Xiaobin Chen, Detmar Meurers
University of Tübingen
{elizabeth.bear, xiaobin.chen, detmar.meurers}@uni-tuebingen.de

The analysis of text to reveal information about the author, such as gender (Koppel et al. 2002), personality (Pennebaker & King 1999), or the identity of the author itself (Mosteller & Wallace 1964), has long been of interest in several fields of research. Based on Allport (1961)'s definition of stylistic behavior as "one's manner of performing adaptive acts" (461), Pennebaker and King (1999) were the first to explore linguistic style as a reliable individual difference across multiple writing samples that can be linked to personality traits.

Bringing this perspective to Second Language Acquisition research, in this paper we investigate whether authors writing in a second language (L2), in this case English, possess a linguistic style consistent across time and writing tasks. We pursue the hypothesis that research on stylistic features (Koppel et al. 2009: 12) can be empirically enriched by considering linguistic complexity features as discussed in research on Complexity, Accuracy, and Fluency (CAF; Skehan 1989; Housen & Kuiken 2009). Different from CAF research studying language development, we investigate whether a writer makes characteristic individual choices in language complexification.

EFCAMDAT (Geertzen et al. 2013) was selected as the corpus due to the availability of multiple writings for each learner and CEFR-aligned labels, which permitted controlling for proficiency differences to the extent possible. EFCAMDAT also contains detailed task information, which has been shown to influence linguistic complexity (Alexopoulou et al. 2017; Michel et al. 2019) and poses "a particularly severe threat to validity in longitudinal designs" (Vyatkina 2012: 595) and, accordingly, to an analysis of a linguistic style across time and tasks.

For our analyses, we selected all learners at the three highest proficiency levels (B2-C2) who had completed the final three writing tasks of the given level. Within each proficiency level, we computed how well a measure for a writer correlated across these tasks for two sets of measures. The first set contained eight lexical and syntactic complexity measures drawn from Pennebaker and King (1999). Responding to calls for the inclusion of more fine-grained measures within L2 linguistic complexity research (e.g., Lu 2011; Vyatkina 2012), our second feature set added six syntactic complexity measures at the clausal level, which was not considered by Pennebaker and King.

In the first set of measures, mean sentence length (MLS) displayed the highest correlations (*rs* ranging from 0.32 to 0.50); in other words, writers with a high MLS in one writing task were more likely to have a high MLS in another writing task. In addition, all measures had at least one significant correlation between two given writing tasks. Among the more fine-grained measures in the second set, some significant correlations were found but less systematically. This result is consistent with research on L2 writing quality (Yang et al. 2015) that found local clausal-level measures to be more impacted by the topic of a task than more global measures such as MLS.

To validate these results, we performed two additional analyses. The first addressed the potential influence of the automatic processing of learner language, which may contain language errors. We took advantage of the availability of annotated and corrected versions of the texts in EFCAMDAT to identify whether the errors impacted the computation of the complexity measures analyzed. For both sets of measures, we found high correlations between measures calculated on the original texts and their corrected versions. The second validation analysis incorporated a wider range of predictors available from EFCAMDAT, such as task and nationality, into mixed-effects models to provide a more complete picture of the individual variation in select linguistic style measures from the two sets. Substantial individual variation remained after task effects and other predictors were taken into account and revealed qualitative insights into the learners who deviated the most from a measure's population mean.

Taken together, the results provide some preliminary evidence for a cross-task linguistic style at both the lexical and syntactic level, with some measures, particularly MLS, emerging as more consistent than others. The findings were notable given the different topics and separation in time between writing tasks. This first step opens possibilities for further research, such as if the linguistic style measures can be linked to demographic information or personality traits of the author.

**References**


Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1), 180-208. https://doi.org/10.1111/lang.12232

Allport, G. W. (1961). *Pattern and growth in personality*. New York: Holt, Reinhart & Winston.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*. Somerville, MA: Cascadilla Proceedings Project, 240-254.

Housen, A. & Kuiken, F. (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics*, 30(4): 461-473.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 401-412.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9-26.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1): 36-62.

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: Evidence from a large learner corpus of A1 to C2 writings. *Instructed Second Language Acquisition*, 3(2): 124-152. https://doi.org/10.1558/isla.38248

Mosteller, F. & Wallace, D. (1964). *Inference and disputed authorship: The Federalist*. Reading, MA: Addison Wesley.

Pennebaker, J. W. & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6): 1296-1312.

Skehan, P. (1989). *Individual differences in second language learning.* London: Arnold.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*, 96(4): 576-598.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28, 53-67.

# Dimensions of grammatical complexity in L1/L2 writing: A comparative analysis of theory-based models

Doug Biber[1], Tove Larsson[2], Gregory Hancock[3], Bethany Gray[4], Randi Reppen[5]
Northern Arizona University[1,2,5], University of Maryland[3], Iowa State University[4],
douglas.biber@nau.edu[1]

Biber et al. (2011) hypothesizes that advanced L2 English writers progress through five developmental stages in their use of complexity features, with each stage being composed of a different set of lexico-grammatical features. This hypothesis was based on previous corpus-based studies, which demonstrate the importance of both structural type and syntactic function for distinguishing between the complexities of spoken and written registers (see Biber and Gray 2016; Biber et al 2021, 2022): Spoken discourse relies on embedded finite dependent clauses, functioning syntactically as clause-level constituents; informational written discourse relies on embedded phrases functioning syntactically as phrase-level modifiers. Other features are intermediate along these two parameters.

By considering both parameters, linguistic complexity features were grouped into the five hypothesized developmental stages. Structurally, the stages progress generally from finite dependent clauses → non-finite dependent clauses → embedded phrases, while syntactically, the stages progress generally from features functioning as clause-level constituents to features functioning as phrase-level modifiers. Since 2011, numerous empirical studies have provided strong descriptive evidence that L2 writing development progresses generally according to these hypothesized stages (see, e.g., Taguchi et al. 2013; Parkinson and Musgrave 2014; Staples et al. 2016; Ansarifar et al. 2018; Staples et al. 2018; Lan and Sun 2019; Gray et al. 2019; Atak and Saricaoglu 2020; Biber, Reppen, Staples, and Egbert, 2020).

However, no study to date has empirically validated the groupings of complexity features associated with each developmental stage, or compared the descriptive adequacy of those groupings to other theory-based models grouping complexity features in different ways. Further, no study to date has empirically tested whether the ways in which complexity features pattern together in L1-English written discourse is the same as the ways in which those features pattern together in L2-English writing. These are the two main objectives of the present study.

We employ Confirmatory Factor Analysis (CFA) for these research goals, a statistical technique that tests the extent to which the observed data actually fits hypothesized models based on theory and/or previous empirical research. In particular, we compare the goodness-of-fit for six different models:

1. a model with a single dimension, which would support the theoretical claim that all complexity features pattern in the same way.
2. a model with three dimensions that represent the three major structural types of complexity features: finite dependent clauses, nonfinite dependent clauses, and dependent phrases. This model would support the theoretical claim that embedded clauses represent a different kind of complexity from embedded phrases.
3. a model with two dimensions that represent the two major syntactic functions of complexity features: clause elements (functioning as objects, complements, or adverbials), and phrase-level modifiers. This model would support the theoretical importance of syntactic function.
4. a model with six dimensions, representing the combinations of structural type and syntactic function. This model would support the claim of Biber et al (2021, 2022) that both grammatical structure and syntactic function are crucially important distinctions for understanding grammatical complexity.
5. a model with five dimensions, representing the major developmental stages hypothesized in Biber et al. 2011. This model would directly test the adequacy of the groupings of complexity features proposed in Biber et al. (2011).
6. a model with six dimensions, informed by previous corpus-based research of register variation, taking into account structural type, syntactic function, lexico-grammatical patterns, and spoken/written register differences.

The descriptive/statistical adequacy of these models is compared in a large multi-register corpus of L1-English and L2-English writing, with samples evenly matched for the registers and proficiency levels included for each group (with registers ranging from personal narratives to research writing, and proficiency levels ranging from first-year university student to professional academics).

In this talk we will, in a non-technical manner, focus on describing the groupings of complexity features that best account for the patterns of use in L1/L2-English writing. Separate CFAs are carried out for L1-English and L2-English groups, identifying the theoretical model that best accounts for linguistic complexity patterns in each group, and ultimately addressing the question of whether L1-English written discourse is governed by the same parameters of complexity as L2-English written discourse.

**References**


Ansarifar, A., Shahriari, H., & Pishghadam, R. (2018). Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes, 31*, 58-71.

Atak, N., & Saricaoglu, A. (2021). Syntactic complexity in L2 learners' argumentative writing: Developmental stages and the within-genre topic effect. Assessing Writing, 47, 100506.

Biber, D., B. Gray, S. Staples, and J. Egbert. (2022). *The Register-Functional approach to grammatical complexity: Theoretical foundation, descriptive research findings, applications.* Routledge.

Biber, D., B. Gray, S. Staples, and J. Egbert. (2021). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes.* 46.

Biber, D., Reppen, R., Staples, S. & Egbert, J. (2020). Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. *International Journal of Learner Corpus Research,* 6(1), 38-71.

Biber, D., and B. Gray. (2016). *Grammatical complexity in academic English: Linguistic change in writing.* Cambridge: Cambridge University Press.

Biber, D., Bethany Gray, Kornwipa Poonpon. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45.5-35.

Gray, B., Geluso, J. & Nguyen, P. (2019). *The longitudinal development of grammatical complexity at the phrasal and clausal levels in spoken and written responses to the TOEFL iBT test*. TOEFL iBT Research Report No. RR-19-45. Princeton, NJ: Educational Testing Service.

Lan, G., & Sun, Y. (2019). A corpus-based investigation of noun phrase complexity in the L2 writings of a first-year composition course. *Journal of English for Academic Purposes,* 38, 14-24.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes,* 14, 48-59.

Staples, S., J. Egbert, D. Biber, and B. Gray. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, 33.149-183.

Staples, S., Biber, D., & Reppen, R. (2018). Using corpus-based register analysis to explore authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *The Modern Language Journal*, 102(2), 310-332.

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, 47, 420-430.

# Development of explicit causal connectives in Italian L1 and L2 student writing: A comparison of argumentative texts from lower and upper secondary school

Arianna Bienati, Jennifer Carmen Frey
Institute for Applied Linguistics, Eurac Research
arianna.bienati@eurac.edu, jennifercarmen.frey@eurac.edu

When writing, discourse or coherence relations (Mann & Thompson 1988; Kehler 2002; Asher et al. 2003; Miltsakaki et al. 2004) are a paramount strategy to logically connect semantically related stretches of text. Formally, languages provide extensive sets of connectives that encode these semantic relations explicitly (Pander Maat & Sanders 2006). Although the use of such explicit cohesive devices is not necessarily correlated with coherence or text quality judgments (Crossley et al. 2016), its acquisition is an important steppingstone in text competence development. Thus, an in-depth analysis of the types and variety of connectives used at different stages of a writer's school education could provide important empirical data for training of textuality features and writing assessment in L2 and L1 teaching practice of a particular language.

In our contribution, we analyzed the quantity and repertoire of explicit connectives found in argumentative texts of L1 and L2 speakers of Italian in the 3rd year of lower secondary school, and after four years of training, i.e., in the 4th year of upper secondary school. In our analysis, we focus on explicit causal connectives as one important means for constructing coherence in argumentative texts, in that they explicitly point out supporting reasons and anticipated consequences, to convince an audience of a statement.

Our research questions are:

- Are there any common trends in the use of explicit connectives employed by students through time, regarding quantity and repertoire of uses?
- Are there any significant differences in the use of explicit connectives by L1 and L2 speakers at the same developmental stage?

To answer these questions, we automatically annotated the explicit causal connectives in a sample of 200 texts, evenly distributed between the four conditions, namely first/second language and lower/upper secondary school. All texts were gathered in the multilingual province of Bolzano/Bozen in Italy and originated from three different learner corpora. Argumentative texts of lower secondary school writers were randomly sampled from the L2 and L1 writers in the Italian sub-corpus of LEONIDE (Glaznieks et al. 2022). The texts of upper secondary school writers were drawn as a random sample from the Italian Kolipsi-2 corpus (L2 data, Glaznieks et al. 2021) and from data collected in the ITACA project (L1 data, https://itaca.eurac.edu/). The automatic annotation follows a dictionary-based approach aided by the Lexicon for Italian COnnectives (LICO) (Feltracco et al. 2016), a repository of Italian connectives aligned with the PDTB 3.0 (Webber et al. 2019). To analyze quantity, we observed both the number of causal connectives per text (normalized per 100 words to account for text length differences) and the ratio of causal connectives of all connectives. Furthermore, we investigated the students' repertoire of causal connectives qualitatively and quantitatively, extracting frequencies from a reference corpus (CORIS, Rossini Favaretti et al. 2002) to understand which kind of connectives (if low or high frequency) were present in the four groups. We calculated both the mean and the standard deviation of the frequencies of connectives used in each group, to measure differences in the repertoires.

Our analysis, aided by linear regression models, shows that the number of causal connectives decreases significantly in the upper grades, independently of L1/L2 variable. However, the category is not internally homogeneous: causal connectives of the result type (e.g., *quindi*, *di conseguenza*) display a remarkable relative growth in upper secondary school, suggesting that result relations are more complex and therefore learned later on. Regarding the kind of connective used, older students of both groups use significantly less common connectives than younger students. Changes in the variety over time exist only in the L1 group, in which new, rarer connectives (e.g., *per via*, *siccome*, *cosicché*), may emerge aside from the high-frequency ones typical of lower grades (e.g., *per*, *perché*, *così*, *quindi*). L2 students, instead, seem to use a narrower range of connectives with similar frequency (higher for lower grades and lower for upper grades). In general, the change in the use of causal connectives over time was similar for both L1 and L2 students, with the only significant difference between L1 and L2 students visible only in the variety and average frequency of connectives used in the upper grades. Results suggest that for both L1 and L2 writers, quantity and variety of connectives employed are in a tradeoff: while through time students may learn other strategies to express coherence relations – determining the decrease

in the quantity of connectives –, they learn to use also rarer connectives, supposedly the ones present in formal, academic language.

**References**


Asher, N. M., & Lascarides A. (2003). *Logics of Conversation*. Cambridge: CUP.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1-16.

Feltracco, A., Jezek, E., Magnini, B., & Stede, M. (2016). LICO: A Lexicon of Italian Connectives. In A. Corazza, S. Montemagni, & G. Semeraro (Eds.). *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016*. Torino: Accademia University Press, 141-145.

Ferrari, A. (2014). *Linguistica del testo. Principi, fenomeni, strutture*. Roma: Carocci.

Glaznieks, A. Frey, J.-C., Nicolas, L., Abel, A. & Vettori, C. (2021). Kolipsi-2 Corpus v1.0, Eurac Research CLARIN Centre, http://hdl.handle.net/20.500.12124/30

Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, 8(1), 97-120.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. Stanford: CSLI Publications.

Mann, W., & Thompson, S. (1988). Rethorical Structure Theory: Toward a functional theory of text organization. *Text*, 8, 243–281.

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2004). The Penn Discourse Treebank. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.). *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon: European Language Resources Association, 2237-2240.

Pander Maat, H., & Sanders, T. (2006). Connectives in Text. In K. Brown (ed.). *Encyclopedia of Language & Linguistics*. Amsterdam: Elsevier, 33-41.

Rossini Favaretti, R., Tamburini, F., & De Santis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In A. Wilson, P. Rayson, & T. McEnery (Eds.). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. München: Lincom-Europa, 27-38.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). The Penn Discourse Treebank 3.0 Annotation Manual.

# A corpus-based approach in foreign-language teacher education: A case study on politeness in instant messages in Italian L2 by Germanophone learners

Nicola Brocca, Maria K. Rudigier, Valentin A. Spielthenner
University of Innsbruck
nicola.brocca@uibk.ac.at, {maria.rudigier, valentin.spielthenner}@student.uibk.ac.at

Many recent research projects (Artoni & al. 2020; Nuzzo & Cortés Velásquez, 2020) have underlined the usefulness of creating and analyzing corpora for teaching pragmatics, which can be explained by reference to tendential values or more or less appropriate choices in a given context. That is even more true for interactions via digital media, such as email and instant-messaging services, which have little place in manuals or L2 courses and for which learners have few reference models (Trubnikova & Garofolin, 2020).

The LADDER corpus has been created to fill these needs (Brocca, 2021): The corpus data consist of emails, written instant messages (IM), and transcribed vocal messages (VM). Data were collected from April 2020 to December 2021 by means of a discourse completion test (DCT). The informants are (i) Tyrolean Germanophone learners of Italian between A2-C1 level according to the CEFR and (ii) native speakers of Italian based in Rome. All informants are students aged 18 to 35. The DCTs have been conducted with online questionnaires. Along with the texts, metadata were also registered with the help of an online questionnaire providing sociolinguistic information about the informant (age, self-assessed language level, place of residence, native language, etc.). The DCTs elicit different speech acts (requests and refusals) with different degrees of formality and are communicated via different media (email, IM, or VM). The scenarios represent authentic communication settings for the students. The sub-corpus of IM consists of a total of 1,204 items from 80 native speakers and 114 learners, amounting to 33,966 tokens. The sub-corpus of emails comprises 235 emails from 78 natives and 38 learners, amounting to 18,935 tokens. Finally, 20 natives and 25 learners produced in total 320 VM, amounting to 11,231 tokens. The collected data are published online (https://doi.org/10.5281/zenodo.6390255) in open access format and allow relevant comparison in sub-corpora e.g. proficiency levels.

The corpus LADDER has been used since 2021 in teacher education seminars at the University of Innsbruck. The student teachers have been involved in the data collection enhancing their experiences in empirical research. Moreover, students have been asked to conduct a data-analysis project based on the collected data. In particular, they have been trained in the use of the corpus for Italian L1 – Italian L2 comparative research in pragmalinguistics, conducting need analyses through the quantitative corpus-based methodology. This methodology allows student teachers to design activities and plan the curriculum on the basis of empirical evidence. Students' seminar papers are published online at https://ladder.hypotheses.org/.

As a case study, the contribution will present the results of a corpus investigation about sociopragmatic competencies in requests in IM and VM, with special emphasis on the comparison between native and non-native usage. The research focuses on the use of external modifiers (syntactic and lexical mitigators) of requests (Blum-Kulka & Olshtain, 1984). These pragmatic devices are required for the mastery of the B1 language level, which is essential for the final exam in Austrian schools and their (non)use correlates with the perception of politeness in requests (Savić, 2018). Thus, the research question is, how external modifiers of request are distributed in the messages of the natives and the learners and which type of modifiers are over- or underrepresented. The analysis follows a quantitative research approach based on the following steps:

- The selected sub-corpus is annotated according to an inductive taxonomy grounded on previous literature (Castineira-Benitez & Flores-Salgado, 2018).
- The annotations' reliability is checked with Cohen's K inter-rater agreement.
- Patterns according to the different variables (language proficiency -L1, B1-, used media-IM, VM-, social distance -friend or acquaintance-) are created.
- The results are presented and interpreted with a special focus on didactic outputs.

Overall, the results display that the learners encounter pragmalinguistic difficulties especially in managing the communication in VMs. Contrary to previous research stating that learners display verbose pragmatic behavior (Hassall, 2001), learners' performances are concise and external modifiers are underrepresented. In settings with higher social distance, learners can only count on a limited lexical repertoire and fail to fine-tune the mitigation devices producing none or too formal supportive moves. In VMs, learners' performance is more "straight to the

point" compared to the native baseline. These results will be interpreted with regards both of the researches on second language pragmatics development (Taguchi & Roever 2017) and of the researches on politeness in German-Italian intercultural comparison (Brocca et al. in preparation, Kunkel 2020, Venuti & Hinterhölzl 2019).

**References**


Artoni, D., Benigni V., & Nuzzo E. (2020). Pragmatic instruction in L2-Russian: a study on requests and advice. *Instructed Second Language Acquisition* 4(1): 62-95.

Blum-Kulka, S. & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics*, 5, 196-213.

Brocca, N. (2021). LADDER: Un corpus di scritture digitali per l'insegnamento della pragmatica in L2. Un esempio di analisi di disdette in WhatsApp. *Italiano Lingua Due*, 1, 241-259.

Brocca, N., Cortés Velásquez D., Nuzzo E., Rudigier M. (in preparation) *Linguistic politeness across Austria and Italy: Backing out of an invitation with an instant message.*

Castineira-Benitez, T. A. & Flores-Salgado, E. (2018). The use of politeness in WhatsApp discourse and move 'requests'. *Journal of Pragmatics*, 133, 79-92.

Hassall, T. (2001). Modifying requests in a second language. *International Review of Applied Linguistics (IRAL)* 39, 259-283.

Kunkel, M. (2020). *Kundenbeschwerden im Web 2.0. Eine kurpusbasierte Untersuchung zur Pragmatik von Beschwerden im Deutschen und Italienischen.* Tübingen, Narr Francke Attempto.

Nuzzo, E. & Cortés Velásquez D. (2020). Canceling Last Minute in Italian and Colombian Spanish: A Cross-Cultural Account of Pragmalinguistic Strategies. *Corpus pragmatics* 4 (ISSN: 2509-9507): 358.

Savić, M. (2018). Lecturer perceptions of im/politeness and in/appropriateness in student e-mail requests: A Norwegian perspective. *Journal of Pragmatics*, 124.

Taguchi N., Rover C. (2017). *Second language pragmatics.* Oxford: Oxford University Press

Trubnikova, V. & Garofolin, B. (2020). *Lingua e interazione. Insegnare la pragmatica a scuola.* Pisa: ETS.

Venuti, I. & Hinterhölzl R. (2019). "Überzeugungs- und Überredungsmittel in mündlichen Aufforderungsakten im deutsch-italienischen Sprachvergleich." *Linguistik Online*, 97(4), 209–224. https://doi.org/10.13092/lo.97.5603

## Challenges in the annotation and analysis of learner corpora

Marcus Callies
Universität Bremen
callies@uni-bremen.de

This talk will highlight and discuss the special characteristics of learner corpus data and the challenges these may present for researchers who want to engage in corpus compilation, annotation, and analysis but are new to LCR. Because learner corpus and SLA researchers use corpus data to study L2 production and development it is of utmost importance that the data are valid, i.e. they represent "authentic" L2 production.

Texts contained in learner corpora have, by definition, been produced by bi- or even multilingual individuals, thus multilingual practices and phenomena induced by language contact, such as code-switching, foreignizing or calquing, are commonplace. These present challenges for annotation and analysis alike (Callies & Wiemeyer 2017). Learner corpora, especially those of academic texts, contain expert terminology, metalinguistic language use, e.g. examples ("mentioned items"), citations, and sometimes even whole abstracts or thematic summaries from other languages. Such instances do not represent 'genuine' learner production as they are typically taken over or copied from secondary sources. They can thus be considered unwanted items or "false positives" as their inclusion in word counts and concordance analyses will distort the data. They should thus be specifically tagged so that they can be excluded from analysis and frequency counts (Callies & Wiemeyer 2017: 90).

A further challenge is unwanted lexical bias. This is introduced either by the topic of the task or because learners use words or phrases from the task description, the writing prompt, or other input material (see e.g. O'Donnell et al. 2013 for a description of how this may affect the use of lexical bundles in argumentative writing). It is important that researchers control for such effects because lexical variation, sophistication, and complexity are often considered proxies for L2 proficiency. Identifying lexical bias can be challenging, but if it is not discovered, its effects threaten the validity of the research findings. Words identified to cause lexical bias are either treated as stopwords, or L2 structures that are likely to have been brought about by lexical bias are excluded from the analysis. Similarly, task- and prompt-material may trigger the recurrent use of whole grammatical construction. For instance, Callies (2008) notes an effect of the writing prompt on the occurrence of raising constructions, and Alexopoulou et al. (2015) discuss various task effects on the use of relative clauses.

Finally, certain annotation methods used in Learner Corpus Research (LCR), but also in other disciplines, may introduce a certain bias. LCR is heavily influenced by SLA and its "discourse of deficit" (Ortega 2013) that is linked to the use of a monolingual native-speaker norm as the benchmark for the assessment of learner data. Error annotation thus often tends to be overly prescriptive. Creative and innovative but "non-standard" interlanguage forms (which are often contact-induced or formed on the basis of semantic or structural analogy to L2 input) may be considered "unwanted items" from an exonormative point of view, but they actually present valuable and highly interesting data for research into SLA and nativization processes in World Englishes (see e.g. Callies 2022). In the ICE corpora family, such cases are described in the tagging manual for written texts in a section on "Normalizing the text" (see Nelson 2002).

**References**

Alexopoulou, T., Geertzen, J., Korhonen, A. & Meurers, D. (2015). Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research* 1(1), 96–129.

Callies, M. (2008). Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety. In G. Gilquin, M.B. Diez-Bedmar & S. Papp (eds.), *Linking Contrastive and Learner Corpus Research* (Language and Computers. Studies in Practical Linguistics, Band 66). Amsterdam: Rodopi, 201–226.

Callies, M. (2022). Errors and innovations in L2 varieties of English: Towards resolving a contradictory practice. In G. Febel, K. Knopf, C. Nolte & M. Nonhoff (eds.), *Contradiction Studies: Mapping the Field*. New York: Springer.

Callies, M. & Wiemeyer, L. (2017). Multilingual speakers, multilingual texts: Multilingual practices in learner corpora. In A. Nurmi, T. Rütten & P. Pahta (eds.), *Challenging the Myth of Monolingual Corpora*. Amsterdam: Brill, 80–94.

Nelson, G. (2002). *International Corpus of English. Markup Manual for Written Texts*. https://www.ice-corpora.uzh.ch/dam/jcr:df7b1e8f-005f-4346-903b-c77b6c1da66a/written.pdf

O'Donnell, M. B., Römer, U. & Ellis, N. C. (2013). The development of formulaic language in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics* 18, 83–108.

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning* 63 (Suppl. 1), 1–24.

# What gets funded? A learner corpus study of grant proposal summaries by L1 Arabic-Syrian academics

Maggie Charles[1], Ahmed Halil[2], Michael Jenkins[3], Karin Whiteside[4]
University of Oxford UK[1], Selçuk University Turkey[2], Independent researcher[3], University of Reading UK[4]
maggiecharles_oxford@yahoo.com[1], ahmethalil8686@gmail.com[2], michaelojenkins.mj@gmail.com[3], k.whiteside@reading.ac.uk[4]

The grant proposal is one of the most high-stakes genres in academia since success in gaining funding enables researchers to pursue their research, publish their work and advance their careers (Connor & Mauranen, 1999; Myers, 1990; Swales, 1990). However, the proposal is an 'occluded genre' and this lack of public visibility makes the production of well-crafted proposals particularly challenging for inexperienced writers (Swales, 1996). The abstract/summary plays a key role in the proposal and has attracted growing research interest (Feng & Shi, 2004; Flowerdew, 2016; Matzler, 2021). Nonetheless, the investigation of proposal summaries written by learners is still very limited. Flowerdew describes a module to teach proposal summary writing to junior scholars in Hong Kong, but her learners' summaries are pedagogical tasks, not submissions for real funding. The summaries examined by Matzler and Feng and Shi were written by experts with considerable experience in applying for grant funding. Moreover, these studies do not present data which allows a comparison between funded and unfunded proposals.

The current paper examines a corpus of proposal summaries written by learners of English with Arabic L1 who are inexperienced in applying for research grants and it compares the summaries of funded and unfunded projects. The research questions are:

- RQ1 What is the generic structure of inexperienced learners' summaries?
- RQ2 What are the differences, if any, between the summaries of funded and unfunded proposals?

The learners are exiled Syrian academics on the Council for At-Risk Academics (Cara) Syria Programme (for details see Brewer & Whiteside, 2019). The charity offers small grants (roughly 650-6,500 Euros) for research projects conducted by Syrian participants. Submission requirements include a detailed proposal and a summary of 500 words maximum. Thirty-two submissions were received for the latest call, which resulted in 12 funded and 20 unfunded projects. The writers have CEFR levels B2 to C1 and their projects cover a wide range of disciplines. Corpora of the detailed proposals and the summaries were compiled using AntFileConverter (Anthony, 2017) to convert the files to plain text and AntConc (Anthony, 2020) to examine the data. This study draws on the corpus of summaries, which consists of 12,292 tokens and comprises two sub-corpora: funded (4,857 tokens) and unfunded summaries (7,435 tokens). The corpora are not currently tagged for moves and steps.

In order to describe and compare the summaries in a pedagogically helpful way, a genre analysis was conducted based on Feng and Shi's (2004) work. First, all four authors independently analyzed a pilot subset of six summaries and after extensive discussions, Feng and Shi's three moves were adopted: 1) *Justifying a research need*; 2) *Describing how to meet the research need*; 3) *Claiming potential contributions*, but their eight steps were re-defined and expanded to ten. The remaining 26 summaries were then analyzed by at least two researchers independently and problematic instances were resolved by discussion with all four authors and with unanimous agreement.

Findings on RQ1 show that 24 summaries (75%) use all three moves. However, only 18 (56%) follow the expected sequence: move 1, move 2, move 3, suggesting that work on the sequencing of moves would be useful. The use of the individual steps is also patchy. In move 1, the key step *Indicating a problem/research gap* is omitted in six summaries (19%), while in move 2, four (13%) fail to mention research methods. These omissions negatively affect the quality of the proposals and indicate areas for pedagogical tasks.

Comparing funded and unfunded summaries (RQ2) reveals that five of the unfunded summaries (25%) omit move 3 entirely compared to just one (8%) of the funded summaries. Move 3 consists of five steps, the most frequent being: *Achievements* (giving anticipated results) and *Benefits* (giving real-world contributions). 65% of the unfunded summaries include a *Benefits* step, but the use of *Achievements* is much lower at 35%. By contrast, 75% of the funded summaries include *Benefits*, while 58% include *Achievements*. Thus the unfunded summaries fail to exploit fully the potential of these two steps, losing the opportunity to evaluate the outcomes of their research positively, thereby making it potentially less attractive to funders. While both steps are important, a pedagogical focus on discussing anticipated results deserves particular attention.

Although these learner corpora are small, such findings shed light on the issues faced by inexperienced proposal writers. This paper reports further results and discusses their pedagogic applications.

**References:**


Anthony, L. (2017). *AntFileConverter* (1.2.1) [Computer software]. Tokyo, Japan: Waseda University. Available from: https://www.laurenceanthony.net/software

Anthony, L. (2020). *AntConc* (3.5.9) [Computer software]. Tokyo, Japan: Waseda University. Available from: https://www.laurenceanthony.net/software

Brewer, S., & Whiteside, K. (2019). The Cara Syria programme – combining teaching of English for Academic Purposes and academic and research skills development. *Language Learning in Higher Education*, *9*(1), 161–172. https://doi.org/10.1515/cercles-2019-0010

Connor, U., & Mauranen, A. (1999). Linguistic analysis of grant proposals: European Union research grants. *English for Specific Purposes*, 18(1), 47–62.

Feng, H., & Shi, L. (2004). Genre analysis of research grant proposals. *LSP and Professional Communication*, *4*, 8–32.

Flowerdew, L. (2016). A genre-inspired and lexico-grammatical approach for helping postgraduate students craft research grant proposals. *English for Specific Purposes*, *42*, 1–12. https://doi.org/10.1016/j.esp.2015.10.001

Matzler, P. P. (2021). Grant proposal abstracts in science and engineering: A prototypical move-structure pattern and its variations. *Journal of English for Academic Purposes*, 49, 100938. https://doi.org/10.1016/j.jeap.2020.100938

Myers, G. (1990). *Writing Biology: Texts in the social construction of scientific knowledge*. Madison WI: University of Wisconsin Press.

Swales, J. M. (1990). *Genre Analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (1996). Occluded genres in the academy: The case of the submission letter. In E. Ventola & A. Mauranen (Eds.), *Academic writing: Intercultural and textual issues*. Amsterdam: John Benjamins, 45-58.

## Disfluencies in L2 peer interaction: A corpus analysis of cognitive fluency

William J. Crawford
Northern Arizona University
william.crawford@nau.edu

Among the wide interpretations and applications of oral L2 fluency research, one useful distinction has been that of Segalowitz (2010, 2018) who differentiates between production (utterance fluency), underlying mental processes (cognitive fluency), and opinions of fluent speech by listeners (perceptual fluency). These distinctions have resulted in research informing the relationship between types of fluency such as studies exploring the connection between various components of utterance fluency (speed, repair, and breakdown measures) and perceptual fluency (Suzuki & Kormos, 2019). Another strand of research of this type focuses on the association between utterance fluency and cognitive fluency (Segalowitz, 2010, 2018; Kahng, 2020). Some studies investigating this relationship use separate measures of cognitive fluency such as speed of morphological processing (de Jong, et al., 2013) and lexical retrieval (Kahng, 2020) while other studies use utterance fluency measures (such as pause position in clauses) as an indirect measure of cognitive fluency (Saito et al., 2018). The present study adopts the latter of these two perspectives and investigates the relationship between utterance and cognitive fluency through a detailed analysis of disfluencies in a corpus of L2 peer interaction.

Disfluencies in L2 speech are an important component of language fluency (Skehan, 2003). Sometimes seen as a deficit behavior obstructing fluency (Levelt, 1989), disfluencies have important communicative functions benefiting both speaker and listener (de Jong, 2018; Foxtree, 2001). The relevance of disfluencies in L2 speech is also acknowledged in Götz (2019) who notes that the frequency and function of disfluencies are used to describe different proficiency level descriptions in the Common European Framework. From a similar perspective, specific types of disfluencies (such as pauses) have been used to describe fluency development in L2 speech. For example, using Levelt's (1989) model of speech production, Kormos (2006) analyzes the clausal positions of pauses to propose a developmental perspective of utterance fluency (e.g., mid-clause pauses are related to articulation at lower proficiencies and final clause pauses are related to conceptualization in higher proficiency speakers).

The present study provides a detailed comparison of disfluencies (operationalized as repeats and filled pauses such as *um*) in the Corpus of Collaborative Oral Tasks (CCOT), a collection of 775 spoken tasks (around 240,000 words) carried out by dyads of L2 English speakers (Crawford & McDonough, 2021). To investigate the relationship between cognitive and utterance fluency, two types of analyses are used. The first uses L1 disfluency as a baseline for comparison with L2 disfluencies. Segalowitz (2018) has maintained that disfluencies are likely related to L2-specific cognitive mechanisms. If disfluencies are, in fact, specific to the L1 or L2, one might expect systematic differences between L1 and L2 users of English. To this end, disfluencies are compared in the CCOT and the face-to-face conversation component of the Longman Grammar of Spoken and Written English. The analysis of disfluencies in the two corpora identifies cases where L2 users produce disfluencies in ways that are similar to native speakers (e.g., repeats of nominative case pronouns), which suggests parallel processing constraints between L1 and L2 performance. However, collocational analysis of other types of disfluencies (filled pauses) shows marked differences between L1 and L2 related to clausal positions which provide evidence for L2-specific processing mechanisms. These findings suggest that some aspects of cognitive fluency are related to general cognitive processing and others may be related to L2-specific cognitive processing. To further demonstrate another component of cognitive fluency, a second analysis illustrates how various tasks used in the CCOT show differential disfluency use, even in cases where proficiency is held constant. The implications of findings from both analyses are then discussed in relation to both L1/L2 cognitive fluency as well as task description in the Task-Based Language Teaching framework.

**References**


Crawford, W.& McDonough, K. (2021). The corpus of collaborative oral tasks. In W. Crawford (Ed.). *Multiple perspective on learner interaction: The corpus of collaborative oral tasks*, pp. 7-16. Mouton DeGruyter.

de Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*:3, 237-254,

de Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34*, 893–916.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory and Cognition*, *29*, 320–326.

Götz, S. (2019). Filled pauses across proficiency levels, L1s and learning context variables: A multivariate exploration of the Trinity Lancaster Corpus Sample. *International Journal of Learner Corpus Research, 5*, 159–180.

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied Psycholinguistics*, *41*, 457–480.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Erlbaum.

Levelt, W. (1989). *Speaking: From intention to articulation.* MIT Press.

Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics, 39*, 593–617.

Segalowitz, N. (2018). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics 54*, 79-95.

Segalowitz, N. (2010). *The cognitive bases of second language fluency*. New York: Routledge.

Skehan, P. (2003). *Task based instruction. Language Teaching, 36*,1–14.

Suzuki, S., & Kormos, J. (2019). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*. Advanced online publication. doi:10.1017/S0272263119000421

## A comparison between colloquial Persian used by English-speaking learners of Persian and Iranian speakers of Persian: Insights from a learner corpus-based study

Sepideh Daghbandan
University of Edinburgh
sepideh.daghbandan@ed.ac.uk

The current paper presentation sets out to describe a corpus-based study conducted to answer the following research question: "What are the differences between the colloquial Persian used by English-speaking learners of Persian and L1 Iranian speakers of Persian?"

To this aim, the paper is divided into three sections. First, the paper provides a brief overview of the contextual and theoretical background which informed the rationale of this study. The contextual background of the study stems from the lack of research on how learners of Persian use colloquial spoken Persian despite the growing interest in this language variety (Sedighi & Shabani-Jadidi 2018). This scarcity of research causes difficulty in identifying the problems that students may face in producing colloquial spoken Persian. In addition, the little research that does inform the teaching of spoken colloquial Persian is either based on small-scale classroom research, which is mainly influenced by teacher-researcher intuitions about the problems that the learners may be facing (Shabani-Jadidi 2020), or from research on first language speakers regarding the differences between colloquial Persian and the form closely associated with formal written Persian (Saffar Moghadam 2013). In addition to expanding in these areas on the research on teaching colloquial Persian, a short description of Conversational Grammar (Carter & McCarthy 2017; Leech 2000) that forms the theoretical base of this study is also provided.

Second, the learner corpus that was compiled for the purpose of analysis, namely, the Learner of Persian Spoken Corpus (LoPSC) is introduced. In this section, the design and data collection phase of the LoPSC are described, with special attention given to the challenges posed when collecting data from informal conversations between language learners that share the same first language background. The learners of this study were L1 speakers of English who were third- and fourth-year university students of Persian.

Since this study is a comparative study of learners' and first language speakers' language use, in addition to describing the learner corpus, a brief description of the first language corpus used as the Reference Corpus (RC) is also provided.

The paper then reports the analysis and results used in order to provide answers to the research question of this study. To analyze the data of this study, corpora tools, namely, Keywords, Collocations, and Concordances were used. First, the Keywords tool was used to indicate the positive and negative keywords of the LoPSC. That is, the words that occurred significantly higher, in the case of the positive keywords, and significantly lower, in the case of the negative keywords, in the LoPSC compared to the RC were identified. After this initial analysis using the Keywords' tools, the top five keywords were selected for further qualitative analysis to determine whether the identified keywords indicated any pragmatic differences. Collocations of the keywords and a closer look at the Concordance lines of the selected keywords aided this section of the analysis.

The results of the study indicated similar findings to previous literature with learners of other languages, especially regarding the use of discourse markers. The results also indicated that the same forms were used to perform different pragmatic functions in the two corpora. These results and their implications in the context of teaching colloquial Persian and the field of Second Language Acquisition are further expanded on in the presentation.

**References**

Carter, R., & McCarthy, M. (2017). Spoken Grammar: Where Are We and Where Are We Going? *Applied Linguistics*, 38(1), 1-20. https://doi.org/10.1093/applin/amu080

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning*, 50(4), 675-724.

Saffar Moghadam, A. (2013). Spoken and Written Variants in Teaching Persian Language to non-Persian Speakers. *Language Studies*, 3(6), 45-68.

Sedighi, A., & Shabani-Jadidi, P. (2018). *The Oxford Handbook of Persian Linguistics*. Oxford: Oxford University Press USA - OSO.

Shabani-Jadidi, P. (2020). *The Routledge handbook of second language acquisition and pedagogy of Persian*. New York: Routledge.

# The differential effect of specificity on definite and indefinite article accuracy in learner English

Kateryna Derkach[1], Dora Alexopoulou[2]
University of Cambridge
kyp22@cam.ac.uk[1], ta259@cam.ac.uk[2]

English articles are notoriously difficult to acquire, especially for learners with article-less, or [-art], L1s (Murakami & Alexopoulou 2016). Several factors, besides the presence/absence of articles in the L1, have been shown to play a role, namely specificity, or referentiality, of the nominal (Ionin et al. 2004), number and count/mass distinction (Snape 2008), prenominal modification (Trenkic 2007), article discourse functions (Liu & Gleason 2002; Robertson 2000), abstractness and syntactic position of the nominal (Hua & Lee 2005). Nevertheless, no study has considered all these features together, so their relative importance and interactions are not known.

**Research Questions**

1. What is the relative importance of the different factors which impact article accuracy in learner English, i.e. which factors have the strongest effect?
2. How do these factors interact in predicting learner accuracy and error types, i.e. article omission, article oversuppliance (using *a* or *the* where no article is needed), article substitution (using *a* instead of *the* and vice versa)?

**Data**

We analyzed 660 written scripts randomly selected across proficiency levels from CEFR level A2 to B2 (inclusive) from a large open-access learner corpus, EFCAMDAT, the EF Education First Cambridge Open Language Database (Geertzen et al. 2013). EFCAMDAT contains writings submitted in response to communicative tasks, such as writing a holiday postcard, a film review, describing a terrifying experience, etc., to EF's online language school. Our sample contains data from learners with four typologically distinct L1s: German and Brazilian Portuguese, both [+art], and Chinese and Russian, both [-art].

**Method**

We manually retrieved all the nominals from our subcorpus, excluding those preceded by demonstratives (e.g. *this*) and quantifier items (e.g. *many*, *a few*), as well as excluding any formulaic sequences (e.g. *do the shopping*), which resulted in a total of 5772 nominals. These were automatically coded for learner L1 and proficiency level (based on corpus metadata) and then manually coded according to target article (*a*/*the*/Ø), error type (omission/oversuppliance/substitution), noun type (count singular/count plural/mass), specificity[1] of the nominal (specific/non-specific), abstractness of the noun (abstract/concrete), syntactic position of the nominal (subject/object/predicate/existential), prenominal modification (present/absent), discourse function for the definite article (anaphoric/situational/explanatory). We used a generalized linear mixed-effects logistic regression model to estimate the effect of the coded variables and their interactions.

**Results**

Our analysis confirmed the lower overall accuracy of learners with [-art] L1s. Apart from this well-established factor, we revealed that L2 English article use is significantly affected by specificity, prenominal modifier presence, and syntactic position, but not by abstractness or discourse-pragmatic context.

Our key finding is the differential effect of specificity on definite and indefinite articles: learners are significantly more likely to use *a* with specific than with non-specific indefinite, which in the case of count singular nouns results in article omission with non-specific nominals (1)[2], and in case of mass nouns result in oversuppliance of *a* with specific nominals (2).

(1) I have many dreams […] I'd make [a] career in my business and have a fulfilled and balanced life. (L1-German)

[1] We define a specific nominal as one that refers to a certain entity which exists in the world (real or imaginary) and which the speaker has in mind.

[2] Learner examples are given with their original spelling and grammar, article errors are corrected in square brackets, oversupplied or incorrectly used articles are marked with an asterisk.

(2) When police got the home they noticed that one servant's face was covered with *a [Ø] red paint. (L1-Russian)

Prenominal modifiers appear to further contribute to perceived specificity and, for mass nouns, perceived countability, leading to significantly higher article overuse with modified indefinite mass nouns (2).

In definite contexts, specificity does not have an effect on article accuracy, but prenominal modifiers are associated with increased article omission (3), suggesting that modifiers and *the* have similar roles in learners' interlanguage grammars as referent identifiers.

(3) I first met my friend, Kolya, when I was working in advertising project five years ago. […] Kostya and I enjoy working on [the] advertising project together. (L1-Russian)

The second key finding is that the effect of specificity applies across all L1s included in this study. The differences between L1s are merely quantitative, which suggests that the effect is not driven by the absence of an article system in learners' L1.

**References**


Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). *Selected Proceedings from 31st Second Language Research Forum*.

Hua, D., & Lee, T. H. (2005). Chinese ESL learners' understanding of the English count-mass distinction. *Proceedings of the 7th Generative Approaches to Second Language Acquisition Conference*, 138–149.

Ionin, T., Ko, H., & Wexler, K. (2004). Article semantics in L2 acquisition: The role of specificity. *Language Acquisition*, *12*(1), 3–69.

Liu, D., & Gleason, J. L. (2002). Acquisition of the article "the" by nonnative speakers of English: An Analysis of Four Nongeneric Uses. *Studies in Second Language Acquisition*, *24*, 1–26. https://doi.org/10.1017/S0272263102001018

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: a learner corpus study. *Studies in Second Language Acquisition*, *38*, 365–401. https://doi.org/10.1017/S0272263115000352

Robertson, D. (2000). Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, *16*(2), 135–172. https://doi.org/10.1191/026765800672262975

Snape, N. (2008). Resetting the Nominal Mapping Parameter in L2 English: Definite article use and the count–mass distinction. *Bilingualism: Language and Cognition*, *11*(01), 63–79. https://doi.org/10.1017/S1366728907003215

Trenkic, D. (2007). Variability in second language article production: Beyond the representational deficit vs. processing constraints debate. In *Second Language Research* (Vol. 23, Issue 3). https://doi.org/10.1177/0267658307077643

# "*The store is wanting staff*": A multifactorial approach to progressive marking in Korean English

Deshors Sandra C.[1], Steven Gagnon[2]
Michigan State University
sdeshors@gmail.com[1], gagnons2@msu.edu[2]

This multifactorial corpus-based study focuses on the usage patterns of progressive marking (specifically the progressive vs. nonprogressive alternation) in Korean Learner English (KLE) and how those patterns differ from those in native English. The semantic complexity of the progressive makes it a highly interesting construction to study across native and learner Englishes: while it generally denotes aspectual information and describes an action in progress and ongoing at the time of speaking, research has shown that over time it has developed a complex set of novel subjective and non-aspectual semantic meanings (Kranich 2010) where the difference between non-progressive and progressive constructions cannot be explained based on notions such as duration or dynamics (Nesselhauf 2007). From a second-language standpoint, this complexity is an acquisitional challenge for learners and while some learners tend to rely on prototypical uses of progressive constructions rather than combine them with, for instance, stative verbs (Hundt & Vogel 2011), learners with aspect-marking native languages (e.g., Chinese and Japanese English learners) extend progressive uses more than others as a result of native language influence (Meriläinen et al. 2017). In this context, this study adds KLE to the list of learner Englishes already explored, based on key formal, morphological and semantic typological differences between progressive marking in Korean and English. Drawing from those differences, the present study explores KLE in a systematic, fine-grained, and context-based fashion with the objective (i) to identify the linguistic contexts that characterize progressive and non-progressive constructions respectively in KLE, (ii) to assess to what degree the linguistic context of the progressive construction in KLE varies from that in native English, (iii) pinpoint the contextual linguistic features that trigger a progressive construction in KLE, and (iv) assess to what extent those features differ from those in native English.

Methodologically, we explored over 2,600 occurrences of spoken and written (non-) progressive constructions as extracted from the Korean and native English subsections of the *International Corpus Network of Asian Learners of English* (ICNALE) and manually annotated for nine linguistic factors, i.e., aspect, lemma, variety, aksionsart, perfect, tense-modality, mode, voice and animacy of the grammatical subject. Statistically, these factors were analyzed with a distinctive collexeme analysis (DCA; Stefanowitsch & Gries 2003) and a generalized linear mixed model (GLMM) tree analysis, inspired by Rautionaho (2020). With the DCA, we conducted a first analysis based on the predictors' aspect and variety to assess to what degree KLE associates more or less strongly with the progressive construction compared to native English. We then conducted a second analysis based on aspect and aktionsart to assess how strongly different types of lexical aspects (accomplishment, achievement, state, process) associate with (non-)progressive constructions to capture potential traces of stative progressives in KLE that native-language influence could trigger. With the GLMM, we analyzed all our predictors simultaneously to assess their joint influence on the progressive in KLE.

Overall, we found that speakers and learners prefer different constructions equally strongly: learners opt for non-progressive constructions more frequently than native speakers. The DCA with lexical aspect yielded clear differences between native speakers and learners, with state verbs specifically. For instance, learners systematically produce more stative progressives than native speakers, which suggests the existence of L1 transfer in this type of linguistic context. The GLMM analysis returned a strong model ($C$ = 0.92). Overall, it emerged that, unlike native speakers, (i) learners prefer to stay away from the progressive regardless of the animacy of the grammatical subject, and (ii) with process verbs in the active voice, they remain unaffected by the animacy of the grammatical subject; however, unlike with modal auxiliaries, they prefer the progressive, (iii) with accomplishment/achievement verbs in the past tense learners prefer the unmarked form. Ultimately, our results reveal that at a fine-grained level of granularity, there is a disconnect between the diversity of the linguistic contexts that characterize progressive marking in native English and the linguistic contexts that trigger a progressive construction in KLE. Situating our results in the context of current pedagogical practices in Korea, altogether, these results bear important implications for the development of data-driven teaching and learning resources to complement the communicative approach currently adopted in the Korean National Curriculum.

**References**

Hundt, M., & Vogel, K. (2011). Overuse of the progressive in ESL and learner Englishes–fact or fiction? In J. Mukherjee & M. Hundt (Eds), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 145-165). Amsterdam: John Benjamins.

Kranich, S. (2010). *Progressive in modern English: A corpus-based study of grammaticalization and related changes*. Amsterdam: Rodopi.

Meriläinen, L., H. Paulasto & Rautionaho, P. 2017. Extended uses of the progressive in Inner, Outer and Extending Circle Englishes. In M. Filppula, J. Klemola, A. Mauranen & S. Vetchinnikova (Eds.), *Changing English: Global and Local Perspectives* (pp. 191–216). Berlin: De Gruyter Mouton.

Nesselhauf, N. (2007). The spread of the progressive and its 'future' use. *English Language & Linguistics*, *11*(1), 191-207.

Rautionaho, P. (2020) Revisiting the myth of stative progressives in world Englishes. *World Englishes*, 1-24.

Stefanowitsch, A., & Gries, S., (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209-243.

# Error identification, normalization and tagging: Three inter-annotator agreement experiments in a picture-elicited learner corpus

Elisa Di Nuovo[1], Bianca Maria De Paolis[1,2], Cristina Bosco[1], Elisa Corino[1]
University of Turin[1], University of Paris[2]
{elisa.dinuovo, biancamaria.depaolis, cristina.bosco, elisa.corino}@unito.it

Annotating learner texts, which often feature non-standardized forms, is an especially challenging task, but with some important side effects: it brings out both the range of valid interpretations of the same non-canonical forms and unforeseen problematic areas and annotator biases (Hovy & Prabhumoye 2021).

Calculating IAA is progressively becoming a common practice in Computational Linguistics, which consists in comparing the decisions of two or more annotators about the same product (Artstein 2017). The first error-annotated learner corpora were usually tagged by a single coder and revised by another one, e.g. CLC (Nicholls 2003), thus not reporting Inter-annotator agreement (IAA) studies. This issue was first raised with respect to learner corpora by Meurers & Müller 2009. Since then, several scholars have started to pay attention to this topic (Rozovskaya & Roth 2010, Boyd 2012, Lee et al. 2012, Dahlmeier et al. 2013, Rosen et al. 2014, Köhn & Köhn 2018, Boyd 2018, Del Río & Mendes 2018). Our three IAA experiments—error identification, normalization, and tagging—are measured using Cohen's κ, germane to other learner corpus-based IAA studies. The annotation is conducted by two annotators on the novel treebank VALICO-UD, i.e. a subcorpus of VALICO (Corino & Marello 2017) in UD format, which associates each Learner Sentence (LS) with a corresponding correct version (called Target Hypothesis - TH) using an error tagging system adapted from Nicholls 2003 (error tagging system description and the treebank are respectively available at https://bit.ly/3xB2WJ3 and https://universaldependencies.org/treebanks/it_valico/index.html).

In this study, we aim at answering three research questions:
- Is error identification more reproducible using picture-elicited texts (i.e. within a constrained linguistic and extra-linguistic context)?
- When different annotators agree on the presence of an error, do they agree also on its normalization?
- Is error tagging more reliable when based on explicit THs?

Concerning the first question, the results show that error identification with picture-elicited texts is more reproducible than with texts elicited under less controlled conditions. In our experiment and Köhn & Köhn 2018 (COMIGS, picture-elicited corpus) annotators achieved an almost perfect and substantial agreement (k=0.82 and k=0.79, respectively), in Landis & Koch's (1977) terms. While as texts are elicited under less controlled situations, agreement decreases (k=0.64 in Boyd 2018 based on a corpus of reading exercises and k=0.39 in Dahlmeier et al. 2013 on a corpus of essays). However, it should be noted that the mentioned studies are not directly comparable since they differ in terms of targeted languages, types of investigated errors, and error annotation systems.

IAA on error normalization with constrained linguistic and extra-linguistic context achieves substantial agreement (our corpus k=0.69, COMIGS k=0.64). The lower IAA agreement on error normalization compared to error identification could be explained by the non-deterministic nature of this task. However, when analyzing the sources of disagreement, it emerged that more than 40% of disagreement was caused by mistakes due to distraction (e.g. one annotator does not notice the spelling error in *all'improviso* instead of *all'improvviso* 'suddenly') or format (i.e. the possibility to create the same normalization with two different sets of annotation, Dahlmeier et al. 2013). The results reported in Boyd 2018 and Dahlmeier et al. 2013 are not comparable because the former reported only an agreement percentage (i.e. 70%), the latter also considered the associated error tag.

As far as error tagging is concerned, IAA was computed twice, again to remove apparent disagreement due to human or format-induced errors. Before the revision, annotators reached a moderate agreement (using a tagset of 148 unique tags, they achieved a k=0.50). After the revision, annotators reached an almost perfect agreement (k=0.95).

In conclusion, our results show that annotation on picture-elicited texts can reach almost perfect agreement and quantitatively confirm what is hypothesized in the literature (Lüdeling 2008; Reznicek et al. 2013; Meurers 2015; Rosen et al. 2014), i.e. explicit THs improve the replicability and reliability of the analysis. In addition, error tagging performed with explicit THs can be used to validate the error tagset. Some flaws emerged in these three experiments and should be considered to improve the annotation procedure: (i) the need for a second

round of annotation to avoid distraction and format errors; (ii) an even more specific set of guidelines that could overcome different possible interpretations of error substance (Dobrić & Sigott 2014).

**References**


Artstein, Ron (2017). "Inter-annotator agreement". *Handbook of linguistic annotation*. Springer, 297–313.

Boyd, Adriane (2012). "Detecting and diagnosing grammatical errors for beginning learners of German: From learner corpus annotation to constraint satisfaction problems". PhD thesis. The Ohio State University.

— (2018). "Normalization in Context: Inter-Annotator Agreement for Meaning-Based Target Hypothesis Annotation". *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning* (NLP4CALL 2018), 10–22.

Corino, Elisa and Carla Marello (2017). Italiano di stranieri. I corpora VALICO e VINCA. Guerra, Perugia.

Dahlmeier, Daniel, Hwee Tou Ng, and Siew Mei Wu (2013). "Building a large annotated corpus of learner English: The NUS corpus of learner English". *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 22–31.

Del Río, Iria and Amália Mendes (2018). "Error annotation in the COPLE2 corpus". *Revista Da Associação Portuguesa De Linguística* 4, 225–239.

Dobrić, Nikola and Guenther Sigott (2014). "Towards an error taxonomy for student writing". *Zeitschrift für interkulturellen Fremdsprachenunterricht* 19(2).

Hovy, Dirk and Shrimai Prabhumoye (2021). "Five sources of bias in natural language processing". *Language and Linguistics Compass* 15(8), e12432.

Köhn, Christine and Arne Köhn (2018). "An annotated corpus of picture stories retold by language learners". *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions* (LAW-MWE-CxG-2018), 121–132.

Landis, J. Richard and Gary G. Koch (1977). "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers". *Biometrics*, 363–374.

Lee, Sun-Hee, Markus Dickinson, and Ross Israel (2012). "Developing learner corpus annotation for Korean particle errors". *Proceedings of the Sixth Linguistic Annotation Workshop*, 129–133.

Lüdeling, Anke (2008). "Mehrdeutigkeiten und kategorisierung: Probleme bei der annotation von lernerkorpora". *Fortgeschrittene Lernervarietäten*, 119–140.

Meurers, Detmar (2015). "Learner corpora and natural language processing". *Cambridge Handbook of Learner Corpus Research*. Ed. by Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier. Cambridge University Press Cambridge, UK, 537–566.

Meurers, Walt Detmar and Stefan Müller (2009). "Corpora and Syntax (Article 42)". *Corpus linguistics*. Ed. by Anke Lüdeling and Merja Kytö 2. BerlMouton de Gruyter, 920–933.

Nicholls, Diane (2003). "The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT". *Proceedings of the Corpus Linguistics 2003 Conference* 16, 572–581.

Reznicek, Marc, Anke Lüdeling, and Hagen Hirschmann (2013). "Competing target hypotheses in the Falko corpus". *Automatic treatment and analysis of learner corpus data* 59, 101–123.

Rosen, Alexandr, Jirka Hana, Barbora Štindlová, and Anna Feldman (2014). "Evaluating and automating the annotation of a learner corpus". *Language Resources and Evaluation* 48(1), 65–92.

Rozovskaya, Alla and Dan Roth (2010). "Annotating ESL errors: Challenges and rewards". *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, 28–36.

## Connector placement in EFL learner writing: Focus on *however*

Dupont Maïté[1], Granger Sylviane[2]
University of Louvain & Université Saint-Louis - Bruxelles[1], University of Louvain[2]
maite.dupont@uclouvain.be[1], sylviane.granger@uclouvain.be[2]

The use of adverbial connectors is a very popular topic in learner corpus research. This is perfectly justified in view of the important role they play in building clear and convincing argumentation and the difficulty even advanced learners experience in using them appropriately in academic writing. Most studies so far have compared the frequency of connectors in learner versus native reference corpora (e.g. Lenko-Szymanska 2008). These comparisons have yielded contradictory results, some pointing to a general overuse (Lei 2012, Güneş 2017), some to overall underuse (Altenberg & Tapper 1998), while others do not reveal any quantitative difference (Granger & Tyson 1996). Native/non-native comparisons have also highlighted differences in learners' preferred semantic categories of connectors (adversative, causal, etc.). One aspect that has rarely been investigated is placement. A key characteristic of adverbial connectors is that they are mobile. According to Quirk et al. (1985: 643), the default position in English is the initial position; the medial position is "quite normal" for conjuncts that cannot be misinterpreted in this position, while the final position is restricted to a handful of connectors. The few studies that have studied connector position in learner corpora point to overuse of the initial position (e.g. Narita & Sugiura 2006, Van Vuuren & Berns 2018). However, they tend to be based on very small datasets and cover a limited number of L1 populations. In addition, while the positioning of connectors induces a range of rhetorical effects (e.g., in the case of, *however*, emphasizing a contrast between two ideas or laying focus on some parts of the message), this aspect has hardly been investigated in learner corpus research.

The aim of our study is to investigate connector placement in a large learner corpus covering a variety of L1 populations through the lens of the highly frequent connector, *however*. Learners' placement patterns will be compared to those observed in the argumentative writing of both (i) expert native writers; and (ii) novice native writers of English. Thus, the influence of both L1 and the degree of expertise (expert vs novice) will be assessed. More specifically, the study attempts to answer the following research questions:

1) Does the placement of *however* by EFL learners differ from the placement preferences of (i) expert native writers and (ii) novice native writers of English?
2) Are there differences between the different L1 learner populations, and do some L1 populations better approximate the placement preferences of expert or novice native writers?
3) Are the different connector positions associated with similar rhetorical effects in the learner, novice, and expert writing?

The study is based on a 5-million-word corpus of argumentative texts written by EFL learners from 24 L1 learner populations extracted from the third version of the International Corpus of Learner English (ICLEv3) (Granger et al. 2020). The placement preferences of expert writers are identified on the basis of a 2-million-word corpus of newspaper editorials, a genre which is arguably close to the argumentative essays of the ICLE in terms of both length and overall communicative purpose (cf. Neff et al. 2004). The LOCNESS (Louvain Corpus of Native English Essays) was used to analyze the patterns of novice native speakers of English. All the occurrences *however* were extracted and disambiguated so as to weed out premodifying uses. The disambiguated data set, amounting to 8,000+ occurrences *however*, was categorized on the basis of Dupont's (2021) study of the placement of connectives of contrast. The adverb classification is inspired by Halliday's Systemic Functional description of thematic structure (see Halliday & Matthiessen 2014) and distinguishes between five positions (thematic 1 and 2 and rhematic 1, 2, and 3), exemplified in (1) to (5).

(1) *However*, members of the royal family tend to act without thinking.
(2) Worryingly, *however*, our survey also found that […].
(3) Russia, *however*, has failed to rise to the challenge of creating a real democracy.
(4) Mr. Adams has, *however*, stopped short of recommending […].
(5) The cute comparisons don't always apply, *however*.

Preliminary results strongly confirm that learners of English markedly overused *however* in thematic 1 (i.e. initial) position as compared to expert writers. By contrast, learners tend to largely underuse *however* in rhematic 1 position, i.e. after the subject or a fronted adjunct (as in example (3)). These two tendencies are observed across all 24 learner populations, thereby pointing to a developmental pattern rather than an L1-induced feature. Once

all the data has been analyzed, the random forest technique (Levshina 2020) will be applied to assess the relative importance of the degree of expertise and L1 on the placement of, *however*. The rhetorical effects achieved through connector placement by the three groups of writers will also be compared. Although our study is focused on only one connector, we will highlight some of its wider implications for the teaching of connectors in general.

**References**


Altenberg, B. & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learner's written English. In S. Granger (Ed.) *Learner English on Computer*. London: Longman, 80-93.

Dupont, M. (2021). *Conjunctive Markers of Contrast in English and French: From Syntax to Lexis and Discourse*. Amsterdam & Philadelphia: Benjamins.

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English. Version 3.* Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. & Tyson, S. (1996). Connector Usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15(1), 17-27.

Güneş, H. (2017). A corpus-based study of linking adverbials through contrastive analysis of L1/L2 PhD dissertations. *International Journal of Curriculum and Instruction*, 9(2), 21-38.

Halliday, M.A.K. & C. Matthiessen (2014). *Introduction to Functional Grammar*. London: Routledge.

Lei, L. (2012). Linking adverbials in academic writing on applied linguistics by Chinese doctoral students. *Journal of English for Academic Purposes*, 11, 267-275.

Lenko-Szymanska, A. (2008). Non-native or non-expert? The use of connectors in native and foreign language learners' texts. *Acquisition et interaction en langue étrangère*, 27: 91-108.

Levshina, N. 2020. Conditional inference trees and random forests. In S. T. Gries & M. Paquot (Eds.). *A Practical Handbook of Corpus Linguistics*. Springer: New York, 611-643.

Narita, M. & Sugiura, M. (2006). The use of adverbial connectors in argumentative essays by Japanese EFL college students. *English Corpus Studies*, 13, 23-42.

Neff, J., Ballesteros, F., Dafouz, E., Martínez, F., Rica, J.P., Díez, M., & Prieto, R. (2004). Formulating writer stance: A contrastive study of EFL learner corpora. In U. Connor & T. Upton (Eds.). *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi Brill, 73-89.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language.* London: Longman.

Van Vuuren, S. & Berns, J. (2018). Same difference? L1 influence in the use of initial adverbials in English novice writing. *IRAL*, 56(4), 427-461.

# '*Growing up students*': A collostructional analysis approach to phrasal verbs in Korean learner English

Deshors Sandra C.[1], Steven Gagnon[2]
Michigan State University
sdeshors@gmail.com[1], gagnons2@msu.edu[2]

This study investigates phrasal verbs (PVs) in (in)transitive constructions across native English and Korean learner English (KLE). As multi-word units, PVs combine a lexical verb and a particle. As such, they are one of the most difficult structures to learn and teach. Their complexity lies both in their semantic and syntactic properties: semantically, although they serve as single units, they can express a range of literal, idiomatic, and semi-idiomatic meanings. Syntactically, they occur in different configurations: Verb Particle (VP), Verb Particle Object (VPO), and Verb Object Particle (VOP) constructions. From an acquisitional standpoint, these semantic and syntactic properties are challenging for learners who experience difficulties establishing the limits between grammar and lexis (Alejo Gonzáles 2010). In this context, studies such as Gilquin (2015) and Deshors (2016) have shown that, as verb-particle combinations, (i) PVs associate more/less strongly with particular syntactic constructions, and (ii) association patterns vary across native and learner Englishes. Although deviant uses of PVs in learner English have been well documented in these studies, the case of KLE remains relatively unexplored despite (a) typological differences between English and Korean that suggest the existence of KLE-specific usage patterns of PVs and (b) the challenges that Korean learners experience using PVs syntactically (Sung & Kim 2016) and regarding form-meaning pairings (Lee 2003). In this context, the present study digs deeper into PV constructions by assessing degrees of mutual attraction between verbs and particles and between PVs and their semantic uses. Specifically, we explore to what extent

(i) lexical verbs and particles attract in KLE;
(ii) phrasal verbs and semantic uses attract in KLE;
(iii) pairings of lemmas, particles, and semantic uses vary across KLE and native English; and to what extent
(iv) the strength of those pairings varies across speaker populations.

Methodologically, we extracted approximately 1,500 occurrences of PVs across [VP], [VPO], and [VOP] constructions from the written *Yonsei English Learner Corpus* for the learner data and the *Louvain Corpus of Native English Essays* for the native data, on the basis of twenty-four particles (*aboard, about, across, ahead, along, apart, around, aside, away, back, by, down, forth, forward, in, off, on, out, over, round, through, together, under,* and *up*). Extracted occurrences were annotated for lemma, particle, and semantic uses (literal, idiomatic, completive, continuative, inceptive). Statistically, association strengths were measured using Stefanowitsch & Gries' (2005) co-varying collexeme analysis approach. Overall, within individual constructions, different verbs and particles (*grow* and *up* in *grow up* or *get* and *along* in *get along*) combine to different degrees, suggesting that, as cognitive routines, those combinations are not equally entrenched. Further, individual constructions affect verb-particle association strengths differently: in KLE, *clean* and *up* associate more strongly in [VPO] constructions than they do in [VOP] constructions. Individual constructions also affect PVs semantically: in [VPO], the strongest association of a PV and a semantic use were observed in KLE with a completive *turn off* and in native English with a literal *pull out*, suggesting that learners operate at a relatively high level of semantic complexity. Overall, for a learner population that has been shown to avoid phrasal verbs due to the lack of PV constructions in Korean, Korean English learners' uses of these constructions are surprisingly varied, particularly with intransitive and [VPO] constructions. With regards to lemma-particle pairings specifically, we found a disconnect between learners and native speakers: even though both populations use similar particles in PV constructions (particularly with [VP]), compared to native speakers, learners combine them with a larger variety of lexical verbs. Further, these combinations are often stronger in the learner data both in [VP] and [VPO] configurations. With [VOP], however, usage patterns are weak across both speaker populations. Semantically, our results confirm that learners' difficulties lie at the grammar-lexis interface. Specifically, learners are yet to fully integrate the notion that the syntactic configurations in which lemmas and particles combine most strongly vary as a function of individual phrasal verbs and individual semantic uses. Pedagogically, this study bears important implications, namely the need to adopt teaching approaches (e.g., input-flooding) and resources that help learners develop more confident uses of PVs in [VOP] constructions, particularly with resultative semantic uses.

**References**

Alejo Gonzáles, R. (2010). Making sense of phrasal verbs: A cognitive linguistic account of L2 learning. *AILA Review* 23:51-71.

Deshors, S. C. (2016). Inside phrasal verb constructions: A co-varying collexeme analysis of verb-particle combinations in EFL and their semantic associations. *International Journal of Learner Corpus Research, 2*(1), 1-30.

Gilquin, G. (2015). The use of phrasal verbs by French-speaking EFL learners A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory, 11*(1), 51-88.

Lee, C. (2013). A corpus-based study of the particle up in Korean high school students; writing. *Language Research, 49*(3).

Stefanowitsch, A., & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory, 1*(1), 1-43.

Sung, M., & Kim, H. (2016). Tracing developmental changes in L2 learners' structuring of phrasal verbs: A corpus study of native and non-native argumentative essays. *3L, Language, Linguistics, Literature, 22*(2).

# Exploring the operationalisation of L2 microsystems as functional complexity metrics for proficiency assessment

Thomas Gaillat
University of Rennes
thomas.gaillat@univ-rennes2.fr

This paper focuses on designing and evaluating new functional complexity metrics for the prediction of CEFR levels. Functional complexity is a sub-construct of structure complexity (Bulté & Housen, 2012, p. 25). It relies on the mappings between forms and functions of linguistic forms. It has been operationalized in various ways such as specific parts of speech or dependency relations (Settles et al., 2018) or syntactic constituents as in CTAP's feature selector module (Chen & Meurers, 2016). The use of functional complexity features offers two advantages for studies in the field of Second Language Acquisition. First, based on learner corpora, these features can be used to design metrics exploited for modelling purposes in prediction tasks such as CEFR classification (Kyle, 2016; Vajjala & Rama, 2018; Yannakoudakis et al., 2011). Secondly, using functional complexity features, which are both descriptive and potentially significant for proficiency, would help with the design of specific linguistic feedback that is meaningful for learners and teachers (Riemenschneider et al., 2021).

To understand the internal systems that learners build (Ellis, 1994, p. 140), it is necessary to explore functional features, i.e. how various forms are mapped to single language functions. Such features can be determined by bijective form-function mappings, but they can also be composed of multiple forms mapped to one function. Our proposal is to compute functional metrics which inform how likely a group of syntactic forms, mapped to the same language function, is likely to occur across CEFR levels. For this purpose, our approach relies on learner-specific microsystems whereby learners' confusions between forms of the same paradigmatic relations and linguistic functions are captured. For instance, learners hesitate between IT, THIS, and THAT as proforms when referring to discourse entities (Gaillat, 2016). We intend to measure the likelihood of use of each of these three proforms in relation to its two other competitors in the microsystem. Our research objective is to assess the internal variations of several microsystems across proficiency. Following Gaillat et al. (2021), we focus on the aforementioned proforms, the modals MAY, MIGHT, and MUST as well as the FOR-TO prepositional microsystem. Experience in correcting learner writings shows evidence of confusion between these forms. Our research question can be formulated as follows: which form variations can be observed within microsystems across CEFR levels?

We apply a data-driven approach grounded in supervised learning. We use the Spanish subset (N = 8,187 writings) of the EFCAMDAT corpus (Geertzen et al., 2013). After preprocessing the data with UDPipe (Straka et al., 2016), we subset the microsystems' forms and extract the linguistic features linked to these forms. Each observation corresponds to one occurrence of each form which is the dependent variable. Independent variables include dependency relations of forms, CEFR level, syntactic parent's Part-of-Speech (POS) and POS of adjacent tokens. Modelling is conducted with the multinomial logistic regression function (datasets and scripts available on Github) included in the nnet library (Ripley, 2022) in R (R Core Team, 2012). The logistic regression method outputs the probabilities for the different levels of the microsystems' forms depending on the independent variables. First, we evaluate the power of the model on a random test set made up of 25% of the data. Secondly, the estimated probabilities of each form in a microsystem are cross-tabulated with the CEFR levels. We can then visualize which of the forms are more likely to be used at a given level.

Preliminary results for the three microsystems show that classification is encouraging but the p-values of features show that more discriminative features need to be found. The proform microsystem shows 0.5587 accuracy (95% CI = 0.5223, 0.5946, $p < .001$, Nagelkerke's R2 = 0.399). The preposition *for_to* microsystem shows 0.7 accuracy (95% CI = 0.6833, 0.7163, $p < .001$, Nagelkerke's R2 = 0.358) and the modal microsystem's is 0.5079 (95% CI = 0.3789, 0.6362, $p = .003119$, Nagelkerke's R2 = 0.482). Visualizing the variations in the probabilities of occurrence of the forms reveals some patterns. Results show clear variations within the preposition and modal microsystems, indicating different patterns according to the CEFR levels. The proform microsystem presents less distinctive variations indicating unclear patterns.

Further work is needed to improve feature selection and to assess UDPipe's performance on all levels of L2 English. Ultimately, we intend to evaluate the microsystem probabilities as metrics in higher-level tasks such as proficiency prediction in writing. If validated, these metrics could also support feedback messages in Computer-

Aided Language Learning (CALL) systems by linking learner's form usage to language functions.

**References**


Bulté, B., & Housen, A. (2012). *Defining and Operationalising L2 Complexity*. John Benjamins Publishing Company.

Chen, X., & Meurers, D. (2016). CTAP : A Web-Based Tool Supporting Automatic Complexity Analysis. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 113119. http://aclweb.org/anthology/W16-4113

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford University Press.

Gaillat, T. (2016). *Reference in Interlanguage : The case of this and that. From linguistic annotation to corpus interoperability* [Thesis]. Université Paris-Diderot.

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (2021). Predicting CEFR levels in learners of English : The use of microsystem criterial features in a machine learning approach. *ReCALL*, 117. https://doi.org/10.1017/S095834402100029X

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In R. T. Miller, K. I. Martin, C. M. Eddington, A. Henery, N. Miguel, A. Tseng, A. Tuninetti, & D. Walter (Éds.), *Proceedings of the 31st Second Language Research Forum*. Cascadilla Press.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Dissertation, Georgia State University]. https://scholarworks.gsu.edu/alesl_diss/35

R Core Team. (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Riemenschneider, A., Weiss, Z., Schröter, P., & Meurers, D. (2021). Linguistic complexity in teachers' assessment of German essays in high stakes testing. *Assessing Writing*, *50*, 100561. https://doi.org/10.1016/j.asw.2021.100561

Ripley, B. (2022). *Nnet* (7.3) [R]. https://www.rdocumentation.org/packages/nnet/versions/7.3-12/topics/multinom

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018). Second Language Acquisition Modeling. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 5665. http://aclweb.org/anthology/W18-0506

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 42904297. https://aclanthology.org/L16-1680

Vajjala, S., & Rama, T. (2018). Experiments with Universal CEFR Classification. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 147153. https://doi.org/10.18653/v1/W18-0515

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In D. Lin, Y. Matsumoto, & R. Mihalcea (Éds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (p. 180189). Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2002472.2002496

## Developmental use of prenominal noun modifiers by Spanish L1 EFL teachers

Roger W. Gee[1], M. Karen Jogan[2], Kathleen S. Jogan[3]
Holy Family University[1], Albright College[2], University of Arkansas[3]
rgee@holyfamily.edu[1], kjogan@albright.edu[2], kjogan@uark.edu[3]

The paper presented in this session gives the results of a corpus-driven study of prenominal noun modifiers (PNMs) in a corpus of essays written by Spanish L1 EFL teachers. Appropriate and inappropriate PNMs were identified, and an error analysis was conducted with inappropriate PNMs to determine cross-linguistic influence (CLI). Error analysis, popular in the 1970s, has seen renewed interest with the advent of learner corpora (Diez-Bedmar 2022).

PNMs are an important feature of academic English. Hyland & Jiang (2021) point out that "noun-noun sequences … have increased dramatically … and represent the main noun phrase pattern" (7). However, Biber et al. (2011) proposed a developmental sequence in which PNMs develop at a late stage. Also, previous research has found that EFL speakers whose L1 is Spanish, a language that does not allow PNMs, underuse PNMs when compared to L1 speakers of languages that allow PNMs (Parkinson 2015) and have more difficulties with comprehension (Priven 2020). The purpose of the current study was to investigate the use of PNMs by teachers, presumably advanced users of English.

Three research questions were addressed:

1. Do the number of appropriate PNMs vary across advanced proficiency ranges?
2. Do the number of inappropriate PNMs vary across advanced proficiency ranges?
3. Do inappropriate PNMs show evidence of CLI?

The corpus contained 48 essays about teaching English during the pandemic with 48,187 words. It was tagged using TagAnt (Anthony, 2015). The essays were divided into three groups, low-, medium-, and high-advanced based on the use of vocabulary beyond the 3000-word level of the BNC/COCA wordlists (Nation 2012) using AntProfiler (Anthony 2021), creating a cross-sectional design that allowed developmental insights. Following Gotz (2022), concordance lines containing the PNMs were extracted with AntConc (Anthony, 2017) to provide context for "the highly context-dependent nature of what constitutes an error" (Thewissen 2020: 05). Analysis found 490 PNMs, which were normalized with the low group producing 8.44 PNMs per 100 words, the medium group 22.43, and the high group 21.31.

Inappropriate PNMs were coded for type of error and evidence of CLI through a unanimous agreement among the three researchers, with 81 PNMs determined to be inappropriate. CLI was operationalized as occurring when it was possible "to describe and/or explain L2 performance that appears to be influenced by, draws on, or uses some type of prior language knowledge (e.g., L1) in the learning and use of a new language" (McManus 2022: 24). Types of errors included number agreement (*adults learners*), translation (*necessity students*), and reversal (*video music*). It was determined that 48 PMIs (59%) of the inappropriate PMNs evidenced some form of CLI.

As the data was not normal, it was analyzed using two Kruskal Wallis tests. The first test showed significant differences among the three groups, $H$ (2) = 9.857, $p$ =.007 for appropriate PNMs. Dunn's test for pairwise comparisons using a Bonferroni correction found significant differences between the low and medium groups and the low and high groups ($p$ =.005 and $p$ = .014 respectively) but no difference between the medium and high groups ($p$ = .360). The second Kruskal Wallis test for inappropriate PNMs found no significant differences among the three groups, $H$ (2) = 2.148, $p$ = .342.

The answers to the research questions are

1. appropriate use PNMs varies across advanced proficiency ranges suggesting continuing development
2. the number of inappropriate PNMs does not vary across proficiency ranges
3. inappropriate PNMs show evidence of CLI.

These findings support Biber et al.'s (2011) assertion that PNMs develop at a late stage as they are still developmental at an advanced proficiency level. Furthermore, the number of PNMs per 1,000 words found in this study, 10.16, is far below the approximately 67 PNMs per 1,000 words in the humanities reported by Biber and Gray (2016: 165, Figure 4.16), suggesting underuse and possible avoidance. These results, moreover, are consistent with those reported by Parkinson (2015) for high intermediate to advanced university

students (5.9-16.07/1,000) and more recently by Bychkovska (2021) for L2 first-year university writers (6.72-16.05/1,000).

Finally, CLI is seen even in the use of PNMs by advanced EFL learners, confirming Schilk (2021), who argued that "L1 influence is very likely to play a role in the production of L2 expressions" (212).

**References**


Anthony, L. (2015). *TagAnt* (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2017). *AntConc* (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2021). *AntWordProfiler* (Version 1.5.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1), pp. 5-35.

Biber, D. & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing.* Cambridge University Press.

Díez-Bedmar, M. B. (2022). Error Analysis. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora*, (2nd edition), (pp. 90-104). Routledge. [Kindle Edition]

Gotz, S. (2022). Analyzing a learner corpus with a concordancer. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora*, (2nd edition), (pp. 68-89). Routledge. [Kindle Edition]

Hyland, K., & Jiang, F. (2021). Academic naming: Changing patterns of noun use in research writing. *Journal of English Linguistics*. June 2021. Doi: 00754242211019080

McManus, K. (2022). *Crosslinguistic influence and second language learning.* New York: Routledge.

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved from https://www.laurenceanthony.net/software/antwordprofiler/

Parkinson, J. (2015). Noun–noun collocations in learner writing. *Journal of English for Academic Purposes*, *20*, pp. 103-113.

Priven, D. (2020). "All these nouns together just don't make sense!": An Investigation of EAP Students' Challenges with Complex Noun Phrases in First-Year College-Level Textbooks. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée*, *23*(1), 93-116. https://doi.org/10.37213/cjal.2020.28700

Schlik, M. (2021). Tracing collocation in learner production and processing: Integrating corpus linguistic and experimental approaches. In S. Granger (Ed.), *Perspectives on the L2 phrasicon: The view from learner corpora*, (pp. 206-231). Multilingual Matters.

Thewissen, J. (2020). Accuracy. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 305-317). Routledge. [Kindle Edition]

## Students' requestive emails to faculty-pragmatic proficiency in elicited and spontaneous Italian L1 and English L2

Sara Gesuato[1], Elisabetta Pavan[2]
University of Padua[1,2], University Ca' Foscari University of Venice[2]
sara.gesuato@unipd.it[1], elisabetta.pavan.1@unipd.it[2], epavan@unive.it[2]

Studies on SL/FL pragmatic skills often describe L2 discourse as not as effective/appropriate as L1 discourse, which appears to be (more) adequate because produced by competent speakers sharing expectations about interactional norms. But communicative expertise is shaped by socialization practices: since these differ from one individual to the next, different speakers will display different levels of effectiveness and appropriateness, whether native or non-native speakers. Thus, a person's degree of discursive refinement cannot be taken for granted either in the L1 or the L2. A case in point is that of email discourse. The literature tends to present L2 email writing as a deviation from the L1 "norm" (e.g. Alcón-Soler 2015, Biesenbach-Lucas 2007, Chen 2006, Economidou-Kogetsidis 2011, Economidou-Kogetsidis et al. 2021, Garrote & Ainciburu 2020, Codina-Espurz & Salazar Campillo 2019, Hartford & Bardovi-Harlig 1996). Yet, our impressionistic observations as faculty recipients of student email requests suggested that both L1 and L2 written discourse is not always effective or appropriate. We thus decided to investigate whether and to what extent comparable L1 and L2 requestive texts might exhibit comparable types of linguistic-communicative inadequacies.

We examined both spontaneously produced and elicited email messages written by Italian students of English. First, through two Written Discourse Completion Tasks, we elicited from 60 student volunteers two requestive email messages to faculty, one in English L2 and one in Italian L1 (120 texts, 9,002 words). Then we collected exchange-initiating email requests received from our students (students of English or Linguistics) over a four-month period, and we each selected the first 30 Italian and the first 30 English requestive messages we received from Italian students (120 texts, 8,800 words), deleting all personal identifying information from them. We assessed the texts in terms of structure and interaction management (e.g. opening, closing, subject heading), content (amount, intelligibility, and relevance of information), requestive strategies (e.g. legitimacy and cost of the request), and form (e.g. paragraphing and accuracy). We coded all values as binary, except for accuracy, which we coded on a three-point scale (inter-coder agreement: 96%).

Three main findings emerged: A) We coded the various features as adequate in a majority of the texts in the four datasets, contrary to our expectations; B) we coded a higher number of features as adequate in the L1 than the L2 texts in the four domains considered, in line with our expectations; but C) in some specific formal and structural features (e.g. use of the sender's institutional email address, informative subject headings, paragraphing) a higher number of the L2 texts appeared to outperform the L1 texts, unlike what we expected.

Since the various features considered were not coded as adequate in all the L1 texts, we conclude that native speaker status is not a fully reliable predictor of effective/appropriate communication. Thus, L1 writers, too, might be considered learners in their own language in specific discursive domains where face wants are addressed. We suggest that not only in L2 but also in L1 language education, students should be alerted to the key determinants of communicative acceptability (i.e. addressee-friendliness, face enhancement) and effectiveness (i.e. Grice's Cooperative Principle), which affect how their discourse is perceived and responded to (Hartford, Bardovi-Harlig 1996).

**References**

Alcon, E. (2015). Teachers' perceptions of email requests: insights for teaching pragmatics in study abroad contexts. In S. Gesuato, Bianchi & W. Cheng (eds.). *Teaching, learning and investigating pragmatics: principles, methods and practices*. Newcastle upon Tyne: Cambridge Scholars Publishing, 13-31.

Biesenbach-Lucas, S. (2007). Students writing emails to faculty: an examination of e-politeness among native and non-native speakers of English. *Language Learning & Technology*, 11(2). 59-81.

Chen, C.-F. E. (2006). The development of e-mail literacy: from writing to peers to writing to authority figures. *Language Learning & Technology*, 10(2), 35-55.

Codina-Espurz, V., Salazar-Campillo P. (2019). Student-to-faculty email consultation in English, Spanish and Catalan in an academic context. In P. Salazar-Campillo & V. Codina-Espurz (Eds.). *Investigating the learning of pragmatics across ages and contexts*. Leden-Boston: Brill, Rodopi, 196-217.

Economidou-Kogetsidis, M. (2011). 'Please answer me as soon as possible': pragmatic failure in non-native speakers' e-mail requests to faculty, *Journal of Pragmatics.* 43(13), 3193-3215.
Economidou-Kogetsidis, M., Savic, M. & Halenko, N. (2021). *Email pragmatics and second language learners*. Amsterdam/Philadelphia: John Benjamins.
Garrote, P. R., Aicinburu, M. C. (2020). Analysis of requests from Italian-speaking students of English and Spanish: reflection on pragmatic strategies in multilingual competence. *Lingue e Linguaggi*, 36, 267-281.
Hartford, B. S., Bardovi-Harlig, K. (1996). At your earliest convenience: a study of student email to faculty. In L. B. Bouton (Ed.). *Pragmatics and Language Learning*, Monograph Series Volume 7, 55-69.

## Syntactic variation in German *weil*-clauses: A comparison between immersed and non-immersed learners of German

Aivars Glaznieks[1], Jennifer-Carmen Frey[2]
Institute for Applied Linguistics, Eurac Research
aivars.glaznieks@eurac.edu[1], jennifer.frey@eurac.edu[2]

Register-related variation has been identified as a challenge for young writers in general and especially for second or foreign language learners (cf. Stollhans 2020). Contemporary German, for example, shows a well-described register-dependent syntactic variation between verb-second and verb-final word order in clauses introduced by the conjunction *weil* 'because'. As a subordinating conjunction, *weil* requires verb-final word order typical for dependent clauses in German. However, in spoken language, *weil* is also used as a coordinating conjunction with verb-second main clause word order. This might lead to controversial inputs for language learners depending on their language environment and exposure to the target language.

In our study, we analyzed *weil*-clauses in texts of young learners with various language backgrounds and exposures within the German sub-corpus of the longitudinal learner corpus LEONIDE (Glaznieks et al. 2022). We focused on two groups of non-native learners from lower secondary schools which differ with respect to the learner setting, i.e., the school's main language of instruction and the status of German in the schools. Pupils of group (A) attended schools in which Italian is the main language of instruction and German is taught as the first additional language from grade one onwards. Pupils of group (B) attended schools in which German is the main language of instruction and German lessons follow native language instruction. While both groups must master word order differences between dependent and independent clauses, we assume different preconditions for learning the target word order for (A) and (B) (cf. Grießhaber 2010). To investigate the impact of the learning settings on both groups, we compare them with each other and to a third group (C) of native speakers of German, also attending schools with German native language instruction (L1 reference data). All pupils reside in the same multilingual (Italian, German) region and are comparable with respect to age (12-14) and grade (6-8).

In our analysis we address the following research questions:

- RQ1: Is there a difference between the interlanguage varieties of the two groups (A) and (B) with respect to word order in *weil*-clauses and how do they compare to the reference group (C)?
- RQ2: Does the proportional use of the target structure change over time and does the development differ between group (A) and (B)?

To address these research questions, we analyzed all occurrences of *weil* in the corpus with respect to syntactic (e.g., word order, grammatical errors) and semantic features (using Sweetser's (1990) levels of semantic/pragmatic relations). We analyzed the overall use of verb-final vs. non-verb final structures as well as developmental trends in the three groups quantitatively with generalized linear mixed-effects models, investigating the main effects of time and group as well as their interaction. We investigated non-verb-final structures for all groups qualitatively, analyzing formal and pragmatical/semantical features of these structures and their distribution for each group, before we identified some linguistic features of the *weil*-clauses that are predictive for the use of verb-final or non-verb-final structures within and over the three groups, such as the presence of additionally integrated subclauses or the semantics of the causal relation. Furthermore, we investigated the same aspects for a second subordinating conjunction (*wenn* 'when, if') which does not show a register-specific syntactic variation, allowing us to separate pragmatical from purely formal learning goals for non-native writers.

Finally, we extended the developmental perspective with cross-sectional data from upper secondary schools, using comparable German non-native speakers' texts from the Kolipsi-2 corpus (non-immersed learners, Frey et al. in preparation) and the KoKo corpus (immersed learners, Abel et al. 2014). Additional information on all corpora can be found at www.porta.eurac.edu. Our results show that group (A) has a significantly lower rate of target-like structures and a high rate of verb-second word order next to diverse clear learner errors in non-target-like clauses. Target structures increase significantly with time. Group (B) produces a higher percentage of target structures than non-target structures. The latter are exclusively verb-second clauses; however, the proportional use of the verb-final target structure does not significantly increase over time. The results suggest that the overall use and developmental paths of the group (B) are more similar to group (C) than to group (A). We will discuss the results in the light of the different learning settings of the groups.

**References**

Abel, A., Glaznieks, A., Nicolas, L. & Stemle, E. (2014). KoKo: An L1 Learner Corpus for German. Calzolari, Nicoletta et al. (Eds.). *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik: European Language Resource Association, 2414-2421.

Frey, J.-C., Glaznieks, A., Nicolas, L., Abel, A. & Vettori, C. (in preparation). *The Kolipsi Corpus Family. Resources for learner corpus research in Italian and German*.

Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L. & Nicolas, L. (2022). LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, 8(1), 97-120.

Grießhaber, W. (2010). *Spracherwerbsprozesse in Erst- und Zweitsprache. Eine Einführung*. Duisburg: Universitätsverlag Rhein-Ruhr.

Stollhans, S. (2020). Linguistic variation in language learning classrooms: considering the role of regional variation and 'non-standard' varieties. *Languages, Society and Policy*. https://doi.org/10.17863/CAM.62274

Sweetser, E. (1990). *From etymology to pragmatics. Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

## Most dispersion measures do not measure dispersion, and the implications of that for LCR

Stefan Th. Gries
UC Santa Barbara & JLU Giessen
stgries@linguistics.ucsb.edu

The two most widely-used corpus statistics by far are probably frequencies (of occurrence and co-occurrence) and association measures (such as MI and log-likelihood). However, over the last ten years or so, a variety of publications have also made a case for a more widespread adoption/use of dispersion measures, i.e. measures that quantify the degree to which (typically) words are distributed evenly or 'clumpily' in a corpus. While I agree with the notion that dispersion information is important, in this paper, I will do four things.

First, I will argue that nearly all dispersion measures that are currently used do in fact not measure dispersion well, at least not if, as I think it should be, dispersion is considered a separate dimension of information. Instead, they merely repackage or tweak, frequency information. To support this seemingly bold/counterintuitive claim, I will discuss results supporting it on the basis of twelve dispersion measures (including Juilland's *D*, Rosengren's *S*, *KLD*, *IDF*, *DP*/$DP_{\text{norm}}$, range) applied to six corpora (including the BNC, the BNCspoken, Brown, and the ICE-GB) that show two things:

- most dispersion measures are 0.9 correlated with the frequency of occurrence based on $R^2$s of generalized additive models (used to capture more than straight-line correlations);
- if frequency and dispersion measures were really measuring different constructs independently of each other, it should be possible to identify words with high and low frequencies with both high and low dispersions, but the way dispersion measures are computed practically rules out findings words that are of low frequency and even dispersion.

Second, I will outline how we can measure dispersion in a way that is truly independent of frequency. I will first use a straightforward example to motivate the proposed way of measuring (two specific words in the Brown corpus), then I will discuss how the measure can be computed and how its computation makes it independent of frequency, and then I will apply it to the same six corpora on which the traditional measures were tested to show how much less than the traditional measures it is correlated with frequency.

Third, I will validate the measure on the basis of psycholinguistic data from the Massive Auditory Lexical Decision (MALD) database. When the new measure (as applied to the six corpora) is compared to existing ones in terms of how well it, together with (logged) frequency as a second dimension, predicts lexical decision times (between 68K and 112K tokens, depending on the corpus used), for five out of six corpora, it beats all other measures' predictive power.

Fourth, these findings have important implications for learner corpus research. On a very general level, this is because the application of the new dispersion measure completely changes our too frequency-biased understanding of what is how widespread in native speaker data which often function as a reference to which learner data are related (e.g. in over-/under-use studies or for the compilation of general academic word lists). On a more particular level, the new measure can also help improve specific measures relevant to SLA research such as measures of lexical sophistication. Traditionally, some measures of lexical sophistication involve the number of files/documents a word is attested in, which produces results that are, again, just repackaged frequency results and do not help us to see how 'sophisticated' words are based on their dispersion in native speaker data.

On the basis of the above results, I will argue that this new measure should be used instead of the traditional ones (at least when 'word commonness' or its counterpart 'specialness/sophistication' is what is being studied) and I will briefly discuss an additional example of how this measure can also augment collocation/association statistics.

## Association measures in learner corpus research: Problems and pointers for improvement

Stefan Th. Gries[1], Magali Paquot[2]
UC Santa Barbara & JLU Giessen[1], Université catholique de Louvain[2]
stgries@linguistics.ucsb.edu[1], magali.paquot@uclouvain.be[2]

Much research in learner language involves studying the degree to which frequencies of use of some linguistic element or its preferred co-occurrences with other elements by learners differ (i) from those of that same element by native speakers or (ii) over time as learners become more proficient. Much of the work involving co-occurrence (association) in particular involves association measures (AMs) that aim to quantify how much words are associated with other words (collocations) or with other, more schematic constructions (e.g. collostructions). Learner corpus studies often use measures such as *MI* and/or *t* score (Gablasova et al., 2017; Forsberg-Lundell, 2021); it is also not infrequent that they rely on cut-off points to identify important collocations (e.g. Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Siyanova-Chanturia, 2015). In this paper, we will discuss several problems we see in much existing work.

First, we will show that much work suffers from a very elementary – in a sense – validity problem because the association measures that are used really reflect co-occurrence frequency more than they do association proper. On the basis of hypothetical collostructional data, observed keyness statistics from the Clinton-Trump corpora, and actual Adj-N collocations (involving four-speed adjectives in the BNC), we show that, in monofactorial tests, esp. the loglikelihood ratio and *t* are extremely predictable from just the co-occurrence frequency alone ($R^2_{GAM}$>0.94) and are hardly correlated with what one might consider a gold standard measure of association (the log odds ratio, $R^2_{GAM}$<0.06). In addition, we will exemplify how the characteristics of some measures (esp. *MI*) are not always properly understood or at least discussed in the literature.

Second, if one tries to tease apart the contribution that co-occurrence frequency and association make together, i.e. in a multifactorial setting, one finds that the loglikelihood ratio and *t* can be nearly predicted perfectly from an interaction of logged frequency and association (the log odds ratio), but with frequency playing a much stronger role than the association. Crucially, that very high correlation of frequency and association also means that these AMs do not permit the user to properly address the separate contributions that association (i.e., contingency) and frequency (i.e. entrenchment) make cognitively or psycholinguistically (Gries & Ellis, 2015): if one's AM conflates contingency and frequency, one can be definition not separate high frequency from high association in one's theoretical explanations; we, therefore, recommend a two-dimensional representation of frequency and association for any kind of association measure scenario.

The question that arises from both of these related issues is how to decouple frequency and association in our studies involving AMs. In other words, how do we make sure that our AMs really reflect association and only association so that we can test and/or develop our theories in such a way that frequency and association make measurable but separate components that our hopefully cognitively-informed SLA models can handle? In this paper, we will present an answer to this question by outlining a three-step procedure of how one can take any AM and completely decouple it from frequency.

First, we quantify the association for a certain collocation in the data (let's call this value *obs*). Second, we take the frequencies of the two co-occurring elements in question (i.e. the totals $a+b$ and $a+c$) and the corpus size (i.e. $a+b+c+d$), hold them constant (which virtually eliminates any way in which co-occurrence frequency can unduly boost/lower the resulting association-only measure), and determine (i) the lowest and the highest possible associations given the values we are holding constant (let's call these *low* and *upp* for lower and upper limit). Third and because these maximal-attraction and minimal-attraction/maximal-repulsion values will exhibit different ranges (due to the marginal totals), we then transform/min-max scale these three values (*obs*, *low*, and *upp*) such that they fit into the interval [0,1], and our new association-without-frequency measure becomes the value that corresponds to *obs* in that [0,1] interval. We exemplify this approach by applying it to one specific AM – conditional probability – for co-occurrences of speed adjectives with nouns in the BNC and show that it is indeed uncorrelated with frequency ($R^2$ with co-occurrence frequency < 0.01) but also supports the use of *MI* and the log odds ratio as true measures of association only. We conclude with some comments on what the results mean for the use of cut-off points for significant/interesting associations.

**References**

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

Forsberg-Lundell, F. (2021). Formulaicity. In Tracy-Ventura, N. & Paquot, M. (Eds). *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 370-381). Routledge.

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, 67 (51): 155-179.

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 in Language Teaching* 52: 229-252.

Gries, S. Th. & Ellis, N. (2015). Statistical measures for usage-based linguistics. *Language Learning* 65 (Supplement 1): 1-28.

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System* 53: 148-160.

# Code glosses in L2 learner writing: Reformulation and exemplification in master's theses by Czech university students

Tereza Guziurová
University of Ostrava
tereza.guziurova@osu.cz

For over thirty years, metadiscourse has been used as an important analytical framework for investigating writer-reader interaction in academic genres. In his recent appraisal of the concept, Hyland (2017: 16) argues that metadiscourse is perhaps now "one of the most commonly employed methods for approaching specialist written texts". Following a large body of research on metadiscourse, this paper explores one feature of textual metadiscourse, code glosses, in English L2 academic texts written by Czech university students. The study draws on Hyland's metadiscourse model (2005), which characterizes code glosses as devices that elaborate propositional meanings by rephrasing or explaining what has been said, including phrases such as *in other words, that is, this can be defined as, for example*. They can help readers understand the writer's intended meaning or contribute to the formation of persuasive arguments. Hyland (2007) distinguished two broad subfunctions of code glosses, reformulation and exemplification. Reformulation promotes textual cohesion and facilitates discursive progression as it provides "a retrogressive interpretation of the previous utterance and allows speakers to explain, rephrase, reconsider, summarize or even distance themselves from it" (Dal Negro & Fiorentini 2014: 95). Exemplification is a "process through which meaning is clarified or supported by a second unit which illustrates the first by citing an example" (Hyland 2007: 270).

The main aims of the paper are:

(1) to compare the use of code glosses in L2 Master's theses written by Czech university students and in L1 research articles in three soft science disciplines, namely linguistics, literature and English language teaching (ELT) methodology;
(2) to find out whether there are cross-disciplinary differences in the use of code glosses in L2 Master's theses.

The corpus consists of 48 English L2 Master's theses representing three disciplines – linguistics (16 theses), literature (16 theses) and ELT methodology (16 theses), totalling almost 950,000 words. The authors are postgraduate students majoring in English language and literature at Masaryk University in Brno, Czech Republic, and their L1 is Czech. The results are compared with professional writing represented by English L1 research articles from the same disciplines. The second corpus consists of 36 research articles (12 linguistic, 12 literary and 12 methodology RAs), totalling 243,000 words.

The results have shown that the general frequency of code glosses was higher in the Master's theses than in the research articles (370.6 to 330.4 occurrences per 100,000 words, respectively). In both corpora, exemplification predominated over reformulation, which corresponds to Hyland's findings that exemplification plays a significant role in 'soft' disciplines as it represents a "heavier rhetorical investment in contextualization" (Hyland 2007: 272). The overall frequency of exemplification markers was similar in both corpora, so it was the function of reformulation which represented the biggest difference. The paper discusses different functions of reformulation markers in detail. The results have also revealed cross-disciplinary variation, as reformulation and exemplification proved to be much more prominent in linguistics and methodology than in literature, irrespective of the genre. This suggests that literature, which traditionally belongs to humanities, has different rhetorical conventions and modes of argumentation than the other two disciplines.

The findings have shown that novice writers recognise the importance of reformulation and exemplification in their argumentative practices since they use code glosses frequently. Indeed, quantitative comparisons indicate an overuse of certain reformulation markers in their theses. This may be given by the character of the genre, which requires that they demonstrate knowledge and understanding of the theories, methods and nomenclature of a given discipline. However, in comparison with the L1 writers, the Czech students overused two expressions, *i.e.* and *such as*, irrespective of a discipline. The results proved to be statistically significant, using log-likelihood tests (Rayson 2008). This suggests a certain tendency to simplification since the students relied on simple, grammaticalized forms, which do not require much processing effort.

**References**

Dal Negro, S., & Fiorentini, I. (2014). Reformulation in bilingual speech: Italian *cioè* in German and Ladin. *Journal of Pragmatics*, 74, 94-108.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing.* London and New York: Continuum.

Hyland, K. (2007). Applying a gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics*, 28(2), 266-285.

Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics*, 113, 16-29.

Rayson, P. (2008). *Log-likelihood and effect size calculator*. [Accessed 19 April 2022]. Available at: http://ucrel.lancs.ac.uk/llwizard.html

# Visual Thinking Strategies (VTS) in online EFL learner discussions: Creating a micro-corpus of spoken learner discourse for qualitative analysis

Sharon Hartle, Giorgia Andreolli, Emanuela Tenca
Università degli Studi di Verona
{sharon.hartle, giorgia.andreolli, emanuela.tenca}@univr.it

This presentation focuses on a micro-corpus of spoken language created as part of a pilot study that was conducted in 2021 at the Department of Foreign Languages and Literatures (University of Verona). The corpus was created following the trial of Visual Thinking Strategies (VTS) for the design of teaching activities and materials which aim to develop fluency and critical thinking skills in EFL learners. The project in its entirety focuses especially on tasks and materials that enable learner expression, the analysis of key emergent language, and its successive integration into the learning design. Specifically, this stage of our study seeks to explore learner discourse by means of the following questions:

1. How are speaker contributions distributed?
2. Do specific VTS questions lead to the use of different critical thinking patterns?

Our sample consisted of 22 university students, who were English users from different lingua-cultural backgrounds, with language levels ranging from B2 to C2. The study was conducted in two sessions with 11 participants in each, who were recruited by convenience sampling. Participant spoken discussions, held in breakout rooms on Zoom, were carried out at two separate stages of the materials trialling process. These were recorded, transcribed, and coded using Qualitative Data Analysis Software (QDAS). The data were then divided into two sub-corpora, reflecting these stages: the first one, which was a semi-structured discussion, and the second a freer one, and the texts used to compile a local corpus (Gilquin, 2015:15).

The Visual Thinking Strategies (VTS) approach is an inquiry-based, instructional method which originated in the field of museum education (Yenawine, 2013) and has attracted interest in a range of fields such as language education (Bomgaars and Bachelors, 2020), teacher training (Smolkowski et al., 2020), and healthcare (Agarwal et al., 2020). This approach was selected for two main reasons. Firstly, VTS focuses on the learner, and, secondly, in its basic format, it enables learners to create meanings from largely non-verbal sources such as images and art, enabling the language produced to be influenced but not constrained by the input. Anchored in the subjective exploration of images, VTS has been seen to have a positive effect on the development of speaking and writing skills and critical thinking (Housen & De Santis, 2009).

Starting from the exploration of an image or piece of art, semi-structured learner discussion is developed in groups, in which participants are guided by three specifically worded questions. Initially, we applied a purist VTS approach using a static image, and then, at a later stage, further questions were provided. Only one of these was selected by the learners themselves to foster both agency and the discussion of emergent themes in greater depth.

Our aim was to investigate patterns in learner discourse, which might be produced following the specific question types used in the VTS approach. This forms a snapshot of the discourse generated during the two specific stages of our learning design, which enabled us to analyze the task types being applied. In spite of its size, a small corpus of this kind allows us to identify a concentrated picture of language features being used in the very specific context of an online discussion between EFL learners. Such specific patterns, in fact, may not emerge in larger corpora (O'Keeffe et al., 2007: 182). The data collected were qualitatively explored through the lens of the Critical Thinking (CT) Framework developed by Dwyer et al. (2014). This framework seemed particularly suited to our purposes given the role played by CT skills in the cognitive processing of complex information, fostering creative problem solving, and increasing the number of ideas and depth of argumentation. The results obtained so far suggest that the questions may in fact relate to the type of critical skills discourse patterns that learners produce in our context in response to the task and question type. There appears to be a progression from description or narrative, where learners simply list what they see, to inference where a deductive process comes into play and evidence is given for the conclusions reached. In the second discussion, there was greater evidence of analytical thinking where evidence was examined in more depth and finally, learners evaluated their own and their peers' insights extending them to wider contexts.

**References**


Agarwal, G.G., McNulty, M., Santiago, K.M. et al. (2020). Impact of Visual Thinking Strategies (VTS) on the Analysis of Clinical Images: A Pre-Post Study of VTS in First-Year Medical Students. *Journal of Medical Humanities*, 41, 561-572.

Bomgaars, J. & Bachelor, J. (2020). Visual Thinking Strategies: Exploring Artwork to Improve Output in the L2 Classroom. *Journal of Foreign Language Education and Technology*, 5(1), 1-34.

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An Integrated Critical Thinking Framework for the 21st Century. *Thinking Skills and Creativity*, 12, 43-52.

Gilquin, G. (2015). From Design to Collection of Learner Corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The Cambridge Handbook of Learner Corpus Research*, Cambridge: CUP, 9-34.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP.

Housen, A. & De Santis, K. (2009). *VTS Visual Thinking Strategies. A Brief Guide to Developmental Theory*. New York: Visual Understanding in Education.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: CUP.

Smolkowski, K., Stryker, L.A., Anderson, L. et al. (2020). The Visual Thinking Strategies Approach to Teaching Argument Writing. A Professional Development Model. *The Elementary School Journal*, 121(1), 100-124.

Yenawine, P. (2013). *Visual Thinking Strategies: Using Art to Design Learning Across School Disciplines*. Cambridge, MA: Harvard Education Press.

## Young writers' use of adverbial intensification in English L1 and L2

Hilde Hasselgård
University of Oslo
hilde.hasselgard@ilos.uio.no

This study concerns the adverbial modification of adjectives in narrative texts by young writers. The material comprises English L2 (EL2) writing in lower secondary school in Norway, culled from the TRAWL corpus, and English L1 (EL1) writing by similarly aged pupils from the Growth in Grammar (GiG) corpus. From TRAWL, narrative texts from years 8 and 9 were selected (age 13-15), and from GiG, literary texts from year 9 (age 13-14). The study aims to explore similarities and differences between Norwegian EL2 writers and British EL1 writers, guided by the following questions:

- To what extent do the young writers use the adverbial modification of adjectives?
- Which adverbs are used for adjective modification in the respective writer groups?
- In what syntactic environments do they use adverb-adjective sequences (attributive/ predicative) and what meanings are expressed by the modifying adverbs (e.g. downtoner, amplifier, descriptor)?

Since the EL2 material represents the same pupils over two years, I will also look for signs of development in the use of adverb-adjective combinations from year 8 to year 9.

Adverbial intensification has been much studied in L1/L2 contexts, but primarily in advanced learner writing. Findings include a tendency to overuse adverbial intensification (e.g. Lorenz 1998) as well as individual intensifiers, especially *very* (Granger 1998; de Haan & van der Haagen (2013). Hasselgren (1994) and Pérez-Paredes & Sánchez-Tornel (2014) focus on young learners. Hasselgren finds that learners use a more limited lexical repertoire than native speakers, leading to a smaller set of 'core' adverbs being used at the expense of more specific words. Pérez-Paredes & Sánchez-Tornel track a longitudinal development in the use of adverbs, identifying a richer use around year 10.

Investigating intensifiers in dialogic and narrative parts of fictional texts, Ebeling & Hasselgård (2020) find that the same intensifiers are frequent in both subregisters: *very>so>too* in dialogue and *so>too>very* in the narrative. As for spoken English, Aijmer (2018) and Tagliamonte (2008) find that *very* is the most frequent intensifier in British English, while *really* is more frequent in American English. These studies are relevant because the young writers' patterns of intensifiers may resemble that of speech (cf. Gilquin & Paquot 2008) and display a low degree of formality.

The analysis indicates that adverb-adjective combinations are more frequent and widespread in EL2 than in EL1 writing. All the pupils greatly favour amplifiers over downtoners, but this preference is stronger in EL2 than in EL1. As expected, the EL2 pupils use a smaller set of adverbs than the EL1 pupils. In both corpora, the most frequent amplifiers are *so* and *very*, illustrated in (1). These account for 60-70% of the adverbs in TRAWL and c. 35% of those in GiG. *Too* is more frequent in GiG than in TRAWL. Moreover, the EL2 writers do not display the kind of descriptive and creative adverb-adjective combinations occasionally found in GiG and exemplified in (2).

(1) The insects were *so good* they were salty but *very sweet*. (TRAWL_P60106_Y09)
(2) With *spine chillingly cold* corridors the place was scary. (GiG_4_236.txt)

The adverbs in the EL2 material seem to become more varied from year 8 to year 9, including a wider range of adverbs and a lower proportion of the tokens comprising the three most frequent types. Unfortunately, the TRAWL subset used here contains too few narrative texts from year 10 to track any further lexical and phraseological development (cf. Pérez-Paredes & Sánchez-Tornel (2014). A conclusion is that the young EL2 writers are rather proficient users of adverb-adjective combinations, although they generally stick to core vocabulary.

**References**


Aijmer, K. (2018). Intensification with *very*, *really* and *so* in selected varieties of English. In S. Hoffmann, A. Sand, S. Arndt-Lappe & L. M. Dillmann (eds), *Corpora and Lexis* (pp. 106–139). Brill Rodopi. https://doi.org/10.1163/9789004361133_006

Ebeling, S. O. & Hasselgård, H. (2020). Intensification in dialogue vs. narrative in a corpus of present-day English fiction. In E. Jonsson & T. Larsson (eds), *Voices Past and Present - Studies of Involved, Speech-related and Spoken Texts. In honor of Merja Kytö* (pp. 302–316). John Benjamins.

Gilquin, G. & Paquot, M. (2008). Too chatty. Learner academic writing and register variation. *English Text Construction* 1:1, 41–61.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A.P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications* (pp. 145–160). Oxford University Press.

De Haan, P. & van der Hagen, M. (2013). The search for sophisticated language in advanced EFL writing. In S. Granger, G. Gilquin & F. Meunier (eds), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead* (pp. 103–116). Presses Universitaires de Louvain.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4: 237–259.

Hasselgård, H. (2015). Lexicogrammatical features of adverbs in advanced learner English. *ITL: International Journal of Applied Linguistics*, 166: 1, 163–189.

Lorenz, G. (1998). Overstatement in advanced learners' writing: stylistic aspects of adjective intensification. In S. Granger (ed.), *Learner English on Computer*, (pp. 53–66). Longman.

Pérez-Paredes, P. & Sánchez-Tornel, M. (2014). Adverb use and language proficiency in young learners' writing. *International Journal of Corpus Linguistics* 19:2, 178–200.

Tagliamonte, S. (2008). *So different* and *pretty cool*! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics* 12 (2), 361–394.


**Corpora**


*Growth in Grammar*.

http://socialsciences.exeter.ac.uk/education/research/centres/writing/projects/growthingrammar/corpus/

TRAWL: https://www.hf.uio.no/ilos/english/research/groups/trawl-tracking-written-learner-language/

# Genres in young learner EFL writing: A genre typology for the TRAWL (tracking written learner language) corpus

Ingrid Kristine Hasund
University of Agder
kristine.hasund@uia.no

In learner corpus research, it is well known that one should control for the genre, as genre is one factor which has been shown to account for language variation (Ädel 2006, 2008; Aijmer 2002; Gilquin & Paquot 2008; Hasselgård 2009; Melissourgou & Frantzi 2017; Paquot et al. 2013; Petch-Tyson 1998; Recski 2004; Ørevik 2019). The concept *genre*, however, is fuzzy and has been used differently by different researchers (Biber 1988; Halliday & Matthiessen 2014; Melissourgou & Frantzi 2017; Paltridge 2002; Swales 1990). It is therefore necessary that genre categories are clearly defined, making it possible for researchers to compare "like with like" (Granger 2012: 12).

The studies above explore written EFL corpora from older learners. Few young learner corpora are openly available, and little research has been done on lower levels (cf. Dirdal 2021; Hasselgren & Sundet 2017). The present study contributes to filling this gap by exploring genres in EFL writing at the lower secondary level in Norway (ages 13-16), using data from TRAWL[1] (Tracking written learner language), a new longitudinal corpus currently under compilation. The author of the present study is part of the TRAWL research team and has worked on developing a genre typology for annotating one part of the corpus, called the *genre subcorpus*. It comprises all EFL texts written by one class from school year eight to ten, in a total of 327 texts (121,000 words) answering 56 writing prompts in several genres. As TRAWL will be openly available for research, it is important that the genre typology is clearly described, which is what the present study aims to do.

When published, TRAWL will be annotated with information about learners and texts, but only the genre subcorpus will have information about genre also. The reason is that most original texts contain answers in more than one genre. One mock exam answer, for instance, constituting one text unit in the corpus, may contain answers to three questions in three different genres. To create the genre subcorpus, most texts had to be split into smaller units separately annotated for the genre.

The genre categories were identified by studying the writing prompts, as recommended by Melissourgou and Frantzi (2017) and Ørevik (2019). Ørevik (2019), who studied EFL material for upper secondary level in Norway, presents a genre typology which categorises prompts in terms of *individual genres* and *main genres*. For example, the prompt "Write a story that takes place in a school" would be assigned the individual genre category *story*, which is part of the main genre *narratives* (ibid.: 105, 316).

As there exists no detailed genre typology for classifying learner texts at the Norwegian lower secondary level, Ørevik's typology for the upper secondary level was tested on the 56 TRAWL prompts and adapted to the lower secondary level. Two research questions were investigated, using a functional, social semiotic perspective (Berge et al. 2016; Martin 2009; Pilegaard & Frandsen 1996; Swales 1990) and a mixed-methods (quantitative and qualitative) approach:

1. Which individual genres and main genres are found in the writing prompts?
2. How do the findings compare to those from Ørevik's study, and which adaptations had to be applied to make the typology for upper secondary level suitable for lower secondary level?

Research question 1 was answered first through a qualitative study of the 56 prompts to assign one individual and one main genre category to each prompt. Then a quantitative analysis showed that all the six main genres from Ørevik were found in TRAWL, with expository and argumentative genres being the most frequently elicited, followed by narrative, descriptive, dialogic, and reflective genres. Regarding individual genres, only 13 of Ørevik's 34 genres were found in TRAWL, probably because learners are given fewer genre options at the lower secondary level. Furthermore, as some prompts did not elicit any specific genres, a category called *open* was created to signal that the learner texts should be checked manually for the genre. The answer to Research question 2 is that Ørevik's typology is largely suitable for classifying the TRAWL prompts, but two adaptations had to be

[1] https://www.hf.uio.no/ilos/forskning/grupper/trawl-%E2%80%93-tracking-written-learner-language/index.html

made: 1) reducing the number of individual genres to fit lower secondary level and 2) adding the *open* category for prompts that did not elicit any specific genres.

**References**


Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.24

Ädel, A. (2008). Involvement features in writing: do time and interaction trump register awareness? In G. Gilquin, S. Papp & M.B. Díez-Bedmar (Eds.). *Linking up Contrastive and Learner Corpus Research*. Amsterdam: Rodopi, 35-53. https://doi.org/10.1163/9789401206204_003

Aijmer, K. (2002). Modality in advanced learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (Eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 55-76. https://doi.org/10.1075/lllt.6.07aij

Berge, K. L., Evensen, L. S., & Thygesen, R. (2016). The Wheel of writing: A model of the writing domain for the teaching and assessing of writing as a key competency. *The Curriculum Journal*, 27(2), 172–189. https://doi.org/10.1080/09585176.2015.1129980

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: CUP.

Cope, B., & Kalantzis, M. (2014). Introduction: How a genre approach to literacy can transform the way writing is taught. In B. Cope, M. Kalantzis, A. Luke, C.B. Cazden & G. Kress (Eds.) The Powers of Literacy : *A Genre Approach to Teaching Writing.* London: Routledge, 1-21. https://doi.org/10.4324/9780203149812

Dirdal, H. (2021). L2 development of *-ing* clauses: A longitudinal study of Norwegian learners. In P. Pérez-Paredes & G. Mark (Eds.). *Beyond Concordance Lines: Corpora in Language Education*. Amsterdam: John Benjamins, 76-96. https://doi.org/10.1075/scl.102.04dir

Gilquin, G. & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction* 1(1), 41-61. https://doi.org/10.1075/etc.1.1.05gil

Granger, S. (2012). How to use foreign and second language learner corpora. In A. Mackey & S.M. Gass (Eds.). *Research Methods in Second Language Acquisition: A Practical Guide*. Oxford: Blackwell Publishing Ltd, 7-29.

Halliday, M. A. K. & Matthiessen, C. M. I. M. (2014*). Halliday's Introduction to Functional Grammar* (4th ed.). London: Routledge.

Hasselgren, A. & Sundet, K.T. (2017). Introducing the CORYL Corpus: What it is and how we can use it to shed light on learner language. *Bergen Language and Linguistics Studies,* 7. https://doi.org/10.15845/bells.v7i0.1107

Hasselgård, H. (2009). Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In K. Aijmer (Ed.). *Corpora and Language Teaching*. Amsterdam: John Benjamins, 121-140. https://doi.org/10.1075/scl.33.12has

Martin, J. R. (2009). Genre and language learning: A social semiotic perspective. *Linguistics and Education: An International Research Journal,* 20(1), 10-21. https://doi:10.1016/j.linged.2009.01.003

Melissourgou, M. N. & Frantzi, K. T. (2017). Genre identification based on SFL principles: The representation of text types and genres in English language teaching material. *Corpus Pragmatics*, 1(4), 373–392. https://doi.org/10.1007/s41701-017-0013-z

Paltridge, B. (2002). Genre, text type and the English for Academic Purposes (EAP) classroom. In A. M. Johns (Ed.). *Genre in the Classroom: Multiple Perspectives*. London: Routledge, 73-90.

Paquot, M., Hasselgård, H. & Ebeling, S.O. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.). *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Louvain-la-Neuve: Presses Universitaires de Louvain, 377-387.

Petch-Tyson, S. (1998). Reader/writer visibility in EFL persuasive writing. In S. Granger (Ed.), *Learner English on Computer*. London: Longman, 107-118. https://doi.org/10.4324/9781315841342-8

Pilegaard, M., & Frandsen, F. (1996). Text type. In J. Verschueren (Ed.), *Handbook of Pragmatics, Vol. 2*. Amsterdam: John Benjamins, 1-13.

Recski, L. J. (2004). Expressing standpoints in EFL written discourse. *Revista Virtual de Estudos da Linguagem – ReVEL*, 2(3), 16 pp. https://biblat.unam.mx/hevila/Revistavirtualdeestudodalinguagem/2004/vol2/no3/3.pdf

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: CUP.

Ørevik, S. (2019). *Mapping the text culture of the subject of English: Genres and text types in national exams and published learning materials*. PhD thesis. Bergen: University of Bergen. http://bora.uib.no/handle/1956/19266.

# Compiling a corpus of written and spoken L2 Chinese: Combining pragmatic and error annotation to study the Chinese *shì* 是…*de* 的cleft construction

Alessia Iurato
Università Ca' Foscari Venezia – Universität Bremen
alessia.iurato@unive.it

The Chinese *shì* 是…*de* 的cleft construction is challenging for L1 Italian learners to acquire because it conveys focus meaning commonly expressed by the Italian cleft sentence, but, with respect to the latter, has specific constraints relating to temporal reference, aspect, and discourse (Garassino 2014). Here, I will refer to the "adjunct focus *shì…de* cleft" (Mai & Yuan 2016: 249); it consists of a positionally determined focused element and a presupposition, where future-oriented temporal adverbs being excluded, with a past-tense reading only (Paul & Whitman 2008, Simpson & Wu 2002), as e.g. in (1):

(1) 他是昨天来的

| *tā* | *shì* | *zuótiān* | *lái* | *de* |
|---|---|---|---|---|
| 3SG | COP | yesterday | come | DE |

'It was yesterday that he came' (Jing-Schmidt 2017: 213).

This construction has received much attention in Second Language Acquisition research (Su & Tao 2018, among others); however, no research has been conducted on its acquisition by Italian-speaking learners. Moreover, it has never been studied within the framework of Learner Corpus Research (LCR), in which L2 Chinese is generally understudied (Iurato forthcoming), and this also applies to research on syntactic and discourse phenomena in LCR. Since currently available corpora of L2 Chinese mainly consist of data from Asian and English-speaking learners (Zhang & Tao 2018), I compiled a new learner corpus for the present research purposes.

I adopted a multi-method triangulated approach consisting of the combination of corpus and experimental data (Gilquin 2021) to 1) provide different insights into the phenomenon under study (Callies 2013), and 2) counterbalance potential construct underrepresentation (Tracy-Ventura & Myles 2015). The contributors to the corpus and the experiments correspond (Gilquin 2021), as they are 103 learners studying at Ca' Foscari University of Venice. Since external factors are considered unreliable factors to assess learners' proficiency and researchers encourage the use of external proficiency measures (Callies et al. 2014, Leclercq & Edmonds 2014), I grouped learners into elementary, intermediate, and advanced proficiency levels according to their HSK Chinese language proficiency test score. As for the corpus study, I collected written data through open-ended tasks (discourse completion test, picture description task). Moreover, I collected spoken data through closed/open role plays and semi-structured interviews. Written and spoken data are comparable as collected from the same learners using the same tasks. As for the experimental study, clinical data were collected through pragmalinguistic judgement tests, interpretation tasks, acceptability judgement tests, and retrospective interviews. A control group of 30 L1 Chinese speakers also completed the same tasks as the learners. I developed a target-oriented error taxonomy to manually annotate the grammatical errors; a pragmatic annotation was also added to detect the inappropriate use of the pragmatic functions (highlighting information and contrastive focus) of the *shì…de* construction. Following Granger (2012) and Díez-Bedmar (2015), the identification of errors was carried out by a bilingual team composed of two expert Chinese native speakers and a researcher whose L1 is the same as the learners.

The study addresses the following research questions:

- Are there differences and similarities between learners at different proficiency levels in the use of the *shì…de* construction in terms of quantity and quality of use?
- Are there any differences in the use of the *shì... de* construction by L1 Chinese speakers and L2 Italian learners in terms of quantity and quality of use, and in the ways the construction is used to highlight information and express contrastive focus?
- Is there evidence of cross-linguistic influence in terms of L1 transfer in the use of the *shì…de* construction, and is the typological distance between L1 and L2 an explanatory factor?

Inferential analyses show that learners use the construction with a lower frequency and a lower accuracy rate than native speakers. Following the Feature Reassembly Hypothesis (Lardiere 2009), results reveal how learners establish an L1 form–L2 form mapping and alter the L2 feature set in their interlanguage grammars, since there

is a persistence of L1 influence in morphosyntactic development. This study also confirms the Interface Hypothesis (Sorace 2005), since knowledge of the simultaneous application of the grammatical and pragmatic properties of the *shì...de* construction has not been developed by learners. The intensity of interaction in the L2 environment and the L1 pragmatic transfer (Bardovi-Harlig 2012) arguably affect the development of pragmatic comprehension of the construction by L1 Italian learners.

**References:**


Bardovi-Harlig, K. (2013). Developing L2 pragmatics. *Language Learning: A Journal of Research in Language Studies*, 63(1), 68–86.

Callies, M. (2013). Triangulation. In S.J. Schierholz & H.E. Wiegand (Eds.). *Wörterbücher zur Sprach- und Kommunikationswissenschaft [WSK] online*. Berlin: De Gruyter Mouton.

Callies, M., Díez-Bedmar, M.B., & Zaytseva, E. (2014). Using Learner Corpora for Testing and Assessing L2 Proficiency. In P. Leclercq, A. Edmonds & H. Hilton (Eds.). *Measuring L2 Proficiency: Perspectives from SLA*. Bristol, Blue Ridge Summit: Multilingual Matters, 71-90. https://doi.org/10.21832/9781783092291-007

Díez-Bedmar, M.B. (2015). Dealing with Errors in Learner Corpora to Describe, Teach and Assess EFL Writing: Focus on Article Use. In E. Castello, K. Ackerley & F. Coccetta (Eds.). *Studies in Learner Corpus Linguistics. Research and Applications for Foreign Language Teaching and Assessment*. Bern: Peter Lang AG, 37-69.

Garassino, D. (2014). Cleft sentences. Italian-English contrast. In A. De Cesare (Ed.), *Frequency, Forms and Functions of Cleft Construction in Romance and in Germanic*. Berlin, München, Boston: DeGruyter Mouton, 101-138.

Granger, S. (2012). How to Use Foreign and Second Language Learner Corpora. In A. Mackey & S.M. Gass (Eds.). *Research Methods in Second Language Acquisition. A Practical Guide.* Oxford: Wiley-Blackwell, 7-29.

Gilquin, G. (2021). Combining learner corpora and experimental methods. In N.Tracy-Ventura & M. Paquot (Eds.). *The Routledge Handbook of Second Language Acquisition and Corpora*. New York: Routledge, 133-144.

Iurato, A. (forthcoming). Learner Corpus Research meets Chinese Second Language Acquisition: Achievements and Challenges. *Annali di Ca' Foscari. Serie Orientale, 58.*

Jing-Schmidt, Z. (2017). Grammatical construction and Chinese Discourse. In C. Shei (Ed.). *Routledge Handbook of Chinese Discourse Analysis*. London: Routledge, 209-230.

Lardiere, D. (2009). Some thoughts on the contrastive analysis of features in second language acquisition. *Second Language Research*, 25(2), 173-227.

Leclercq, P. & Edmonds, A. (2014). How to Assess L2 Proficiency? An Overview of Proficiency Assessment Research. In P. Leclercq, A. Edmonds & H. Hilton (Eds.). *Measuring L2 Proficiency: Perspectives from SLA.* Bristol, Blue Ridge Summit: Multilingual Matters, 3-23. https://doi.org/10.21832/9781783092291-004

Mai, Z. & Yuan, B. (2016). Uneven reassembly of tense, telicity and discourse features in L2 acquisition of the Chinese shì…de cleft construction by adult English speakers. *Second Language Research*, 32(2), 247-276.

Paul, W. & Whitman, J. (2008). *Shi...de* focus clefts in Mandarin Chinese. *The Linguistic Review*, 25 (3/4), 413-451.

Simpson, A. & Wu, Z. (2002). From D to T- Determiner incorporation and the creation of tense. *Journal of East Asian Linguistics*, 11, 169-209.

Sorace, A. (2005). Selective optionality in language development. In L. Cornips & K.P. Corrigan (Eds.). *Syntax and Variation: Reconciling the Biological and the Social.* Amsterdam: John Benjamins, 55-80.

Su, D. & Tao, H. (2018). Teaching the shi... de construction with authentic materials in elementary Chinese. *Chinese as a Second Language Research*, 7(1), 111-140.

Tracy-Ventura, N. & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, 1(1), 58-95.

Zhang, J. & Tao, H. (2018). Corpus-based research in Chinese as a second language. In C. Ke (Ed.). *The Routledge Handbook of Chinese Second Language Acquisition.* London & New York: Routledge, 48-62.

# Register effects and morphosyntactic complexity affecting the use of the preterite construction in advanced L2 Finnish

Ilmari Ivaska
University of Turku
ilmari.ivaska@utu.fi

This paper reports on a study on the Finnish preterite constructions in texts written by advanced second language (L2) users. In Finnish, the preterite construction comprises a suffixal morpheme *-i-* attached to the verb stem before the markers of the grammatical person:

*Minä puhu-**i**-n suome-a*.
I speak-**pret**-1sg Finnish-prt
'I spoke Finnish.'

The use of the preterite construction in L2 Finnish is interesting for three inter-related reasons: First, it adds to the construction's morphosyntactic complexity via the interplay of frequency and salience when contrasted with the non-marked present tense constructions, making it prone to learner-sensitive patterning (Ellis 2016). Second, as with many languages (Biber 2014), different registers in Finnish vary drastically in terms of tense distribution (Pallaskallio 2003; Ivaska 2015), which makes tense a clear stylistic indicator that adds to the discourse-interactional complexity of the construction. Third, in relation to the discourse-interactional complexity, awareness of the register differences is described as an indicator of advanced linguistic proficiency (Council of Europe 2001). The research questions are the following:

1) Do advanced L2 users of Finnish diverge from L1 users in their use of the preterite construction in general?
2) How do different registers differ from one another?
3) Is the relationship between L2 and L1 users similar across registers?

The data comprise a balanced sample of 318 texts by advanced (C1–C2) L2 learners of three registers (academic, argumentative, narrative) and three language backgrounds (Czech, German, Russian) with comparable L1 data. The data stem from two corpora (The Corpus of Advanced Learner Finnish [Ivaska 2014]; The International Corpus of Learner Finnish [Jantunen 2011]). Two regression models are employed to address the use from a text-based and from a construction-based viewpoint. First, a linear mixed-effects model is used to model the frequency of the preterite construction as a function of two fixed variables: Variety (L2 and L1) and register and their interaction. The informant ID is included as a random variable to control for the dependency structures between observations (Winter & Grice 2021) and to account for idiolectal variation. Second, a logistic mixed-effects model is used to model the distribution between the preterite and the present tense constructions as a function of variety, register and their interaction, and the grammatical person of the construction and its interaction with variety as fixed variables, and the lemma of the verb, the text unit ID, and the informant ID as random variables. The interpretation focuses on analyzing the estimates provided by the models, but the statistical significance for each fixed variable is also measured by contrasting the full model with a model without that variable by means of a Likelihood Ratio Test. As for the random variables, the quantitative analysis focuses additionally on their contribution to the overall model fit.

The two models capture the variance in the data relatively well: $R_2$ of the text-based linear model is 0.39, and $R_2$ of the construction-based logistic model is 0.60. The results suggest that L2 users of Finnish use the preterite construction less frequently than L1 users ($X(1)= 5.561$, $p =0.001$), but that the underlying mechanisms are more complex. This difference is far outweighed by the difference across registers ($X(2)= 41.995$, $p<0.0001$). In addition, L2 users and L1 users use the construction in a similar fashion in the academic (estimates: 24.4 and 25,7 / 1,000 words, respectively) and in the argumentative register (estimates: 16.0 and 20.5 / 1,000 words, respectively), whereas there is a clear difference in the narrative register (estimate in L2: 55.2 / 1,000 words; median in L1: 78.6 / 1,000 words). This observation is corroborated by the statistically significant effect of the interaction between variety and register ($X(2)= 3.910$, $p=0.021$). Contrasting the preterite constructions with the present constructions reveals that L2 users differ from L1 users, especially in their use of the less frequent conjugational forms: the grammatical plural and the passive voice. Finally, the estimates of the random variables suggest that the variation across texts is relatively greater than the variation across informants, which in turn has a greater impact than the lemma of the verb. All in all, the difference between L2 and L1 users seems to be related

both to the construction's general morphosyntactic complexity and to its register-related distributional preferences.

**References**


Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast,* 14(1), 7-34. https://doi.org/doi:10.1075/lic.14.1.02bib.

Council of Europe. (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: CUP.

Ellis, N. C. (2016). Salience, Cognition, Language Complexity, and Complex Adaptive Systems. *Studies in second language acquisition,* 38(2), 341-351. https://doi.org/10.1017/S027226311600005X.

Ivaska, I. (2014). The Corpus of Advanced Learner Finnish (LAS2): Database and toolkit to study academic learner Finnish. *Apples – Journal of Applied Language Studies,* 8(3), 21-38.

Ivaska, I. (2015). Longitudinal changes in academic learner Finnish: A key structure analysis. *International Journal of Learner Corpus Research,* 1(2), 210–241. https://doi.org/10.1075/ijlcr.1.2.02iva.

Jantunen, J. (2011). Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. *Lähivõrdlusi. Lähivertailuja,* 21, 86-105. https://doi.org/10.5128/LV21.04.

Pallaskallio, R. (2003). Uutisaika. Finiittiverbin aikamuodoista katastrofiuutisissa 1892-1994. *Virittäjä,* 107(1), 27-45.

Winter, B., & M. Grice. (2021). Independence and generalizability in linguistics. *Linguistics,* 59(5), 1251-1277. https://doi.org/10.1515/ling-2019-0049.

## Empirical translation studies: Contrasting learner translations in a diglossic environment

Marlén Izquierdo[1], Naroa Zubillaga[2]
University of the Basque Country, UPV/EHU
marlen.izquierdo@ehu.eus[1], naroa.zubillaga@ehu.eus[2]

Empirical approaches to translation research have traditionally focused on product-oriented studies that observe and describe real-world translation phenomena from a corpus approach. Echoing Olohan's call for "contextualising translation by combining corpus-based investigations with other kinds of methodologies and analyses" (2003: 419), many scholars have underlined the need to integrate with such empirical observations more social, contextual, and cognitive data (Sutter & Lefer 2019). In this regard, for a few years now the convergence between corpus-based and process-oriented translation studies is shaping current empirical translation studies (Kotze 2019), thus connecting the three branches of Translation Studies, namely, product-, process-, and function-oriented research (Holmes 1972). This effort requires new-generation corpora that are "more carefully designed to take consideration of translators' backgrounds and the circumstances of text production" (Kotze 2020: 356). A prime example of such an endeavour is the Multilingual Student Translation (MUST) Project, a learner translation corpus enriched with standardized metadata related to the source text, the translation, and the students (Granger & Lefer 2020). Within this framework, our aim is to describe and contrast the results of a contrastive descriptive translation study that has been carried out to deepen our knowledge of both similarities and differences in the translation of the same English (EN) source text (ST) into two target languages (TL) in contact, namely, Basque (EU) and Spanish (ES). Considering the *diglossia* situation in which the languages under contrast are used for translation and/or other communicative purposes, our research questions relate to i) whether a given ST poses the same problems to learners translating it into different languages. Likewise, we wonder ii) whether the same chunks trigger the same or different translational errors in each TL, and finally, iii) to what extent translation products differ depending on whether the learner's mother tongue is EU or ES. Data for the analysis was taken from two MUST sub-corpora, i.e., English-Basque (EN-EU) and English-Spanish (EN-ES). Each of these corpora is a multiple translation corpus (Espunya 2014), as there are more than one translation for the same ST. In particular, we have selected a specific translation task common to both language pairs, thus narrowing a comparable parallel corpus (Hareide 2019). This corpus would guarantee the *tertium comparationis* for the study based on students' profile –demographic (gender, education) and linguistic (L1/L2)-; task instructions, and conditions of completion (e.g. time, setting, software used, grading); and size of datasets (number of translations, number of types). All the translations were aligned at the paragraph and sentence level and annotated not only for errors but for good translation choices too, for which we used the first version of the MUST-developed Translation-oriented Annotation System (TAS 1.0) (Granger & Lefer 2020). We then juxtaposed the TAS annotations suggested for each corpus, to identify both similarities and differences at two levels: content and language. Generally speaking, the findings of the study reveal some common problematic patterns in the ST. While the translational solutions given might coincide in nature, the error annotations point to an abundance of language problems for the EN-EU pair, while EN-ES translations feature more content-related phenomena. The interpretation of the results was done taking into account the sociolinguistic context, paying attention to the underlying diglossia situation in which the EU would be considered a "constrained language" (Kruger & Van Rooy 2016). This was also explained in light of the students' metadata on their linguistic background. The dominance of the ES language over EU is reflected in the fact that our students, enrolled in the Degree in Translation and Interpreting at the UPV/EHU, have comparatively fewer opportunities to work on their EN-EU translational competence than they do for the EN-ES language pair. The sociolinguistic reality of the native speaker seems to explain that the learner's command of EU is considerably poorer than that of ES. Teaching implications on the basis of the results obtained could involve the development of data-driven activities for remedial purposes of the most recurrent errors. Likewise, translation trainers and researchers may capitalize on the insights gained for the assessment of translations involving the two target languages, EU and ES.

**Keywords:** empirical translation studies, learner translations, contrastive descriptive translation studies, diglossia, English-Basque-Spanish.

**References**


De Sutter, G., & Lefer, M.A. (2019). On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*, DOI: https://doi.org/10.1080/0907676X.2019.1611891.

Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation*, 48, 33-43. DOI: https://doi.org/10.1007/s10579-013-9260-1.

Granger, S., & Lefer, M.A. (2020). The Multilingual Student Translation corpus: a resource for translation teaching and research. *Language Resources and Evaluation 54*, 1183-1199. DOI: https://doi.org/10.1007/s10579-020-09485-6.

Hareide, L. (2019). Comparable parallel corpora: A critical review of current practices in corpus-based translation studies. In I. Doval & M.T. Sánchez Nieto (Eds.). *Parallel Corpora for Contrastive and Translation Studies. New resources and applications SCL 90*, Amsterdam/Philadelphia: John Benjamins, 19-38. DOI: https://doi.org/10.1075/scl.90.02har.

Holmes, J. (1972). The name and nature of translation studies. In L. Venuti (Ed.). *The Translation Studies Reader*, 1st ed. London: Routledge, 172–185.

Kotze, H. (2019). Converging what and how to find out why: An outlook on empirical translation studies. In L. Vandevoorde *et al*., (Eds*). New Empirical Perspectives on Translation and Interpreting.* New York: Routledge, 333-371.

Kruger, H., & Van Rooy, B. (2016). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide 37*(1), 26–57. DOI: https://doi.org/10.1075/eww.37.1.02kru.

Olohan, M. (2002). Corpus linguistics and translation studies: Interaction and reaction. *Linguistica Antverpiensia*, New Series- Themes in Translation Studies. DOI: https://doi.org/10.52034/lanstts.v1i.29.

## Exploring the effect of target-language extramural activities on students' written production

Henrik Kaatari[1], Tove Larsson[2], Ying Wang[3], Seda Acikara Eickhoff[2], Pia Sundquist[4]
University of Gävle[1], Northern Arizona University[2], Karlstad University[3], University of Oslo[4]
henrik.kaatari@hig.se[1]

Frequent engagement in extramural English (EE) activities (i.e., English-language activities that students engage in outside of the classroom) has been shown to positively influence not only high school students' vocabulary size and listening and reading comprehension, but also their oral proficiency (see, e.g., Sundqvist 2009; 2019; Sylvén & Sundqvist 2012). However, while previous studies have greatly contributed to our understanding of the relationship between EE and students' *receptive* knowledge as measured through formal tests (e.g., of vocabulary, Sundqvist 2019), our understanding of the relationship between such activities and students' *production* remains somewhat rudimentary (though see Sundqvist & Wikström 2015 and Olsson & Sylvén 2015). What is more, whereas vocabulary knowledge (both receptive and productive) features prominently in studies on EE, syntactic and broader lexical aspects have received very limited focus. As both syntactic and lexical complexity have been shown to be strongly correlated with writing quality (Casal & Lee 2019; Kyle & Crossley 2016), examining the relationship between EE activities and linguistic complexity would help us better understand the role that such activities play for students' language development.

Against this background, the present study examines the effect of EE activities on both lexical and grammatical/syntactic features in high school student writing. Specifically, we focus on examining the effects of EE on lexical diversity and noun phrase (NP) complexity, as detailed below. The following research questions are investigated:

- What effect (if any) do EE activities have on lexical diversity and/or NP complexity?
- Are there differences between purely receptive EE activities and other types of EE activities in terms of the effect of lexical diversity and NP complexity, and, if so, what are the differences?

Based on previous research, we hypothesize that there will be a positive relation between EE activities and lexical diversity and NP complexity and that EE activities of the same type will behave similarly. The study uses data from the Swedish Learner English Corpus (SLEC), a corpus currently under compilation, which, as of now, consists of around 1,100 argumentative texts written by Swedish junior and senior high school students (grades 7–12). Here, we limit the focus to grades 9-11 using a subsample of SLEC. What sets SLEC apart from many other learner corpora is the fact that it contains detailed information about EE activities. Specifically, the corpus includes information on how many hours per week students (i) read in English, (ii) watch TV shows or movies in English, (iii) engage in conversations in English, (iv) spend time on social media with English content, and (v) communicate in English while playing computer/video games. For the purpose of the present study, we consider activities (i) and (ii) as (purely) receptive.

To measure lexical diversity, MATTR (moving average type-token ratio; Covington & McFall 2010) is used, as it has been demonstrated to produce stable results for short texts (see Zenker & Kyle 2021; the mean text length in SLEC is 458 words). To measure NP complexity, the rate of occurrence of attributive adjectives and prepositional phrases as modifiers in NPs is used. Frequent use of these features has been shown to be a sign of syntactic maturity, and as a key factor for distinguishing speech from writing (see, e.g., Biber 1988, and Biber et al. 2011).

In order to test the effect of EE on lexical diversity and NP complexity, we applied measured variable path analysis from the Structural Equation Modeling framework (SEM; see Larsson et al. 2021). Specifically, we fitted five competing models with hypotheses based on theory and previous studies to test their fit relative to our data. Given the flexible nature of this framework, we were able to look at the effect of the EE activities on all three of our complexity measures in a single model, as well as test relationships among all of these variables. The best-fitting model ($\chi^2$: 16.8, df: 15, CFI: 96.6, RMSEA: 0.023[0.00–0.067], SRMR: 0.039) confirmed our hypotheses that (a) participation in EE activities has a (mostly) positive effect on lexical diversity and NP complexity, (b) that the activities grouped differently based on type, where the purely receptive activities (in particular reading) each had an effect on lexical diversity, in a way that the other EE activities did not.

**References**


Biber, D. (1988). *Variation across Speech and Writing*. New York: CUP.

Biber, D., Gray, B & K. Poopon. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1), 5-35.

Casal, J. E., & J.J. Lee. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, 44, 51-62.

Covington, M. A., & McFall, J.D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94-100.

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12-24.

Larsson, T., Plonsky, L., & Hancock, G. (2021). On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory*, 17(3), 683-714.

Olsson, E., & Sylvén, L.K. (2015). Extramural English and academic vocabulary. A longitudinal study of CLIL and non-CLIL students in Sweden. *Apples - Journal of Applied Language Studies*, 9(2), 77–103.

Sundqvist, P. (2009). *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary*. Karlstad University Studies, 2009:55.

Sundqvist, P. (2019). Commercial-off-the-shelf games in the digital wild and L2 learner vocabulary. *Language Learning & Technology*, 23(1), 87-113.

Sundqvist, P., & Wikström, P. (2015). Out-of-school digital gameplay and in-school L2 English vocabulary outcomes. *System*, 51, 65-76.

Sylvén, L.K. & Sundqvist, P. (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL*, 24(3), 302-321.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, 100505.

## Use of English negation in the Slovene subcorpus of ICLE

Monika Kavalir[1], Gašper Ilc[2]
University of Ljubljana
monika.kavalir@ff.uni-lj.si[1], gasper.ilc@ff.uni-lj.si[2]

Negation is a complex phenomenon and a common challenge for L2 speakers. In English, it is most commonly expressed with the syntactic negator *not* (*not*-negation), modifying both verbal and non-verbal elements. Moreover, negation can be incorporated into lexical items, resulting in synthetic negation (*no*-negation), as in *nobody*, *nothing*, *nowhere*, etc. There are additional options such as the use of approximate negators (e.g. *barely*, *seldom*) and affixal negation (e.g. *un*-). In English, the negator licenses the occurrence of negative polarity items in its scope (NPIs; Huddleston & Pullum 2002, 799*ff*; Quirk et al. 1985, 780*ff*). The first major corpus analysis of negation in English by Biber et al. (1999, 169*ff*) suggested that *not*-negation is the prevailing type, accounting for approximately 75% of occurrences in academic prose and generally from 65% to 90% of occurrences depending on the register. Furthermore, *not*-negation can, by and large, replace *no*-negation, whereas the reverse is frequently impossible. In analysing English learner interlanguage, the focus has mainly been on the order of L2 acquisition of negation compared to L1 learners, and on transfer effects in the form of typical learner errors, particularly in the initial stages of language learning. Some common research topics include L1 vs. L2 interpretation of the scope of negation, the challenges in using and teaching NPIs and *do*-support in negated sentences, and double negation (e.g. Gil et al. 2019; Grüter et al. 2010; Milon 1974; Perales 2010).

Corpus studies of negation in learner English generally use the International Corpus of Learner English (ICLE; Granger et al. 2002; Granger et al. 2009). García-Fuentes' (2008) study found that when compared to L1 English speakers, Spanish students overused clausal negation, but underused subclausal and affixal negation. Negative transfer from L1 Spanish and limited knowledge of L2 English were seen to be responsible for mixing up *no* and *not*, lack of *do*-support, problems in the use of NPIs, and inconsistencies in the use of negative prefixes. Herriman (2009) examined the semantic functions of negation and discovered that both L1 and L2 student texts use negation less than professional writing at the level of content, but Swedish advanced learners use negation much more than the other two groups to express subjective interpersonal meanings, which makes their language more emphatic and closer to speech; the proportion of *not*-negation over *no*-negation is very close to L1 usage. Finally, Rankin (2012) investigated verb placement and found no V2 interference effects for Dutch and German L1 speakers, although such effects were found in declarative clauses. Word order errors in sentential negation mainly seemed to be related to a lack of distinction between auxiliary and lexical uses of *have*, *do*, and modal verbs.

The present paper is the first learner corpus study of Slovene English and seeks to explore the use of negation in the Slovene subcorpus of ICLE, which is currently being compiled at the University of Ljubljana. The main research questions relate to the choice between *not*- and *no*-negation, including the proportion of the two types and their distribution according to their preferred grammatical, semantic and lexical environments, such as the existential *there*-construction, combinations with mental verbs, or with the lexical verb *have*. The preliminary results based on an investigation of the incomplete ICLE-SI corpus (120,000 words) in Sketch Engine suggest that at 78% the proportion of *not*-negation is quite close to both LOCNESS and the Written Academic subcorpus of the British National Corpus (BNC), but that overall Slovene L1 speakers use negation considerably more than English L1 speakers and particularly professional writers, with *not*-negation twice as frequent and *no*-negation 50% more frequent in the ICLE-SI corpus compared to BNC Written Academic. The only *no*-negators that are slightly underused in comparison are *none* and *nor*, whereas *never* and *nothing* are the most overused negators. While the enthusiastic use of negation by Slovene L1 speakers may partly be attributed to register issues (e.g. *No beating around the bush*) similar to those noticed by Herriman (2009), it may also be related to the greater assertiveness/uncertainty avoidance of Slovene speakers (e.g. *There is no doubt whatsoever that*…), as attested for instance by lower frequencies of hedging devices (cf. Pisanski Peterlin 2010).

**References**


Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

García Fuentes, A. (2008) The use of English negation by Spanish students of English: A learner corpus-based study. *Revista de Lingüística y Lenguas Aplicadas*, 3(1), 49–58. https://doi.org/10.4995/rlyla.2008.689.

Gil, K. H., Marsden, H., & Whong M. (2019). The meaning of negation in the second language classroom: Evidence from 'any.' *Language Teaching Research*, 23(2), 218–236. https://doi.org/10.1177/1362168817740144.

Granger, S., Dagneaux, E., & Meunier, F. (2002). *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S., Dagneaux, E., Meunier, F., & Paquot M. (2009). *International Corpus of Learner English Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Grüter, T., Lieberman, M., & Gualmini A. (2010). Acquiring the scope of disjunction and negation in L2: A bidirectional study of learners of Japanese and English. *Language Acquisition*, 17(3), 127–154. https://doi.org/10.1080/10489223.2010.497403.

Herriman, J. (2009). Don't get me wrong! Negation in argumentative writing by Swedish and British students and professional writers. *Nordic Journal of English Studies*, 8(3), 117–140.

Huddleston, R. D., & G. K. Pullum. (2002). *The Cambridge Grammar of the English Language*. Cambridge: CUP.

Milon, J. P. (1974) The development of negation in English by a second language learner. *TESOL Quarterly*, 8(2), 137. https://doi.org/10.2307/3585537.

Perales, S. (2010). The status of the auxiliary do in L1 and L2 English negative clauses. *IRAL - International Review of Applied Linguistics in Language Teaching*, 48(1), 1–23. https://doi.org/10.1515/iral.2010.001.

Pisanski Peterlin, A. (2010). Hedging devices in Slovene-English translation: A corpus-based study. *Nordic Journal of English Studies*, 9(2), 171–193.

Quirk, R., Greenbaum, S., Leech, G., & J. Svartvik. (1985). *A Comprehensive Grammar of the English Language*. London; New York: Longman.

Rankin, T. (2012). The transfer of V2: Inversion and negation in German and Dutch learners of English. *International Journal of Bilingualism*, 16(1), 139–158. https://doi.org/10.1177/1367006911405578.

## Lexical bundles and L2 Spanish writing development: A case of dual language immersion

Elnaz Kia[1], Fernando Rubio[2]
The University of Utah
elnaz.kia@utah.edu[1], fernando.rubio@utah.edu[2]

Formulaic sequences play an important role in L2 writing development. Research shows that formulaic sequences are processed as single units (Ellis 1996) and often unanalyzed chunks; Therefore, they can be produced with lower processing time and fewer errors. Several studies have compared the use of formulaic sequences in L2 English writing with L1 English writing (Chen & Baker 2010). The majority concludes that L2 writers use fewer numbers and types of formulaic sequences than L1 writers. Other studies have examined the use of formulaic sequences by L2 English writers across proficiency levels (Staples et al. 2013); although inconsistent, findings show meaningful differences in the types of sequences used by lower-level and higher-level learners. These findings have important pedagogical implications for language teaching and learning. Yet, there is very little research on the developmental aspect of formulaic sequences in languages other than English. The present study aims to fill this gap by examining the developmental use of formulaic sequences in L2 Spanish writing. The study adopts a frequency-based approach to identifying multi-word units, i.e., lexical bundles (Biber et al. 1999). Lexical bundles are the most frequent recurring sequences of three or more words in a register (Biber et al. 1999).

The present study is based on analyzing texts from the three-million-word written Spanish subsection of the Corpus of Utah Dual Language Immersion (CUDLI; Rubio & Schnur 2019-). CUDLI is a multilingual corpus of second language writing. Texts in CUDLI are collected from the presentational writing section of the ACTFL Assessment of Performance toward Proficiency in Languages (AAPPL), administered in Utah's Dual Language Immersion (DLI) program in grades four, six, eight, and nine (age range 9-15). The writing portion of AAPPL includes six prompts and is scored holistically. Test takers receive an overall rating for the entire writing section. The AAPPL ratings run from Novice Low to Advanced, with four sub-levels in the Novice range (N1, N2, N3, N4), five in the Intermediate range (I1, I2, I3, I4, and I5), and one in the Advanced level (A).

The sub-corpus used in this study includes texts from eighth-grade learners (90% L1 English and 10% heritage Spanish speakers) with intermediate and advanced proficiency ratings. The ACTFL intermediate level roughly corresponds to A2 to B1.2 in the Common European Framework of Reference for Languages (CEFR), and the advanced level corresponds to CEFR B2.1 to C1 (ACTFL n.d.). Most of the eighth-grade students tested (97%) are rated intermediate or advanced, with the most common rating being I3 (approximately B1.1 on the CEFR scale). Therefore, we consider this group an ideal population to explore differences between the intermediate and advanced proficiency levels. For this study, we compare students in the mid-range of the intermediate level (ratings of I2, I3, and I4) with students rated advanced. Specifically, we address the following research questions:

1. What differences exist in the number of types and tokens of lexical bundles used in intermediate- and advanced-level L2 Spanish writing?
2. What differences exist in the structural types of lexical bundles used by intermediate and advanced-level L2 Spanish writings?

To answer these questions, we extract three-word bundles that frequently occur at two proficiency levels—intermediate and advanced. Then, we classify the bundles based on their structural characteristics—noun phrases (NP), verb phrases (VP), prepositional phrases (PP), clauses, and pronoun-based bundles (Pro). To compare the use of lexical bundles by intermediate and advanced learners, we report statistical analyses and interpret the results using linguistic descriptions.

Preliminary results demonstrate structural differences in bundles used by intermediate and advanced students. Intermediate writers mostly use NP bundles (e.g., *mi escuela es*), whereas advanced writers use VP (e.g., *mi me gusta*) and PP (e.g., *en el futuro*) most frequently. As expected, most of the NP and PP bundles in both groups are related to concrete topics in the immediate environment of these learners (friends, school, family) and topics elicited by the prompts (e.g., social media). Of note is the fact that a strong predominance of NP bundles at the intermediate level is replaced by a more balanced mix of VP, PP, NP, and Pro bundles at the advanced level. These findings have important implications for L2 Spanish teaching and materials development that will be discussed.

**References:**


ACTFL. (n.d.). *Assigning CEFR Ratings to ACTFL Assessments.* https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *14*(2), 30-49.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in second language acquisition*, *18*(1), 91-126. https://doi.org/10.1017/S0272263100014698

Rubio, F. & Schnur, E. (2019-). *The Corpus of Utah Dual Language Immersion (CUDLI).*

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*(3), 214-225.

# Multidimensional analysis of syntactic complexity development in L2 learner writing in an American university EAP programme

Sangeun Kim
University of Exeter
sk688@exeter.ac.uk

Syntactic patterns are a foundational element of linguistic production and therefore, measures of syntactic complexity are a valid indicator of English writing development (Ortega 2015). Empirical research findings suggest significant correlations between syntactic complexity and holistic evaluation of English writing quality (e.g., Larsson & Kataari 2020; Bulté & Housen 2014; Kyle & Crossley 2018). Longitudinal studies have also shown the increased importance of phrasal features in more advanced English writing (e.g., Grey et al. 2019; Biber et al. 2020). However, relatively little is known about patterns of syntactic development comprising clausal and phrasal levels in academic writing produced by speakers of English as an additional language (L2) over time, and this longitudinal study aims to fill this gap.

This study takes longitudinal, corpus-based approaches to explore syntactic variations in academic writing produced by L2 adult learners over three academic semesters. The fine-grained evidence gained from an extensive analysis of linguistic variations and their underlying communicative functions is expected to provide a fuller picture of the syntactic development appropriate to academic written discourses.

This study aims to answer two research questions:

1) How do the syntactic complexity features and the functional characteristics of academic texts written by L2 English learners systematically vary across different academic semesters?

2) How do the writing topics, language background, and linguistic proficiencies interact with different dimensions of syntactic complexity?

The primary data is the University of Pittsburgh English Language Institute corpus (PELIC; Juffs et al. 2020), an L2 longitudinal learner corpus collected in an intensive English programme. This corpus consists of written texts produced by learners of 30 linguistic backgrounds in a language classroom in a non-experimental setting. This study selects the written essays produced by the same participants for at least three consecutive academic semesters in writing classes which provides a pivotal reference point to track linguistic changes over time. In addition, the Michigan Corpus of Upper-Level Student Papers (MICUSP; O'Donnell & Römer 2012; Römer & O'Donnell 2011) is used as reference data to provide a comparable point of native speaker use in similar academic contexts.

This study adopts a multidimensional (MD) analysis, a corpus-based approach to text analysis pioneered by Biber (1988). The 46 syntactic complexity measures used in this study incorporate phrasal and clausal levels, which reflects the recent findings that support the importance of phrasal complexity measures as an index of syntactic development in academic written registers (Biber et al. 2011; Kyle & Crossley 2018).

The basic procedures are as follows:

First, the PELIC corpus is pre-processed and coded by the personal course of study, reflecting the temporal order of different semesters in which a student submitted written texts. Then, the corpus is tagged to 46 syntactic features using the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle 2016) to extract the raw frequency of each syntactic feature in each text. Next, the frequencies of each syntactic feature in the tagged texts are standardised to check for their correlations among features. Features with low correlations are excluded and the retained features are used for Principal Component Analysis, a type of factor analysis, which identifies the functional dimensions underlying co-occurring syntactic features within the corpus. Subsequently, the dimension scores are computed for each text in both the PELIC and the MICUSP corpora as the basis of comparison among the PELIC sub-corpora sorted by the personal course of study (semesters) and the MICUSP corpus. Finally, the Mixed-effect model is fitted to rule out the mediating effects and measure the actual effect size of the 'time' variable.

The preliminary findings confirm the previously noted spoken and written register distinction (evidence of spoken register in novice academic writers) as an indicator of academic English development in L2 writing (e.g., Biber & Gray 2013; Kobayashi & Abe 2016; Kim & Nam 2019). More specifically, as their studies progressed, the learner texts demonstrated a closer association with phrasal complexity. Further analysis is expected to provide more explicit trajectories that learners follow as they progress through different semesters.

This MD analysis will contribute to the strand of MD analyses of learner corpora, providing longitudinal evidence of the previous hypothesis of English grammatical development with a more consistent and specific focus on patterns in syntactic measures.

**References**


Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511621024

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly*, *45*(1), 5-35.

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL ibt® test: a lexico-grammatical analysis*. ETS Research Report Series, 2013(1), i-128.

Biber, D., Reppen, R., Staples, S., & Egbert, J. (2020). Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. *International Journal of Learner Corpus Research*, *6*(1), 38-71.

Bulté, B., & Housen, A. (2014). Conceptualising and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42-65. https://doi.org/10.1016/j.jslw.2014.09.005

Gray, B., Geluso, J., & Nguyen, P. (2019). The longitudinal development of grammatical complexity at the phrasal and clausal levels in spoken and written responses to the TOEFL iBT® test. *ETS Research Report Series*, *2019*(1), 1-51.

Juffs, A., Han, N-R., & Naismith, B. (2020). The University of Pittsburgh English Language Corpus (PELIC) [Data set]. http://doi.org/10.5281/zenodo.3991977. Retrieved from: https://eli-data-mining-group.github.io/Pitt-ELI-Corpus. Accessed 24 June 2021.

Kim, J. E., & Nam, H. (2019). How do textual features of L2 argumentative essays differ across proficiency levels? A multidimensional cross-sectional study. *Reading and Writing*, *32*(9), 2251-2279.

Kobayashi, Y., & Abe, M. (2016). A Corpus-Based Approach to the Register Awareness of Asian Learners of English. *Journal of Pan-Pacific Association of Applied Linguistics*, *20*(2), 1-17.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Doctoral dissertation). Retrieved from: https://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1035&context=alesl_diss

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *The Modern Language Journal*, *102*(2), 333-349. https://doi.org/10.1111/modl.12468

Larsson, T., & Kaatari, H. (2020). Syntactic complexity across registers: Investigating (in)formality in second-language writing. *Journal of English for Academic Purposes*, *45*, 100850. https://doi.org/10.1016/j.jeap.2020.100850

O'Donnell, M. B., & Römer, U. (2012). From student hard drive to web corpus (part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP). Corpora, 7(1), 1-18.

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, *29*, 82-94. https://doi.org/10.1016/j.jslw.2015.06.008

Römer, U., & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). Corpora, 6(2), 159-177.

# Non-canonical syntax in learner language: Between language transfer, language universals and idiosyncrasies

Kathrin Kircili
University of Marburg
kathrin.kircili@uni-marburg.de

Both in speech and writing, language users have at their disposal numerous ways of meeting information structural needs and achieving particular communicative goals, such as the placement of emphasis or contrast or the linking of discourse units. However, while in spoken interactions intonation frequently suffices to achieve said goals, writers commonly have to resort to other means, such as the use of non-canonical patterns to convey their communicative interests. Naturally, these phenomena, which deviate from the commonly acknowledged basic clause patterns (cf. Quirk et al. 1985), can be particularly challenging for learners, not only due to their inherent syntactic complexity and their information structural peculiarities but also potential transfer-related influences.

While a considerable amount of research has been dedicated to the description of non-canonical sentence patterns in the ENL context (e.g. Quirk et al. 1985; Birner & Ward 1998; Biber et al. 1999, Huddleston & Pullum 2002, Gómez-González 2001 or Martínez Lirola 2009 (all of which discuss multiple phenomena), Prince 1997 (left dislocation), Martínez Insua 2004 (existential-*there*) or Kreyer 2006 (inversion), to name but a few), in EFL, the majority of studies have focused either on individual phenomena rather than comprehensive overviews (cf. e.g. Larsson 2016/2017 on introductory-*it*, Leńko-Szymańska 2008 or Van Vuuren & Laskin 2017 on fronting, Balhorn 1996 or Palacios-Martínez & Martínez-Insua 2006 on existential-*there*, or Lozano & Mendikoetxea 2007/2008/2010 on inversion) or on individual language backgrounds only (cf. e.g. Callies 2009 on German learners).

Against this backdrop, this study reports on findings of a *Contrastive Interlanguage Analysis* (cf. Granger 2015) taking six non-canonical structures (i.e. fronting/preposing, inversion, existential-*there*, introductory-*it* as well as right- and left dislocation) among learners of four L1 backgrounds (German, Spanish, Turkish, Japanese) into account. The analysis is based on randomly sampled essays from the *International Corpus of Learner English* (Granger et al. 2009) along with the British and American components from the *Louvain Corpus of Native English Essays*.[1]

In total, 270 essays, and more than 12,000 T-units (cf. Hunt 1965) were manually annotated for various syntactic and pragmatic variables, including, among others, CONSTITUENT LENGTHS, INFORMATION STATUS, TYPE and FUNCTION of the non-canonical phenomenon. The data were then subjected to multifactorial analyses and regression modelling to test for possible predictors accounting for the occurrence of non-canonical patterns in the learner vs. the native-speaker data and to shed light on whether the use of particular structures can be traced back to language transfer or -universals (cf. e.g. Gass 1984).

The analyses suggest that there are, indeed, commonalities among the learner populations, concerning both syntactic and discourse-pragmatic choices. Fronting, for instance, was identified as the most frequent non-canonical pattern, while the learners also seem to share the preference of particular clusters of phenomena within one T-unit, such as the combination of fronting, introductory-*it*, and the embedded existential in (1):

(1) *Despite these countries having been awarded with more number of votes* […], *it should not be forgotten that there are more countries knocking at the European Union door* […]. < ICLE-SP-UCM-0014.2>

As regards information structure – which has been acknowledged as a language universal in the past (cf. Zimmermann & Féry 2010) – it has, indeed, been found that its principles seem to be universally realizable. One example includes the use of the *given-new* progression successfully applied by Japanese learners even though their L1 follows the reverse order.

Other observations also point at certain transfer-related choices, however, ranging from the utter avoidance of structures that are not present in a learner's L1, including the absence of inversion in the Turkish and Japanese data, to over-representations of patterns that are familiar from their L1 as in (2), showing a case of left dislocation in the Spanish data or (3), an introductory-*it* construction produced by a German learner:

(2) *Children instead of reading books, they play computer games*. <ICLE-SP-UCM-0003.5>
(3) *It is highly recommended to keep a watchful eye on small toddlers*. <ICLE-GE-AUG-0004.3>

[11] Cf. http://www.learnercorpusassociation.org/resources/tools/locness-corpus/

In particular, the overrepresentations seem to suggest the existence of 'syntactic teddy bears' (in line with Hasselgren's (1994) concept of 'lexical teddy bears') as well as certain idiosyncratic tendencies. The results shall be discussed on both a quantitative and qualitative basis, shedding light on the implications they might have for the acquisition of non-canonical patterns in SLA.

**References**


Balhorn, M. (1996). *Existential-presentational sentences in Second Language Acquisition*. Paper presented at the Annual Meeting of the American Association for Applied Linguistics (18th, Chicago, IL, March 23-26, 1996).

Biber, D., S. Johannson, G. Leech, S. Conrad & E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Birner, B. J. & G. Ward (1998). *Information Status and Noncanonical Word-Order in English*. Amsterdam/Philadelphia: John Benjamins.

Callies, M. (2009). *Information Highlighting in Advanced Learner English: The syntax- pragmatics interface in second language acquisition*. Amsterdam/Philadelphia: John Benjamins.

Gass, S. (1984). A Review of Interlanguage Syntax: Language Transfer and Language Universals. *Language Learning*, 34(2), 115-132.

Gómez-González, M. A. (2001). *The theme-topic interface. Evidence from English*. Amsterdam/Philadelphia: John Benjamins.

Granger, S., E. Dagneaux & F. Meunier (2009). *International Corpus of Learner English*. Louvain: UCL.

Granger, S. (2015). Contrastive Interlanguage Analysis: A Reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, 4(2), 237-258.

Huddleston, R. & G. K. Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hunt, K. (1965). *Grammatical Structures Written at Three Grade Levels: NCTE Research Report No. 3*. Champaign (Illinois): National Council of Teachers of English.

Kreyer, R. (2006). *Inversion in Modern Written English: Syntactic Complexity, Information Status and the Creative Writer*. Tübingen: Gunter Narr.

Larsson, T. (2016). The introductory it pattern: Variability explored in learner and expert writing. *Journal of English for Academic Purposes* 22, 64-79.

Larsson, T. (2017). A functional classification of the introductory it pattern: Investigating academic writing by non-native-speaker and native-speaker students. *English for Specific Purposes* 48, 57-70.

Leńko-Szymańska, A. (2008). Non-native or non-expert? The use of connectors in native and foreign language learners' texts. *Acquisition et Interaction en Langue Étrangère* 27, 91-108.

Lozano, C. & A. Mendikoetxea (2007). Learner corpora and the acquisition of word order: A study of the production of Verb-Subject structures in L2 English. In M. Davies, P. Rayson, S. Hunston & P. Danielsson (Eds.). *Proceedings of the Corpus Linguistics Conference 2007*. University of Birmingham.

Lozano, C. & A. Mendikoetxea (2008). Postverbal subjects at the interfaces in Spanish and Italian learners of L2 English: a corpus analysis. In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds). *Linking up Contrastive and Learner Corpus Research*. Amsterdam: Rodopi, 85-125.

Lozano, C. & A. Mendikoetxea (2010). Interface conditions on postverbal subjects: a corpus study of L2 English. *Bilingualism: Language and Cognition*, 13(4), 475-497.

Martínez-Insua, A. (2004). *Existential There-Constructions in Contemporary British English*. München: LINCOM.

Martínez Lirola, M.(2009). *Main processes of thematization and postponement in English*. Bern: Peter Lang.

Palacios-Martínez, I. & A. Martínez-Insua (2006). Connecting linguistic description and language teaching: native and learner use of existential there. *International Journal of Applied Linguistics*, 16(2), 213-231.

Prince, E. (1997). On the Functions of Left-Dislocation in English Discourse. In A. Kamio (Ed.). *Directions in Functional Linguistics*. Amsterdam/Philadelphia: John Benjamins, 117-144.

Quirk, R., S. Greenbaum, G. Leech, & J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. Harlow: Longman.

Van Vuuren, S. & L. Laskin (2017). Dutch learner English in close-up: A Bayesian corpus analysis of pre-subject adverbials in advanced Dutch EFL-writing. *International Journal of Corpus Research*, 3(1), 1-35.

Zimmemann, M. & C. Féry (2010). Introduction. In M. Zimmermann & C. Féry (Eds.). *Information Structure: Theoretical, Typological, and Experimental Perspectives*. Oxford: Oxford University Press, 1-12.

# Lexical and syntactic complexity development in L2 Russian texts and correlations with curricular levels and raters' scores

Olesya Kisselev[1], Rossina Soyan[2], Dmitrii Pastushenkov[3], Jason Merrill[4]
University of Texas at San Antonio[1], Carnegie Mellon University[2], Michigan State University[3,4]
olesya.kisselev@utsa.edu[1], rsoyan@andrew.cmu.edu[2], pastushe@msu.edu[3], merril25@msu.edu[4]

Linguistic complexity has served as an important measure of the second language (L2) writing development and an important construct in language assessment (Larsen-Freeman 2006; Lu 2011; Verspoor et al. 2012). Complexity indices, however, rarely feature in the studies of learner languages other than English, especially the Less Commonly Taught Languages (LCTLs). Additionally, studies utilizing complexity measures in assessing L2 data have been criticized for the lack of consistency in defining proficiency (Gablasova et al. 2017; Ortega 2012). In this proposed paper, we attempt to address these gaps by exploring writing development in one LCTL, Russian, while paying specific attention to the operationalization of the notion of proficiency.

The study is based on a corpus of 601 essays (103,150 words total) written by 133 L2 Russian learners at different levels of proficiency, before and after an 8-week intensive language program. The learners were asked to write one narrative and two argumentative essays at the entrance and exit tests in the span of 90 minutes. While more proficient learners submitted all three essays, less proficient students submitted only one or two essays.
With the help of the compiled corpus, we investigate which lexical and syntactic complexity measures can help a) track writing development over the course of an instructional program and b) distinguish proficiency levels, operationalized as program-internal curricular levels as well as ratings on a standardized writing proficiency test (based on the ACTFL Proficiency guidelines).

Following previous research (e.g., Bulté & Housen 2014; Knoch et al. 2014; Yang et al. 2015) we employed descriptive statistics, paired samples t-tests, and the Wilcoxon signed-rank test to address the question of which lexical and syntactic complexity measures help index writing development over the course the program (RQ1). To investigate the relationships between the complexity indices and the program's curricular levels (RQ2) and the relationships between the complexity indices and proficiency level rankings (RQ3), we conducted correlation and multiple linear regression analyses.

Our results demonstrate that at least nine complexity indices (i.e., mean word length, MTLD by wordform and by lemma, the percentage of high-frequency words, mean sentence length, number of clauses per sentence, syntactic depth, number of subordinate clauses, and proportion of relative clauses) changed significantly over the course of the program and can reliably track language development (RQ1). The same nine indices showed significant correlation with the initial curricular placement (RQ2), and all–with the exception of the proportion of clauses per sentence–modestly or highly correlated with the final ratings on the writing proficiency test, with three indices serving as strongest predictors of proficiency levels (mean word length ($\beta = .25$), mean sentence length ($\beta = .21$), and proportion of subordinate clauses ($\beta = .15$)) (RQ3).

The findings largely confirm the developmental patterns identified in previous L2 complexity research: consistent with the results of many previous studies, sentential, clausal, and phrasal complexity in the texts of our learners increased with the time spent in the program. For example, we found some evidence in favor of gradual progression from coordination to subordination. However, we also established that the index of subordination is best interpreted when various specific types of subordinate structures are assessed separately; for instance, we found no changes with growing proficiency or with time spent in the program in such structures as infinitive clauses or adverbial clause modifiers, but relative clauses increased in number with growth in proficiency. These findings add to the possible interpretation of the different results of some previous studies which either find growth in subordination overall (Huang et al 2021; Mazgutova & Kormos 2015; Polat et al 2020) or fail to register any such growth (Bulté & Housen 2014; Lu 2011). It appears that not all subordinate structures are "created equal" and that some specific types of subordinate structures can better index proficiency levels than others.

Since the results for nine lexical and syntactic indices were stable across the three research questions and the different statistical procedures, we conclude that these complexity measures have significant implications for the development of (semi)automated assessment tools, granted these measures are further tested on larger sets of learner data.

Overall, the study offers insights into interlanguage development in an instructional context and adds to a better understanding of linguistic correlates/predictors of proficiency in a second language.

**References**


Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, *67*(S1), 130–154.

Huang, T., Steinkrauss, R., & Verspoor, M. (2021). Variability as predictor in L2 writing proficiency. *Journal of Second Language Writing*, *52*, 100787.

Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing*, *21*(1), 1–17.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, *27*(4), 590–619.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, *29*(C), 3–15.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: De Gruyter, 127–155.

Polat, N., Mahalingappa, L., & Mancilla, R. L. (2020). Longitudinal growth trajectories of written syntactic complexity: The case of Turkish learners in an intensive English program. *Applied Linguistics*, *41*(5), 688-711.

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*(3), 239–263.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53–67.

Bulté, B. and Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26:42–65.

Gablasova, D., Brezina, V., and McEnery, T. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, 67(S1):130–154.

Huang, T., Steinkrauss, R., and Verspoor, M. (2021). Variability as predictor in L2 writing proficiency. *Journal of Second Language Writing*, 52:100787.

Knoch, U., Rouhshad, A., and Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing*, 21(1):1–17.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27(4):590–619.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1):36–62.

Mazgutova, D. and Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, 29(C): 3–15.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: De Gruyter, pages 127–155.

Polat, N., Mahalingappa, L., and Mancilla, R. L. (2020). Longitudinal growth trajectories of written syntactic complexity: The case of Turkish learners in an intensive English program. *Applied Linguistics*, 41(5):688-711.

# Towards standardizing LCR metadata

Alexander König[1], Jennifer-Carmen Frey[2], Egon W. Stemle[3], Aivars Glaznieks[4], Magali Paquot[5]
CLARIN ERIC[1], Eurac Research[2,3,4], Université Catholique de Louvain[5]
alex@clarin.eu[1], {jennifercarmen.frey, egon.stemle, aivars.glaznieks}@eurac.edu[234],
magali.paquot@uclouvain.be[5]

Over the last decades, research data management has become a central task in the scientific enterprise. Research infrastructures such as CLARIN (de Jong et al. 2020) have been developed to provide services and technologies to improve data sustainability. Many communities have taken important steps to ensure interoperability and reusability of research data (e.g. the CMC community, see Beißwenger & Lüngen 2020). In learner corpus research (LCR), however, research data management has attracted less attention, with much room for improvement in terms of sustainable use of resources, comparability, and interconnectivity of individual studies (Tracy-Ventura et al. 2021; Stemle et al. 2019).

One area that would benefit significantly from standardization is corpus description, which includes metadata at the level of the learner corpus as a whole and metadata used to describe the individual learners and task types/registers the corpus is meant to represent. There are a number of reasons why this is important. First, standardized and well-structured metadata increases the findability and usability of existing learner corpora. Second, it should enhance the comparability of datasets and comparability of LCR studies, provided researchers agree on a common set of definitions. Extensive metadata that follow - at best - a standardized vocabulary, and have a strong focus on findability, accessibility, interoperability, and reusability (FAIR) are an essential aspect of FAIR research data (Wilkinson et al. 2016). Today, however, it is still unclear to what extent standardization of metadata would be possible in Learner Corpus Research and preliminary work on the topic (Granger & Paquot, 2017) shows the complexity of this issue.

To estimate the feasibility of such an approach, we tried to apply the metadata schema proposed by Granger & Paquot (2017) to five learner corpora available for research purposes. In this effort, we identified a set of core metadata fields that we consider necessary to describe learner corpora consistently and informatively, while also leaving room for optional information. Although all corpora were collected in the context of school education, they represent a variety of learners and language samples, thus providing a rich testbed.

The main objective of this presentation is to introduce this revised metadata schema for learner corpora, which is the result of extensive collaboration between a research data infrastructure expert and member of CLARIN's metadata taskforce, and data owners for the five resources. In line with Granger & Paquot (2017), our proposed metadata schema is divided into a number of different sections for Corpus metadata (itself divided into administrative metadata (e.g. authors or license) and design metadata (e.g. date and place of collection or type of task)), Text metadata (fine-grained per-text information), Author metadata (details about the learners, e.g. age, languages spoken), Annotator metadata (e.g. professional and language background), Transcriber metadata (e.g. native language or language repertoire) and Task metadata (e.g. instructions, time constraints). While basic information about learners (authors) and language samples (texts) are typically found as part of metadata associated with a learner corpus, other aspects such as those related to the annotation or transcription procedure or the specificities of a task are often found elsewhere (e.g. corpus manual) or are just absent from currently available learner corpora. In our presentation, we argue in favour of a systematic description of all these aspects as part of core metadata.

While the metadata schema was initially created in a simple tab-separated format, it is currently being transformed into the CMDI metadata format (Broeder et al. 2012) using the CMDI Core Components (https://clarin-eric.github.io/cmdi-core-components/). This will serve as a viable use case for the creators of the core components and as an "off-the-shelf"-profile for any researcher seeking one for their learner corpus project.

The schema will be made available as CMDI in the CLARIN Component Registry (https://catalog.clarin.eu/ds/ComponentRegistry/) and as a resource on the research data repository of the Eurac Research Clarin Center (ERCC, https://clarin.eurac.edu/), where the corpora and their accompanying metadata that were used for the development of the metadata schema are also available. Additionally, a detailed schema description will be provided to the research community at the learner corpus portal PORTA (https://www.porta.eurac.edu/).

**References**


Beißwenger, M., & Lüngen, H. (2020). *CMC-core: A schema for the representation of CMC corpora in TEI.* Corpus 20. https://doi.org/10.4000/corpus.4553

Broeder, Daan, et al. CMDI: A component metadata infrastructure. *Proceedings of the workshop describing language resources with metadata: towards flexibility and interoperability in the documentation of language resources*. LREC 2012, 1-4.

de Jong, F., Maegaard, B., Fišer, D., van Uytvanck, D., & Witt, A. (2020). Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN. *Proceedings of the 12th Language Resources and Evaluation Conference*, 3406–3413. https://www.aclweb.org/anthology/2020.lrec-1.417

Granger, S. & Paquot, M. (2017). Towards standardization of metadata for L2 corpora. *Invited talk at the CLARIN workshop on Interoperability of Second Language Resources and Tools*, 6-8 December 2017, University of Gothenburg, Sweden. https://sweclarin.se/sites/sweclarin.se/files/event_atachements/Granger_Paquot_Metadata_G%C3%B6teborg_final.pdf

Stemle, E. W., Boyd, A., Janssen, M., Lindström Tiedemann, T., Mikelić Preradović, N., Rosen, A., Rosén, D., & Volodina, E. (2019). Working together towards an ideal infrastructure for language learner corpora. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (Eds.). *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference 2017*. Louvain-la-Neuve: Presses universitaires de Louvain, 437–478.

Tracy-Ventura, N., Paquot, M. & F. Myles. (2021). The future of corpora in SLA. In N. Tracy-Ventura & M. Paquot (Eds). *The Routledge Handbook of Second Language Acquisition and Corpora*. New York: Routledge, 409-424.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.

# Exploring early L2 writing development: A register-functional approach to grammatical complexity

Tove Larsson[1], Tony Berber-Sardinha[2], Bethany Gray[3], Doug Biber[4]
Northern Arizona University[1,4], Pontifical Catholic University of Sao Paulo[2], Iowa State University[3]
tove.larsson@nau.edu[1], tonycorpuslg@gmail.com[2], begray@iastate.edu[3], douglas.biber@nau.edu[4]

Writing development has received a great deal of attention in studies on second-language (L2) users of English in recent years (e.g., Gray et al. 2019; Parkinson and Musgrave, 2014). Many studies in this line of research have turned to grammatical complexity as "an index of language development and progress" (Bulté & Housen, 2014: 43). The present study looks at writing development in first-language (L1) Brazilian Portuguese speakers of (L2) English through the lens of grammatical complexity, as outlined below.

Grammatical complexity, here defined as the addition of optional structural elements to 'simple' phrases and clauses, has been studied through different frameworks. For example, many studies have focused on the effectiveness of omnibus measures, such as the mean length of T-units, for predicting language proficiency or development (see, e.g., Lu, 2017; Bulté & Housen, 2018). These measures have, however, been criticized in relation to their linguistic interpretability and their usefulness for descriptive studies of language (Biber et al., 2020). The researchers in the latter tradition instead argue in favor of a register-functional approach to complexity (Biber et al., 2020), directly analyzing the use of specific lexico-grammatical complexity features that combine particular structures with particular syntactic functions (e.g., finite *that* complement clauses controlled by a verb, non-finite participial clause post-modifying a noun) rather than employing omnibus measures (see, e.g., Biber & Gray, 2016; Biber et al., 2011; Biber et al., 2020). Within this general theoretical framework, Biber et al., 2011 build on prior corpus analyses of the complexity features that are common in conversation versus informational writing to propose a series of developmental stages, with each stage being defined by a group of complexity features. Since 2011, numerous empirical studies have provided strong descriptive evidence that L2 writing development progresses generally according to these hypothesized stages (see, e.g., Taguchi et al. 2013; Parkinson and Musgrave 2014; Staples et al. 2016; Ansarifar et al. 2018; Lan & Sun 2019; Gray et al. 2019; Biber et al., 2020).

However, while we know a fair bit about the use of complexity features by L2 English writers at the advanced end of the proficiency level spectrum, our knowledge is far more limited when it comes to beginner-level writers. That is, little is known about whether development along the hypothesized developmental sequence is evident already in beginner-level students' production, and if so, how quickly they progress from one stage to the next. Against this background, the present cross-sectional study aims to employ a register-functional approach to explore the development of grammatical complexity in low-proficiency L2 writing across six proficiency levels to answer the following two research questions:

- To what extent are the developmental stages proposed in Biber et al. (2011) evident in low-proficiency L2 writing?
- To the extent that these stages are detectable, what patterns of progression in terms of grammatical complexity can be identified across the six levels in the data? Do these stages correspond to a movement away from speech-like production toward more advanced written production?

The study uses data from COBRA, a corpus of L1 Brazilian Portuguese adult learner production. All the data were tagged using the Biber tagger (Biber, 1988) and the Developmental Complexity tagger (Gray et al. 2019), and subsequently analyzed using a technique developed in Staples et al. (under review) to quantify developmental profiles across levels. The technique considers not only the overall change in frequency across levels but also the incremental variation across each adjacent level (based on % frequency changes).

Early results show that while no substantial changes could be noted across the proficiency levels in the data, the overall trends were consistent with the hypothesized stages in that phrasal features (in particular attributive adjectives) became more prominent, whereas dependent clause features (e.g., non-finite complement clauses) remained infrequent. These results were consistent with gradual movement away from speech-like production toward more advanced written production. It thus seems as if the hypothesized stages remain relevant for early writing development.

**References**


Ansarifar, A., Shahriari, H., & Pishghadam, R. (2018). Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes*, 31, 58–71.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Biber, D., Gray, B., & Poopon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly*, 45(1), 5–35.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *International Journal of English for Academic Purposes*, 46.

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics*, 28, 147–164.

Bulté, B. & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language* Writing, 26, 42–65.

Gray, B., Geluso, J., & Nguyen, P. (2019). The longitudinal development of grammatical complexity at the phrasal and clausal levels in spoken and written responses to the *TOEFL iBT* test. ETS Research Report No. RR-19–45. Princeton, NJ: Educational Testing Service.

Lan, G., & Sun, Y. (2019). A corpus-based investigation of noun phrase complexity in the L2 writings of a first-year composition course, *Journal of English for Academic Purposes*, 38, 14–24.

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing. *Language Testing*, 34(4), 493–511.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, 14, 48–59.

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, 33(2), 149–183.

Staples, S., Gray, B., Biber, D., & Egbert, J. (under review). Writing trajectories of grammatical complexity at the university: Comparing L1 and L2 English writers in BAWE. [Manuscript submitted Nov 2021 to *Applied Linguistics*].

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program, *TESOL Quarterly*, 47, 420–430.

## Teaching pre-service teachers to create corpus-informed materials: The effectiveness of different types of tasks in an e-learning setting

Elen Le Foll
Osnabrück University
elefoll@uos.de

Recent studies confirm that corpora have yet to be widely used in the foreign language classroom (e.g., Callies 2019; Chambers 2019; Kavanagh 2019). We know that teacher training is key to closing the much-discussed research-practice gap (e.g. Mukherjee 2004; Hüttner, Smit & Mehlmauer-Larcher 2009; Breyer 2009); yet such endeavours face numerous challenges (e.g., Breyer 2009; Farr 2008; Leńko-Szymańska 2015; 2017).

The present study reports on the strengths, limitations, and challenges of two iterations of an M.Ed. project-based seminar entitled "Creating corpus-informed teaching materials" as part of which pre-service EFL teachers with no previous experience of working with corpora had the opportunity to contribute to an open access textbook (Le Foll 2021) that aims to empower foreign language teachers to create corpus-informed materials autonomously using online tools and corpora.

These two iterations of the seminar were held exclusively online which means that, in addition to pre- and post-seminar surveys, students' responses to a variety of tasks conducted in an e-learning portfolio could be tracked in detail. The present study triangulates these data to reflect on the effectiveness of different types of tasks designed to support pre-service teachers' development of corpus skills to design teaching and learning corpus-informed materials. Which types of tasks and activities are most effective? In what order are they best presented? To what extent do students react differently to these tasks and activities?

Mixed methods are used to seek answers to these research questions. The results of pre- and post-seminar surveys are compared. The results of the university's official course evaluation questionnaire are also examined. In addition, students' answers to a variety of reflective (e.g., what is 'authentic' language? How can we determine what is 'correct' English?), hands-on (principally using english-corpora.org and Sketch Engine), and creative tasks (students' attempts to design their own corpus-informed materials) are analysed. These tasks comprise both closed and open-ended question types so that both quantitative and qualitative methods are employed.

Preliminary results indicate that tasks must provide considerable scaffolding for them to be effective. The combination of short videos with multiple-choice quizzes that require students to immediately try out their newly acquired corpus skills and provide them with immediate feedback proved to be particularly popular and effective. However, many students struggled with the interpretation of corpus results and, in particular, their pedagogical implications.

This is illustrated in a case-study analysis of a task requiring students to query the Open Cambridge Learner Corpus (Cambridge University Press 2017) and to draw inferences from the results they obtained. Whilst the vast majority of the 48 students who completed the task successfully queried the learner corpus, the analysis of their open-ended answers makes clear that some would need more support to draw meaningful pedagogical conclusions from the data. Thus, when asked to compare the most frequent collocates of the verb EXPLAIN in a subcorpus of the Open Cambridge Learner Corpus and the Spoken British National Corpus 2014 (Love et al. (2017), some students successfully identified the most obvious difference, e.g.:

"A typical learner error seems to be leaving out the necessary 'to' between the verb 'explain' and the pronoun that this construction requires. example: 'explain us' instead of 'explain to us'" (P4). However, some observed the phenomena of interest yet stopped short of drawing any pedagogical conclusions, e.g.:

"B1 learners of English most frequently used the word forms 'explain to' and 'explain you'. Whereas the BNC Spoken Corpus indicates that Brits use the word forms 'explain it' and 'explain to' most frequently" (P18).

The results suggest that the careful scaffolding of tasks and immediate, automated feedback can help learners reach more meaningful conclusions which can ultimately support them in making more informed pedagogical decisions.

In light of these results, the paper concludes by considering how future courses can be improved. It outlines how tasks can most effectively be scaffolded for students of different proficiencies and discusses the potential of such scaffolded e-learning tasks in online, offline, and hybrid instructional settings. In addition, it considers the role that Open Educational Resources (OERs) and OER-enabled pedagogy (Wiley & Hilton 2018)

can play in ensuring that the knowledge and skills gained in such a university seminar are genuinely transferred to students' future teaching practice.

**References**


Breyer, Y. (2009). Learning and teaching with corpora: reflections by student teachers. *Computer Assisted Language Learning*, 22(2), 153–172.

Callies, M. (2019). Integrating corpus literacy into language teacher education: The case of learner corpora. In S. Götz & J. Mukherjee (eds.), *Learner Corpora and Language Teaching*. Amsterdam: John Benjamins, 245–263.

Cambridge University Press. (2017). Open Cambridge Learner English Corpus, available on Sketch Engine. https://www.sketchengine.eu/cambridge-learner-corpus/ (29 November 2019).

Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching*, 52(4), 460–475.

Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness*, 17(1), 25–43.

Hüttner, J, Smit, U. & Mehlmauer-Larcher, B. (2009). ESP teacher education at the interface of theory and practice: Introducing a model of mediated corpus-based genre analysis. *System*, 37(1), 99–109.

Kavanagh, B. (2019). Using 'what already works' to 'bridge the gap' between corpus research and corpora in schools. *Learner Corpus Research Conference*, September 2019, Warsaw.

Le Foll, E. (2021). *Creating Corpus-Informed Materials for the English as a Foreign Language Classroom: A step-by-step guide for (trainee) teachers using online resources (Open Educational Resource (OER)*. 3rd edn. https://elenlefoll.pressbooks.com (30 July 2021).

Leńko-Szymańska, A. (2014). Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education. *ReCALL*, 26(2), 260–278.

Leńko-Szymańska, A. (2015). A teacher-training course on the use of corpora in language education: Perspectives of the students. In A. Turula, B. Mikolajewska & D. Stanulewicz (eds.), *Insights into Technology Enhanced Language Pedagogy*, Berlin: Peter Lang, 129–144.

Leńko-Szymańska, A. (2017). Training teachers in data-driven learning: Tackling the challenge. *Language Learning & Technology*, 21(3), 217–241.

Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics*, 22(3), 319–344.

McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context. *Language Teaching*, 41(4), 563–574.

Mukherjee, J. (2004). Bridging the Gap between Applied Corpus Linguistics and the Reality of English Language Teaching in Germany. In U. Connor & T. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi, 239–250.

Wiley, D. & Hilton III, J. (2018). Defining OER-Enabled Pedagogy. *International Review of Research in Open and Distributed Learning*, 19(4).

## Changing patterns of linking adverbials in L2 university student writing

Joseph J. Lee[1], Robert Bern[2]
Dalarna University[1], Ohio University[2]
jole@du.se[1], rb467019@ohio.edu[2]

Appropriate use of linking adverbials (LAs) is a key feature of successful academic writing because these devices (e.g., *furthermore, however, thus*) enhance meaning and establish textual cohesion explicitly (Shaw 2009). Previous research has shown that LAs appear prominently in academic prose. In fact, these studies have revealed that academic writing includes more LAs than other registers including conversation, fiction, and news (Biber et al. 1999; Liu 2008). Despite their importance in academic writing, second language (L2) writers of English have been reported to struggle to use LAs appropriately. Over the past few decades, considerable research has compared the use of LAs between first-language (L1) English writers and various L2 English groups including L1 Chinese (e.g., Gao 2016), L1 Korean (e.g., Ha 2016), and L1 Spanish writers (e.g., Carrió-Pastor 2013), as well as among specific L1 groups (e.g., Appel & Szeib 2018). These studies have shown that L2 English writers frequently overuse, underuse, and/or misuse these devices. While these studies have been important in understanding L2 writers' challenges with LAs, surprisingly little attention has been given to whether L2 students' use of LAs in their writing changes over time or the degree to which their behaviors change with experience. Using corpus-based methods, this study reports findings of an analysis of the developmental trajectory of English-as-a-second-language (ESL) university students' use of LAs in their academic writing. The study was guided by the following research question: To what extent does L2 university students' use of linking adverbials in their writing change over time? Through this analysis, this study aims to provide a greater understanding of the relationship between educational experience and L2 writing development.

Data consist of a specialized corpus of 126 high-rated source-based argumentative essays written by 63 ESL undergraduate students in US-based first-year writing (FYW) courses at two different points in time. The first subcorpus (ESL-1) includes 63 argumentative essays (66,424 words) written by these students in the first of two FYW courses, while the second subcorpus (ESL-2) consists of 63 argumentative papers (87,638 words) written by the same student writers in the second FYW course. To analyze LAs in the student essays, Liu's (2008) taxonomy of LAs was used because his list is considered to be one of the most comprehensive (Gao 2016), with a total of 110 lexical items. The framework consists of four broad semantic categories: additive (e.g., *additionally*, *similarly*), adversative (e.g., *however*, *in contrast*), causal (e.g., *as a result*, *hence*), and sequential (e.g., *first, in conclusion*). Each category in this framework is classified further into subcategories. Using the concordance tool *Antconc* (Anthony 2018), every LA item in Liu's (2008) list was searched in both subcorpora, and then we manually examined each example in its textual context to ensure every item functioned as an LA. Item frequencies were counted per text and normalized per 1,000 words. To determine whether the differences were statistically significant, paired samples t-tests, with Bonferroni correction, were performed, with the alpha set at .05 (two-tailed).

Analysis reveals statistically significant changes in the overall frequency of LAs, with the ESL-2 subcorpus consisting of fewer LAs than the ESL-1 subcorpus. Upon closer analysis, the results show that the use of additive and causal LAs decreased over time, while adversative and sequential LAs increased. However, a statistically significant difference was only found for the additive category. Analysis of the proportional distributions of the categories shows that with experience ESL student writers rely less on additive and more on adversative, yet the distributions of causal and sequential do not seem to change. With a few exceptions, the most frequently used words/phrases for all the categories are strikingly similar in both subcorpora, though the frequencies at which they are used changes. Thus, the preliminary findings suggest that the distribution of LAs appears to change and matches more closely with published academic prose (cf. Liu 2008) as ESL students gain more experience with academic writing; however, the specific linguistic LA devices used do not seem to markedly change. The paper begins by reporting and discussing the results, followed by implications for L2 writing research and pedagogy.

**References**


Anthony, L. (2018). *AntConc* (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/antconc/

Appel, R., & Szeib, A. (2018). Linking adverbials in L2 English academic writing: L1-related differences. *System*, 78, 115-129.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Carrió-Pastor, M. L. (2013). A contrastive study of the variation of sentence connectors in academic English. *Journal of English for Academic Purposes*, 12, 192-202.

Gao, X. (2016). A cross-disciplinary corpus-based study on English and Chinese native speakers' use of linking adverbials in academic writing. *Journal of English for Academic Purposes*, 24, 14-28.

Ha, M.-J. (2016). Linking adverbials in first-year Korean university EFL learners' writing: A corpus-informed analysis. *Computer Assisted Language Learning*, 29, 1090-1101.

Liu, D. (2008). Linking adverbials: An across-register corpus study and its implications. *International Journal of Corpus Linguistics*, 13, 491-518.

Shaw, P. (2009). Linking adverbials in student and professional writing in literary studies: What makes writing mature. In M. Charles, D. Pecorari & S. Hunston (Eds.), *Academic Writing: At the Interface of Corpus and Discourse*. London: Continuum, 215-235.

## Phraseology in the assessment of L2 writing

Agnieszka Leńko-Szymańska[1], Piotr Pęzik[2], Michał Adamczyk[3]
University of Warsaw[1], University of Łodź[2,3]
a.lenko@uw.edu.pl[1], piotr.pezik@uni.lodz.pl[2], michal.adamczyk@uni.lodz.pl[3]

The last two decades have witnessed a surge of interest in the role of phraseological competence in second language acquisition and assessment (cf. Wray 2002; Meunier & Granger, eds. 2008; Leńko-Szymańska 2020). It is believed that broadly understood formulaic language has a great influence on the perception of the quality of L2 production. Learner corpus methodology in particular offers an extensive range of data and tools to investigate the development and use of L2 phraseology both quantitively and qualitatively (e.g., Bestgen & Granger 2014; Paquot 2019). This presentation aims at exploring the link between phraseological complexity and raters' assessment of L2 writing at the B2 level.

The data used in this study were 497 argumentative essays on the same topic, randomly selected from a pool of over 2200 scripts in a high-stake English certification exam at the B2 level. The essays were evaluated holistically by 5 tandems of raters in the exam's regular marking procedure and after three months evaluated again using an analytical rubric with four marking categories: content, organization, accuracy, and vocabulary. The learner texts were parsed with an online tool spaCy (https://spacy.io/) and five types of relational collocations were extracted from the corpus: noun + verb, 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, ΔPforward, and ΔPbackward. 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' 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 as the outcome variable. The models were computed for all types of relational collocations jointly and then again only for the adjective + noun collocation type.

The results demonstrated that the predictive power of the models built for all the relational collocations was very low ($R2 = 0.012$ for the holistic marks and $R2 = 0.002$ for the vocabulary marks) and the only statistically significant metric in predicting the holistic marks was the frequency in the reference corpus. The predictive power for the adjectival collocations was slightly higher but still low ($R2 = 0.077$ for the holistic marks and $R2 = 0.090$ for the vocabulary marks), and more indices were statistically significant in the model: the number of items, their BNC frequency, MI and LogDice scores for holistic marks, and the number of items and MI scores for vocabulary marks. The qualitative analysis of selected high-ranking and low-ranking essays demonstrated that learner texts with high scores contain instances of creative word associations which are attested in the reference corpus and potentially "eye-catching" for raters, but their association scores are low (e.g., deep consideration, numerous responsibilities).

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 can be proposed that current methods of capturing phraseological complexity are not intricate enough to capture its rather complex nature and its tricky role in contributing to the perceived quality of L2 writing.

**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. https://doi.org/10.1016/j.jslw.2014.09.004

Leńko-Szymańska, A. (2020). *Defining and Assessing Lexical Proficiency*. New York/London: Routledge.

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

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121–145. https://doi.org/10.1177/0267658317694221

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: CUP.

# Redundancy in subject anaphora resolution: A corpus-based study of L1 Japanese learners of L2 Spanish

Nobuo Ignacio López-Sako[1], Cristóbal Lozano[2]
University of Granada
nilsako@ugr.es[1], cristoballozano@ugr.es[2]

Recent research in anaphora resolution (AR) has revealed that learners of an L2, even at very advanced levels, show deficits in the production of anaphoric forms (typically, null pronouns, overt pronouns, but also full NPs) in subject position at the syntax-discourse interface, as predicted by the **Interface Hypothesis** (IH) (Sorace 2011). IH predicts that L2 learners may find it difficult to reach native-like performance due to the processing overload required to integrate the syntactic form of the anaphor and its discourse-pragmatic context. This non-native-like attainment of AR has been attested in L2 Spanish with different L1 combinations. Some studies have focused on the null vs. overt subject alternation (Lozano 2016, Montrul & Rodríguez-Louro 2006, Rothman 2009). Other studies have tested the ultimate attainment of AR in L2 Spanish among null-subject L1 speakers, e.g., L1 Arabic (Garcia-Alcaraz & Bell 2011), L1 Farsi (Judy 2015), or L1 Greek (Lozano 2018).

Furthermore, using corpus data from CEDEL2 (http://cedel2.learnercorpora.com) (Lozano 2021), it has been shown that not all discourse scenarios are equally problematic in L2 Spanish (Lozano 2016). Particularly, over-explicitness in Topic-Continuity contexts, resulting in redundant or uneconomical uses of anaphors, has been reported to be more frequent than the use of under-explicit forms in Topic-Shift scenarios. This unbalance, which is not predicted by the IH, is accounted for by the **Pragmatic Principles Violation Hypothesis** (PPVH) (Lozano 2016), which articulates the ambiguity-redundancy dichotomy as a continuum, ranging from a mild violation for redundancy to a strong violation for ambiguity, since the latter might result in a communication breakdown. Lozano (2016) formulated the PPVH in order to account for the results obtained in a corpus-based study of L1 English-L2 Spanish, i.e., a non-null-subject language vs. a null-subject language. Later on, Lozano (2018) obtained similar results in an acceptability judgement experimental study with two null-subject languages: L1 Greek-L2 Spanish.

This paper aims at putting the PPVH to the test by using the same corpus as Lozano (2016) (CEDEL2) but pairing two null-subject languages which, to our knowledge, have not been previously tested for AR: L1 Japanese – L2 Spanish. Japanese is a null-subject/topic language and, hence, null pronouns are syntactically licensed in subject/topic position (1), as in Spanish.

1. *Chappurin$_i$ ga roji o aruite iru to, Ø$_i$ michibata ni suterareta akachan o mitsuketa.*
   'Chaplin$_i$ was walking down an alley when [Ø$_i$] found a baby abandoned by the wayside'.
   [CEDEL2 corpus, Japanese native: JP_WR_30_14_AO.txt]

A cross-linguistic facilitating effect might be expected in the realization of AR in the subject position, but the PPVH (and the IH) predicts that the L1-L2 similarity (in terms of AR in the subject position) is no guarantee for a successful performance in L2. Thus, we analysed advanced L1 Japanese-L2 Spanish production data (124 target items) and contrasted it with Spanish native control data (84 target items) regarding AR, including the following variables: topic continuity/shift, number and gender of potential antecedents, and antecedent-anaphor distance, following recent work in this line (Martín-Villena & Lozano 2020; Quesada & Lozano 2020). The data were finely annotated and analysed with UAM Corpus Tool (O'Donnell 2009), accounting for the syntactic and discursive features affecting the choice of pragmatically (in)felicitous anaphors.

As predicted by PPVH, Topic-Continuity results suggest that there is no facilitating effect of the L1 on the production of AR in L2 even when the null-subject feature is shared. By contrast, Topic Shift is less problematic than Topic continuity in the felicitous realization of AR. That is, learners tend to produce significantly more uneconomical anaphors in Topic Continuity scenarios, whereas the informativeness violation (ambiguity) among Japanese learners in Topic Shift scenarios is non-significant. Thus, our findings confirm that the PPVH can be extended to typologically distant L1-L2 combinations (L1 Japanese-L2 Spanish), and the L1 transfer explanation be discarded. Results are in line with findings for other null-subject L1s vs. L2 Spanish (García-Alcaraz & Bell 2011, Lozano 2018).

Additional findings will be discussed, such as the relatively high incidence of uneconomical NPs in Topic-Shift contexts, which might be explained as an effect of Japanese preference for full NPs over explicit

pronouns (Warnick 1991). Finally, a revised version of the PPVH, more finely reflecting the gradience of the pragmatic-principles violations, will be proposed.

**References**


García-Alcaraz, E. & Bel, A. (2011). Selección y distribución de los pronombres en el español L2 de los hablantes de árabe. *Revista de Lingüística y Lenguas Aplicadas* 6: 165-179. http://dx.doi.org/10.4995/rlyla.2011.901

Judy, T. (2015). Knowledge and processing of subject-related discourse properties in L2 near-native speakers of Spanish, L1 Farsi. In T. Judy & S. Perpiñán (Eds.). *The acquisition of Spanish in understudied language pairings*. Amsterdam: John Benjamins, 169-199. https://doi.org/10.1075/ihll.3.07jud

Lozano, C. (2016). Pragmatic principles in anaphora resolution at the syntax-discourse interface: advanced English learners of Spanish in the CEDEL2 corpus. In M. Alonso Ramos (Ed.). *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam: John Benjamins, 236-265. https://doi.org/10.1075/scl.78.09loz

Lozano, C. (2018). The development of anaphora resolution at the syntax-discourse interface: Pronominal subjects in Greek learners of Spanish. *Journal of Psycholinguistic Research* 47: 411-430. https://link.springer.com/article/10.1007%2Fs10936-017-9541-8

Lozano, C. (2021). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research* (online article: 16 October), 1-19. https://doi.org/10.1177/02676583211050522

Martín-Villena, F. & Lozano, C. (2020). Anaphora resolution in topic continuity: evidence from L1 English–L2 Spanish data in the CEDEL2 corpus. In J. Ryan & P. Crosthwaite (Eds.). *Referring in a second language*. New York: Routledge, 119-141. http://doi.org/10.4324/9780429263972-7

Montrul, S. & Rodríguez-Louro, C. (2006). Beyond the syntax of the Null Subject Parameter: A look at the discourse-pragmatic distribution of null and overt subjects by L2 learners of Spanish. In V. Torrens & L. Escobar (Eds.). *The Acquisition of Syntax in Romance Languages*. Amsterdam: John Benjamins, 401-418. https://doi.org/10.1075/lald.41.19mon

O'Donnell, M. (2009). The UAM Corpus Tool: Software for corpus annotation and exploration. In C.M. Bretones et al. (Eds). *Applied Linguistics Now: Understanding Language and Mind/La lingüística aplicada actual: Comprendiendo el lenguaje y la mente*. Almería: Universidad de Almería, 1433-1447.

Quesada, T., & Lozano, C. (2020). Which factors determine the choice of referential expressions in L2 English discourse?: New evidence from the COREFL corpus. *Studies in Second Language Acquisition*, 42(5), 959-986. https://doi.org/10.1017/S0272263120000224

Rothman, J. (2009). Pragmatic deficits with syntactic consequences?: L2 pronominal subjects and the syntax-pragmatics interface. *Journal of Pragmatics* 41(5): 951-973. https://doi.org/10.1016/j.pragma.2008.07.007

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism* 1(1): 1-33. https://doi.org/10.1075/lab.1.1.01sor

Warnick, P. (1991). The use of personal pronouns in the language of learners of Japanese as a second language. *Proceedings of the 17th Deseret language and linguistic society symposium*, Vol. 17(1), Article 16: 109-121.

## Does mode affect referring expression selection? A corpus-based study of advanced L1 Spanish-L2 English narratives

Jorge Montaño[1], Ana Díaz-Negrillo[2]
Universidad de Granada
jorgemont@correo.ugr.es[1], anadiaznegrillo@ugr.es[2]

In their production of narratives, speakers are bound to refer to entities mentioned in earlier parts of their discourse by selecting specific anaphoric referring expressions (REs) in a phenomenon known as anaphora resolution (AR). Crucially, research has shown that second language (L2) learners show deficits in the acquisition of the syntactic and pragmatic principles that govern AR. More specifically, L2 learners tend to show an overexplicit selection of REs in topic continuity (cf. *inter alia*, Crosthwaite 2011, Hendriks 2003, Kang 2004, Leclercq & Lenart 2013, Quesada & Lozano 2020, Ryan 2015). However, despite the growing interest in the linguistic and extralinguistic factors that affect the acquisition of AR in L2 learners (for an overview, see Quesada & Lozano 2020), AR research has mostly ignored the potential role of mode, as there is no single study on AR in L2 English that simultaneously analyses both written and spoken data.

In this corpus-based study, we explore the effects of mode on AR by comparing written and spoken discourse and analysing previously studied factors constraining referential selection (i.e., information status, coordination, character status, potential antecedents). Given the exploratory nature of this study, the main factors analysed in Quesada and Lozano (2020) were included to test whether L2 deficits remained constant in different factor-mode combinations or not. Thus, we formulated the following research questions:

RQ1: To what extent do advanced L2 learners match natives' choice of REs when constrained by information status? Does medium play a role in the choice of REs when constrained by information status?

RQ2: What is the effect of coordinate clauses on L2 learners' and natives' production of REs? Does medium play a role in the production of REs in coordinate clauses?

RQ3: What is the effect of character status on L2 learners' and natives' selection of REs? Does medium play a role in the production of REs when constrained by character status?

RQ4: What is the effect of potential antecedents on L2 learners' and natives' use of REs? Does medium play a role in the production of REs when constrained by the number of potential antecedents?

To answer these questions, we analyse the production of third-person singular subject REs of lower-advanced (C1) L1 Spanish-L2 English learners and compare them to a control group of English native speakers from the COREFL corpus (Lozano et al. 2020). Following previous research (Jarvis 2002, Ryan 2015), the production of written and spoken narratives were elicited by a film-retelling task based on Charles Chaplin's film The Kid. The resulting data was tagged using a fine-grained tagset created in UAM CorpusTool (O'Donnell 2009), which was modelled after the one used in Lozano (2016). Descriptive and inferential ($\chi^2$) statistics were later applied to raw frequencies using the statistical tool in UAM CorpusTool.

Results on the effect of information status show that L2 learners and natives have a similar production pattern in new introductions and topic shift contexts. However, they differ in topic continuity contexts, where learners show an overexplicit selection of REs. Crucially, these differences are only present in the spoken data. Given the interaction between information status and the remaining constraining factors (i.e., coordination, character status, potential antecedents), it is not surprising that they show a similar pattern when information status is included in the analysis. In other words, no clear effects were seen for coordinate clauses, character status, or the number of potential antecedents in new introductions or topic shift contexts. Thus, over-explicitness was only found in learners' spoken narratives in topic continuity scenarios as a by-product of including information status in our analysis.

Overall, these results suggest a mode effect on learners' REs selection and production. That is, in topic continuity learners are more overexplicit in their spoken narratives, while natives' selection of REs remains constant regardless of the factor being analysed. These results can be interpreted in light of the advantages of L2 linguistic processing in the written mode, which could account for the native-like selection of REs in the written mode.

**References**


Bel, A., Perera, J., & Salas, N. (2010). Anaphoric devices in written and spoken narrative discourse: data from Catalan. *Written Language & Literacy*, 13(2), 236-259.

Crosthwaite, P. (2011). The effect of collaboration on the cohesion and coherence of L2 narrative discourse between English NS and Korean L2 English users. *Asian EFL Journal*, 13(4), 135-166.

Hendriks, H. (2003). Using nouns for reference maintenance: a seeming contradiction in L2 discourse. In A. Giacalone (Ed.), *Typology and Second Language Acquisition*. Berlin: Mouton De Gruyter, 291-326.

Jarvis, S. (2002). Topic continuity in L2 English article use. *Studies in Second Language Acquisition*, 24(3), 387-418.

Kang, J. Y. (2004). Telling a coherent story in a foreign language: analysis of Korean EFL learners' referential strategies in oral narrative discourse. *Journal of Pragmatics*, 36(11), 1975-1990.

Leclercq, P., & Lenart, E. (2013). Discourse cohesion and accessibility of referents in oral narratives: a comparison of L1 and L2 acquisition of French and English. *Discours*, 12.

Lozano, C. (2016). Pragmatic principles in anaphora resolution at the syntax-discourse interface: advanced English learners of Spanish in the CEDEL2 corpus. In M. Alonso-Ramos (Ed.), *Spanish Learner Corpus Research: Current trends and future perspectives*. Amsterdam: John Benjamins Publishing Company, 235-265.

Lozano, C., Díaz-Negrillo, A., & Callies, M. (2020). 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*. Bern: Peter Lang, 21-46.

Ngo, B., Kaiser, E., & Simpson, A. (2019). Effects of grammatical roles and parallelism on referential form production in Vietnamese spoken and written narratives. In T. Duffield, T. Phan, & T. Trinh (Eds.), *Interdisciplinary Perspectives on Vietnamese Linguistics*. Amsterdam: John Benjamins, 211-275.

O'Donnell, M. (2009). The UAM CorpusTool: Software for corpus annotation and exploration. In C. M. Bretones Callejas, J. F. Fernández Sánchez, J. R. Ibáñez Ibáñez, M. E. García Sánchez, M. E. Cortés de los Ríos, M. S. Salaberri Ramiro, M. S. Cruz Martínez, N. A. Perdu Honeyman, & B. Cantizano Márquez (Eds.), *Applied linguistics now: Understanding language and mind/La lingüística aplicada actual: Comprendiendo el lenguaje y la mente*. Almería: Universidad de Almería, 1433-1447.

Quesada, T., & Lozano, C. (2020). Which factors determine the choice of referential expressions in L1 English discourse? New evidence from the COREFL corpus. *Studies in Second Language Acquisition*, 42(5), 959-986.

Ryan, J. (2015). Overexplicit referent tracking in L2 English: strategy, avoidance, or myth? *Language Learning*, 65(4), 824-859.

# The influence of L1 typology on the acquisition of the L2 English articles: A large-scale learner corpus study

Dogus Can Oksuz[1], Dora Alexopoulou[2], Kate Derkach[3], Ianthi Maria[4]
University of Cambridge
dco24@cam.ac.uk[1], ta259@cam.ac.uk[2], kate.derkach17@gmail.com[3], imt20@cam.ac.uk[4]

**Abstract**
Linguistic distances between learners' L1s and L2 sheds light on L2 learnability, as well as how far an L1 facilitates or impedes the learning of an L2 (Schepens, van der Slik & van Hout, 2016). The smaller the linguistic distance between the two, the easier it is to learn an L2. There is compelling empirical evidence that the linguistic distance including numerous lexical and morphosyntactic features, between learners' L1s and L2s/L3s predicted L2/L3 speaking proficiency scores in Dutch (Schepens et al., 2016). One crucial question is how linguistic distances might affect the acquisition of individual features like articles, rather than broad outcomes like proficiency. Does the acquisition of individual features depend solely on the presence/absence of a congruent element in the L1 (e.g. Murakami & Alexopoulou, 2016), or do broader typological differences guide how learners approach the input, influencing their acquisition? In this talk, we investigate the L1 influence on article acquisition through the lenses of the presence/absence of a congruent element in the L1 and linguistic distances between L1-L2.

**Research questions**
1. Is learner accuracy in the use of L2 English articles linked to a) the absence or presence of congruent forms in learners' L1s? b) the linguistic distance between learners' L1s and L2?
2. Does L1-L2 linguistic distance and/or presence/absence of a congruent element affect learner accuracy in the use of definite and indefinite articles similarly?

**Method**
Data. We used a subset of the EF-Cambridge Open Language Database (Alexopoulou, Geertzen, Korhonen & Meurers, 2015), a written learner error-tagged corpus of English, which was 34 million words including 527,758 different scripts written by 104,541 learners. We targeted 11 native languages (Portuguese, Chinese, German, French, Italian, Japanese, Arabic, Russian, Mexican Spanish, Korean and Turkish, with proficiency levels from A1 to B2) providing a typologically diverse set for comparison. R scripts were written to convert error-tagged texts to corrected texts. Obligatory contexts were defined as article use in corrected texts. Using R scripts we counted the number of obligatory contexts and each type of error. For instance, if a learner wrote, *She is wearing black t-shirt* and it was corrected to *She is wearing a black t-shirt*, this was classified as an omission error. As a measure of accuracy, we used the ratio between the number of correct uses and obligatory contexts.
Measuring the linguistic distance. We compared the binary classification of the presence/absence of articles in learners' L1s with continuous lexical (Shatz, 2022) and syntactic distance scores in the nominal domain (Ceolin Guardiano, Irimia, & Longobardi, 2020). The lexical distance scores are based on the Levenshtein Distance, and syntactic distance scores are based on the Parametric Comparison Method, which examines similarities and distances of properties.

**Results**
Mixed-effects regression modelling revealed that L1 typology affected L2 learners' accuracy. We found that learners whose L1s have articles used both types of articles more accurately than learners whose L1s do not. Overall, article accuracy is higher in learners with higher proficiency. However, this effect was stronger for learners whose L1s have articles. The accuracy increase over proficiency was smaller in definite than indefinite articles. The linguistic distance scores showed weaker correlations with accuracy scores. In sum, L1 influence is clearly observable in the acquisition of articles. The findings generally confirm the effect of proficiency and L1 typology reported in Murakami and Alexopoulou (2016) for a different set of L1s and in a different corpus. We are currently considering further measures of morphosyntax to shed light on what impacts the acquisition of articles beyond the presence/absence of congruent forms.

**References**

Alexopoulou T., Geertzen, J., Korhonen, A., & Meurers, D. (2015). Exploring big educational corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research* 1(1), 96-129.

Ceolin, A., Guardiano, C., Irimia, M., and Longobardi, G. (2020). Formal Syntax and Deep History. *Frontiers in Psychology* 11, 488871.

Murakami, A. & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition* 38(3), 365-401.

Schepens, J., van der Slik, F., & van Hout R. (2016). L1 and L2 distance effects in learning L3 Dutch. *Language Learning* 66(1), 224-256.

# Introducing the CLAP project: Adaptive comparative judgment as a community-based solution for enriching learner corpora with crowdsourced L2 proficiency assessment

Magali Paquot[1], Rachel Rubin[2], Nathan Vandeweerd[3]
FNRS[1], Centre for English Corpus Linguistics – UCLouvain[2], Vrije Universiteit Brussel[3]
magali.paquot@uclouvain.be[1], Rachel.Rubin@vub.be[2], nathan.vandeweerd@uclouvain.be[3]

Though proficiency is one of the most important constructs in Second Language Research, its measurement has not always received the attention it deserves, and practices of proficiency level assignment have been the subject of continued criticism (e.g. Hulstijn et al. 2010). In Learner Corpus Research, more particularly, Carlsen's (2012) review of some of the most commonly used methods of proficiency-level assignments of texts showed that many learner corpora still rely on variables such as institutional status or year of study as a proxy for proficiency, despite the fact that these external criteria are largely regarded as unreliable (Thomas 1994). While a handful of learner corpora prove the exception by including reliable text-based proficiency scores (e.g. AndreSpråksKorpus, a learner corpus of Norwegian as a second language; Carlsen 2012), the time and cost difficulties typically associated with analytical scoring means that it is often absent from, or operationalized in unreliable ways, in learner corpora.

This presentation has two main objectives. First, we will introduce the technique of adaptive comparative judgment (ACJ), coupled with a crowdsourcing approach, as a practical solution to the reliability issues as well as the time and cost difficulties associated with a text-based approach to proficiency assessment in learner corpus research. The method of CJ is based on Thurstone's (1927) 'Law of Comparative Judgment', which builds on the assumption that people are able to compare two performances more easily and reliably than to assign a score to individual performance (Lesterhuis et al. 2017). The CJ approach involves the consensus of a panel of judges who are asked to compare two performances of any kind (be they dance performances, design portfolios, or, as in the present study, written learner productions) and to simply decide which of them is better. Under the ACJ framework, performances are paired adaptively, reducing the overall amount of comparisons required to achieve a reliable scale of performance abilities. A second critical assumption underpinning CJ is its reliance on holistic judgment: Judges do not receive criteria to guide their judgment process, but at best a general description regarding the competence to be assessed. We showcase this method by reporting on the methodological framework implemented in the CLAP project and presenting the results of a first pilot study that demonstrate that a crowd of 43 judges is able to assess (i.e. rank) 50 learner texts with high reliability (SSR = .95). No effect of language skills or language assessment experience was found on the assessment task, but there was a difference in the decisions made by judges who received formal language assessment training and those who did not. Nevertheless, the scores generated by the crowdsourced task exhibited a strong correlation with the rubric-based scores released with the learner corpus used (ETS Corpus of Non-Native Written English; Blanchard et al., 2014).

The second objective of this presentation is to launch a collaborative initiative that aims to replicate and extend the pilot study described above by addressing some of the most pressing theoretical issues and avenues for future L2 research identified therein (Paquot et al. forthcoming). To that end, the project will be guided by the following three main research questions:

- RQ1. To what extent do specific characteristics of learner texts (topic, length, homogeneity in terms of proficiency) have an effect on the reliability of an ACJ task?
- RQ2. To what extent do specific characteristics of judges (language assessment training and expertise) have an effect on the reliability of an ACJ task?
- RQ3. To what extent do specific characteristics of learner texts and characteristics of judges have an effect on the validity of an ACJ task?

By the end of the project, we will be in a position to provide guidelines about the conditions in which the ACJ method can be used to enrich L2 data, with the hope that colleagues will replicate our work on other learner corpora, including learner corpora of other L2s than English. We will also distribute the comparative rank order for a set of ICLE texts as an open access resource. However, for this project to be successful, we will need to recruit a large crowd of judges. We hope that LCR participants will be as enthusiastic about this project as we are and contribute!

**References**


Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & M. Chodorow (2014). *ETS corpus of non-native written English*. LDC2014T06. Philadelphia: Linguistic Data Consortium, 2014. https://catalog.ldc.upenn.edu/LDC2014T06

Carlsen, C. (2012). Proficiency level – a fuzzy variable in computer learner corpora. *Applied Linguistics*, 33, 161–183.

Hulstijn, J., Alderson, C. & R. Schroonen (2010). Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them? In Bartning, I., Martin, M. & I. Vedder (Eds.). *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research*. EUROSLA Monographs Series 1, 11-20.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & S. De Maeyer (2017). Comparative judgement as a promising alternative to score competences. In Cano, E. & G. Ion (Eds.). *Innovative Practices for Higher Education Assessment and Measurement*. IGI Global, 119-138.

Paquot, M., Rubin, R. & N. Vandeweerd (forthcoming). Crowdsourced Adaptive Comparative Judgment: A community-based solution for proficiency rating. *Language Learning*.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, 44, 307–336.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273-86.

# "Let's say maybe it's our Italian culture": Expressions of uncertainty in Italian learners of English

Francesca Poli
Università Cattolica del Sacro Cuore
francesca.poli@unicatt.it

Interactive communication leads to the production of linguistic items named *expressions of uncertainty* (ExU) as the speaker attempts to generate responses and adjust their stance. For L2 learners, spoken language may be more challenging as they are required to maintain the conversation flowing during rapidly developing discourse (Gablasova et al. 2017). Thus, ExU may either reflect the learners' uncertainty about the truth of a sentence or their language and speech. This study compares the use of ExU in spoken English in a group of Italian learners and native English speakers (NS). The purpose is to explore whether learners' ExU differ from those of NS following Granger's Contrastive Interlanguage Analysis (2015). The study relies on a recently compiled Italian spoken learner corpus and applies a partly corpus-driven and corpus-based approach. The work addresses the following research question: are there any differences in the frequency and/or type of ExU produced by Italian speakers and NS of English?

In conversation, we not only communicate propositions but also attitudes towards them. The notion of stance, i.e., the expression of "attitudes, thoughts, and feelings of the speaker" (Biber et al. 1999: 966), and the concept of epistemic modality, which is defined as the speaker's judgements or assumptions about the factual status of a proposition (Coates 1987) are strongly linked to the idea of expressing uncertainty. Research into learners' spoken ExU is intriguing since expressing commitment to an assertion requires significant pragmatic skills (Holmes 1982) and studies have demonstrated that even advanced learners have a limited pragmatic repertoire (Romero-Trillo 2018). Although research has found differences between learners and NS regarding the use of adverbs of certainty (Perez-Paredes & Camino Bueno-Alastuey 2019), the pragmatics of spoken communication remains generally under-investigated (Gablasova et al. 2017).

Following Callies' urge (2015) for more corpus-driven research, n-grams were first extracted from the two corpora: the Italian Spoken Learner Corpus (ISLC) (Author 2020) and the native-speaker reference corpus LOCNEC (De Cock 2004). The ISLC contains data from ≥ C1 learners of English. The minimum n-gram size for the extraction was set to min. two and max. five; the frequency threshold was set to five and the minimum distribution to three. This yielded an inventory of approximately 10,000 n-grams overall which were manually sorted and cleaned of any irrelevant occurrences (e.g., *but I*, *overlap and*) resulting in 12 simplified expressions: *I think, I don't think, I'm not sure, I don't know, I would say, I guess, I suppose, maybe, probably, perhaps, let's say, how can I say*. The final dataset included the relative frequency per 100,000 words for each of the expressions for the Italian and NS. To address the RQ, Wilcoxon rank sum tests followed by effect size calculations were carried out in R.

The results are mixed: aside from *I think, I don't know, I'm not sure, I would say, I guess* which were not statistically different, there was significant difference between the groups in *I suppose* ($W = 1400$, $p < .0001$, $r = -0.62$), *maybe* ($W = 112$, $p < .0001$, $r = -0.74$), *probably* ($W = 478$, $p < .001$, $r = -0.37$), *perhaps* ($W = 556.5$, $p = 0.003$, $r = -0.33$), *let's say* ($W = 675$, $p < .001$, $r = -0.36$), and *how can I say* ($W = 750$, $p = 0.014$, $r = -0.27$). The Italian learners overuse *maybe, probably, perhaps, let's say* and *how can I say*, while they tend to use fewer instances of *I don't think* and *I suppose*.

The learners display a higher degree of uncertainty compared to the NS as the results demonstrate the overuse of at least four ExU. The underuse of *I don't think* could be traced back to the L1, which usually avoids the use of negatives with epistemic modality, while a lack of exposure to typically British input may have resulted in the poor mastery of *I suppose*. Although additional (L1 contrastive) research is needed to better frame this pattern, it could be hypothesised that Italians show greater uncertainty in English despite their advanced proficiency and high level of Uncertainty Avoidance in the Italian culture (Hofstede 2010). However, closer scrutiny is needed to rule out other phenomena such as pragmatic fossilization (Romero-Trillo 2018) or personalisation of talk (Baumgarten & House 2010).

**References**


Author (2020). *Adverb + adjective collocations in a spoken learner corpus: A quantitative and qualitative approach* [unpublished doctoral dissertation]. Milan: Università Cattolica del Sacro Cuore.

Baumgarten, N., & House, J. (2010). I think and I don't know in English as lingua franca and native English discourse. *Journal of Pragmatics*, 42, 1184–1200.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & R. Quirk (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP, 35–55.

Coates, J. (1987). Epistemic Modality and Spoken Discourse. *Transactions of the Philological Society*, 85, 110–131.

De Cock, S. (2004). Preferred Sequences of Words in NS and NNS Speech. *Belgian Journal of English Language and Literatures (BELL)*, (New Series 2), 225–246.

Gablasova, D., Brezina, V., McEnery, T. & E. Boyd (2017). Epistemic Stance in Spoken L2 English: The Effect of Task and Speaker Style. *Applied Linguistics*, 38(5), 613–637.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7–24.

Hofstede G., Hofstede G. J. & M. Minkov (2010). *Cultures and Organizations: Software of the Mind, Intercultural Cooperation and its Importance for Survival* [3rd edition]. New York: McGraw-Hill.

Holmes, J. (1982). Expressing doubt and certainty in English. *RELC Journal*, 13(2), 9–28.

Perez-Paredes, P. & Camino Bueno-Alastuey, M. (2019). A corpus-driven analysis of certainty stance adverbs: Obviously, really and actually in spoken native and learner English. *Journal of Pragmatics*, 140, 22–32.

Romero-Trillo, J. (2018). Corpus Pragmatics and Second Language Pragmatics: A Mutualistic Entente in Theory and Practice. *International Journal of Corpus Linguistics and Pragmatics*, 2(2), 113–127.

## F0 range in L2 discourse as evidence for the existence of a prosody interlanguage system

Karin Puga
Justus Liebig University Giessen
karin.puga@anglistik.uni-giessen.de

The concept of interlanguage, coined by Selinker in 1972, has attracted immense scholarly attention, particularly since Granger's (2015) Contrastive Interlanguage Analysis was published. The investigation of the interlanguage systems of learners from different L1 backgrounds has covered many linguistic areas, e.g. phraseology, syntax, and pragmatics. However, research into interlanguage prosody in general and the f0 range, in particular, has remained an exception. Overall, the fundamental frequency (f0) range in L2 speech is narrower than in L1 English speech, irrespective of the learners' L1, speaking style, and speech function (e.g. Ramírez-Verdugo 2022; Gut 2009; Volín et al. 2015). Many scholars attribute their results to L1 influence, uncertainty, or a lack of confidence, and other explanations are rarely offered. L1 influence is postulated because learners often produce an f0 with intermediate values between their own L1 and those of native speakers, their f0 span deviating more greatly than their f0 level. The present study seeks to answer the following research questions:

1. Is f0 range always narrower in L2 speech?
2. Are there alternative interpretations of a deviating f0 range?

A mixed-methods approach and a multivariate analysis are adopted in the examination of L1 (n=90) and L2 data (n=135). The database consists of prosodically annotated versions of the Czech, German, and Spanish components of LINDSEI, alongside British (LOCNEC) and American English (NWSP & New South Voices Collection (NSV)) control corpora. Using an autosegmental-metrical approach (Beckman & Pierrehumbert 1986), the study investigates acoustic properties of the f0 range (level and span) of declarative utterances extracted from spontaneous speech (dialogic and monologic) on similar topics (country traveled to, a special experience, and movie description). Although the NSV includes peer-to-peer interactions with a more narrative style, the L1 groups do not show large deviations from each other in terms of f0 range. Regression modeling was used to predict the f0 range of tune patterns by L1/L2 speaker groups and to investigate the effect of several (extra)linguistic factors, e.g. gender, L2 proficiency (B1-C2: based on post-hoc ratings by Huang et al. 2018), duration of stay abroad, speaking style, and intermediate phrase length in the f0 range.

The results show that, while learners approximate their targets for high-low tunes (a high pitch accent at the beginning of an intermediate phrase ending in a low tone) at the f0 level, they produce a significantly narrower f0 span for the same tunes (-0.7 to -1.9 semitones). Further significant tune-based differences in L2 speech are higher and wider high-ending tunes (1-2 semitones).

A combination of (extra)linguistic variables explains the results; for instance, female L2 speech deviates more than male L2 speech, and the longer the intermediate phrases, the higher and wider the f0. L1 influence cannot be ruled out as a factor determining the narrower f0 in high-low tunes (underhitting) and higher and wider f0 in low-high tunes (overhitting). However, L2 proficiency levels seem to be more revealing; all the learners manifested similar trends, but deviation from native speakers was more pronounced in the lower-proficiency learner group. Besides signaling insecurity, the extremely high f0 range produced in high-ending tunes in L2 speech was also found to fulfill a discourse management function to possibly compensate for weaknesses in intonational phrasing and to make cohesion between intonation units more explicit. These results seem to support my claim for a prosody interlanguage system that is characterized by prosodic drift (overhitting and underhitting).

**References**

Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3(1), 255-309.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24.

Gut, U. (2009). *Non-Native Speech: A Corpus-Based Analysis of Phonological and Phonetic Properties of L2 English and German*. Frankfurt: Peter Lang.

Huang, L.-F., Kubelec, S., Keng, N. & Hsu, L.-H. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. *Language Testing in Asia*, 8(14), 1-17.

Ramírez-Verdugo, M. D. (2022). *Intonation in L2 Discourse. Research Insights*. New York: Routledge Studies in Applied Linguistics.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(3), 209-231.
Volín, J., Poesová, K., & Weingartová, L. (2015). Speech melody properties in English, Czech and Czech English: Reference and interference. *Research in Language*, 13(1), 107-123.

# Using two comparable learner corpora to investigate the production of referring expressions bidirectionally: L1 Spanish-L2 English vs. L1 English-L2 Spanish

Teresa Quesada[1], Cristóbal Lozano[2]
Universidad de Granada
teresaquesada@ugr.es[1], cristoballozano@ugr.es[2]

The production of referring expressions (REs) in discourse (e.g., null/overt pronouns and repeated noun phrases in subject position) is constrained by different factors such as the type of language (i.e., null vs. non-null subject languages), the information status (i.e., topic continuity and topic-shift), or the number of activated antecedents, amongst other factors (inter alia: Huang 2000; Lozano 2021a; M. L. Quesada 2015). The L2 English and L2 Spanish literature show that the acquisition of this phenomenon is difficult for L2 learners (L2ers) because they are overexplicit/redundant (i.e., they produce fuller REs than pragmatically required) (inter alia: Blackwell & Quesada 2012; Lozano 2016; T. Quesada & Lozano 2020; Ryan 2015), as shown in (1).

(1) *El hombre$_i$ está caminando alrededor de la ciudad. El hombre$_i$ encuentra un bebé$_j$. El hombre$_i$ trata encontrar la madre$_k$.* [EN_WR_17_20_2.5_14_EO] 'The man$_i$ is walking around the city. The man$_i$ finds a baby$_j$. The man$_i$ tries to find the mother$_k$'.

The cause of L2ers' redundancy is controversial and far from settled. To our knowledge, previous studies have not investigated this phenomenon both bidirectionally (i.e., L1 Spanish-L2 English vs. L1 English-L2 Spanish) and developmentally (i.e., across proficiency levels). Our aim is to ascertain whether: i) all factors are equally problematic for L2ers; ii) proficiency level and language pair modulate the choice of RE; iii) L2ers' redundancy strategy is eventually overcome by showing native-like attainment, and iv) there are cross-linguistic effects.

We used two written learner corpora: COREFL (Corpus of English as a Foreign Language) (Lozano et al. 2021) and CEDEL2 (*Corpus Escrito del Español como L2*) (Lozano 2021b). We analysed the written production of L1 Spanish-L2 English and L1 English-L2 Spanish adult L2ers across proficiency levels (A2-C2) plus two control groups of English and Spanish natives (N= 152 texts) based on a silent film-retell task (Chaplin video). We tagged the multiple factors constraining the production of REs via a linguistically-informed and theoretically-motivated tagset based on previous work (Lozano 2016; T. Quesada & Lozano 2020).

Results show that not all factors are equally problematic for L2ers and this also depends on the language pair. Considering the information-status factor, results reveal that L2 English L2ers are less redundant in topic-continuity contexts than L2 Spanish L2ers, while all groups are more felicitous in topic-shift contexts. L2 English L2ers are initially redundant because they produce more overt but less null pronouns in topic-continuity at beginner and intermediate levels, whereas English natives show a higher production of null pronouns. But these differences amongst the L2 English L2ers are less marked than amongst L2 Spanish L2ers, as the latter start showing lower rates of felicitous null pronouns in topic-continuity contexts compared to their high production by Spanish natives. Crucially, native-like attainment is eventually feasible in a particular context in very-advanced L2 English L2ers but is not observed in very-advanced L2 Spanish L2ers. Additionally, L2 Spanish L2ers show cross-linguistic influence in topic-continuity contexts as they only use null pronouns in topic-continuity and coordinate contexts, whereas their L1 English allows null pronouns, as in (2). By contrast, L2 English L2ers know the regulations of null pronouns in the L2 and do not show cross-linguistic effects. Finally, the number of antecedents factor affects the production of REs equally across language pairs and proficiency levels, so it seems to be a universal factor.

(2) *…el hombre$_i$ recibe el niño$_j$ y **Ø$_i$** camina. Durante, **el hombre$_i$** camina in la calle con el niño$_j$, **Ø$_i$** ve otra el hombre$_k$. **Él$_i$** da el niño$_j$...* [EN_WR_23_21_3_14_NWH] '…the man$_i$ takes the baby$_j$ and **Ø$_i$** walks. In the meantime, **the man$_i$** walks in the street with the baby$_j$, **Ø$_i$** sees another man$_k$. **He$_i$** gives the baby$_j$…'

In short, L2 English and L2 Spanish L2ers are more redundant than ambiguous from beginner levels. These results are in line with the Pragmatic Principle Violation Hypothesis (PPVH) (Lozano 2016) as we show that the informativeness/economy principle is more frequently violated than the clarity/manner principle. Crucially, this study adds new insights into the PPVH because i) we include two language pairs and different proficiency levels; and ii) we incorporate different factors to the pragmatic scale. This allowed us to reveal that L2ers show a developmental acquisition of REs because the higher the competence in the L2, the less redundant they are, but not all of them achieve native-like competence.

**References**


Blackwell, S. E., & Quesada, M. L. (2012). Third-Person Subjects in Native Speakers' and L2 Learners' Narratives: Testing (and Revising) the Givenness Hierarchy for Spanish. *Selected Proceedings of the 14th Hispanic Linguistics Symposium*, 142-164.

Huang, Y. (2000). Discourse anaphora: Four theoretical models. *Journal of Pragmatics*, *32*(2), 151-176. https://doi.org/10.1016/S0378-2166(99)00041-7

Lozano, C. (2016). Pragmatic principles in anaphora resolution at the syntax-discourse interface: Advanced English learners of Spanish in the CEDEL2 corpus. En M. Alonso-Ramos (Ed.), *Spanish Learner Corpus Research: State of the Art and Perspectives* (pp. 236-265). John Benjamins Publishing Company. https://doi.org/10.1075/scl.78.09loz

Lozano, C. (2021a). Anaphora Resolution in Second Language Acquisition. En *Oxford Bibliographies in Linguistics*. Oxford University Press. https://www.oxfordbibliographies.com/

Lozano, C. (2021b). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, *online first*. https://doi.org/10.1177/02676583211050522

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

Quesada, M. L. (2015). The L2 Acquisition of Spanish Subjects: Multiple Perspectives. En *The L2 Acquisition of Spanish Subjects*. De Gruyter Mouton. https://www.degruyter.com/view/title/304558

Quesada, T., & Lozano, C. (2020). Which factors determine the choice of referential expressions in L2 english discourse?: New evidence from the COREFL corpus. *Studies in Second Language Acquisition*, 1-28. https://doi.org/10.1017/S0272263120000224

Ryan, J. (2015). Overexplicit Referent Tracking in L2 English: Strategy, Avoidance, or Myth? *Language Learning*, *65*(4), 824-859. https://doi.org/10.1111/lang.12139

# Studying individual longitudinal development in a corpus of 'natural' disciplinary writing

Randi Reppen[1], Doug Biber[2]
Northern Arizona University
randi.reppen@nau.edu[1], douglas.biber@nau.edu[2]

Second language research often has the goal of describing learners' linguistic development as they gain proficiency with writing in a second language (L2). Although most previous studies have been cross-sectional (for practical reasons), several scholars have recently carried out longitudinal investigations of writing development.

Two major research issues for such longitudinal investigations are the 'naturalness' of the learner language being studied and the extent to which the research design is truly longitudinal. Regarding the first issue, data for longitudinal studies are usually collected in language classrooms or exams, with the writing tasks/topics being tightly controlled. However, the major disadvantage is that such tasks can be unrepresentative of the kinds of writing required in disciplinary content courses (see, e.g., Staples et al. 2018).

In addition, studies differ in the extent to which they are truly longitudinal. One approach is to collect data from a group of students at two points in time. When comparisons are generalized to the group, this approach can be characterized as quasi-longitudinal. To address this concern, several scholars have advocated a focus on individual longitudinal development (e.g., Bulté & Housen 2018 and Lowie & Verspoor 2015).

The present study analyzes writing development as it occurs 'naturally' in university disciplinary content courses, with the primary research goal of comparing the kinds of findings possible in a quasi-longitudinal design versus a true longitudinal design. We began with a quasi-longitudinal comparison of complexity features used by a group of 22 university students at two points in time (separated by two years). However, our regression models indicated that differences across academic disciplines and levels were more important than development across time for most linguistic features (see Biber et al. 2020).

This led us to conduct a true longitudinal study using four complexity features (i.e., adverbial clauses, relative clauses, attributive adjectives, and nouns as pre-modifiers). We focused on six students whose papers from Time 1 and Time 2 could be matched for discipline and task. While the total sample size is small (only eight sets of papers that could be matched for discipline and register), this approach allowed a focus on individual learner variation. When comparing the linguistic complexity features across the two time periods, we generally found the expected pattern of development: an increase in the use of phrasal features accompanied by a decrease in clausal features. However, there also were some puzzling results that prompted a more detailed analysis focusing on methodological issues.

For example, in the case of one student, we had two different texts from the same discipline/register (Informational essays from Social Science) for both Time 1 and Time 2. It turned out that the linguistic characteristics of these individual texts were surprisingly different, reflecting the effect of the specific topic. For example, one essay elicited a high use of attributive adjectives for describing personality traits (*social disinterest*, *parental attitudes*, *cultural meaning*), while the second essay focused more on reporting on theories and therefore used more nouns as premodifiers (e.g., *community membership*, *research area*, *partners' perspectives*).

In conclusion, we see from this close examination of linguistic writing development that controlled for discipline and task that in general, the students showed an increase in proficiency as measured by these linguistic indicators. However, even when discipline and task are matched, variation in topics can create a picture that goes against the expected trends. As we continue to research writing development with naturalistic data (e.g., writing from academic coursework) in an effort to get a realistic picture of the multifaceted aspects of writing development, we need to be prepared to explore the impact of variables like register, discipline, and topic interacting with individual variation.

**References**

Biber, D. Reppen, R. Staples, S. & Egbert, J. (2020). Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. *International Journal of Learner Corpus Research,* 6(1), 38 - 71.

Bulté, B. & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics. 28,*147-164.

Lowie, W., & Verspoor, M. (2015). Variability and variation in second language acquisition orders: A dynamic reevaluation. *Language Learning, 65*(1), 63–88.

Staples, S., Biber, D., & Reppen, R. (2018). Using Corpus-Based Register Analysis to Explore Authenticity of High-Stakes Language Exams: A Register Comparison of TOEFL iBT and Disciplinary Writing Tasks. *The Modern Language Journal 102*(2): 310-332.

## On the other side of the error tag: The nature and functions of the corrected texts

Lisa Rudebeck[1], Gunlög Sundberg[2]
Department of Swedish Language and Multilingualism, Stockholm University
lisa.rudebeck@su.se[1], gunlög.sundberg@su.se[2]

Any detection or categorization of an "error" in a learner text depends on an idea about an alternative, "correct", version of the text segment often referred to as a reconstruction, a target hypothesis, or a target form (see e.g. Lüdeling & Hirschmann 2015). Lüdeling & Hirschmann (2015: 141) emphasize the importance of providing explicit target hypotheses, and more and more learner corpora follow this practice. As pointed out by Tenfjord, Hagen & Johansen (2009: 63–64), this means that the learner corpus also becomes a parallel corpus, consisting of a corpus of original learner texts and a corpus of their "reconstructions".

In the recently released Swedish learner corpus SweLL (Volodina et al 2019) we have followed the implications of these insights even further. A fundamental aspect of the SweLL methodology is the systematic and clear separation between 1) the creation of a corrected text version and 2) all annotation. We have used the term *normalization* to refer both to the process of creating a "correct version" of a text and to the resulting text, and we have chosen the term *correction annotation* rather than *error annotation*. In this methodologically focused paper, we will describe the practical implementation of this far-reaching separation between normalization and annotation, including work practices and the design of a new annotation tool, *Svala*, which makes the separation manageable. We will further discuss the theoretical motivations for and implications of a methodology to which this separation is central.

The correction annotation is precisely a categorization of corrections, i.e. of differences between the original texts and their normalizations. By the choice of the term *correction*, rather than *error*, we emphasize the fact that these corrections are not inherent properties of the original texts but only arise through the introduction of an alternative. It is crucial that in relation to the correction annotation, the normalizations are data – just as much as the original texts on which they are based. In order to properly understand and exploit the information given by the manual correction annotation, as well as by the automatic linguistic annotation of the normalizations, it is, therefore, necessary to understand the nature of the normalizations: What kind of texts are they, and how do they relate to the original texts?

The normalizations have been carried out with the aim to create a comparable text version which adheres to the norms of standard Swedish while staying as close to the original text string as possible and communicating the perceived intended content as effectively as possible. The result of this balancing act can be seen as an *interpretation* or *translation* of the original text into "standard Swedish", or, more specifically, normalized learner Swedish. This variety of Swedish should not be assumed to be of the same kind as the Swedish found in similar texts originally written by native speakers. Rather, it has its own characteristics, influenced both by (highly variable) learner language traits and standard Swedish norms.

The translation-like normalization process is very different from the process of correction annotation, and we will argue that while category-based consistency and inter-annotator reliability should definitely be aimed at in correction annotation, there are clear benefits with a normalization process which is carried out according to broad values (norm adherence, fidelity to the original text) rather than through an attempt at a stricter, more rule-governed procedure. Any such attempt will necessarily introduce an artificial element to the normalizations which will decrease their value as data for research on normalized learner language and the relations between this and other kinds of Swedish.

Our contribution to the recurring discussion of the "comparative fallacy" (Bley-Vroman 1983) in part consists in stating that SweLL's explicit parallel corpus character should hopefully reduce the risk that comparisons be taken for something else; the comparative fallacy is only a fallacy if something other than a comparison is intended.

To sum up: The SweLL normalizations are samples of *normalized learner Swedish*. In relation to the original texts, they are *interpretations*, and in relation to the *correction annotation*, and for the corpus user who chooses to use this half of the parallel corpus, they are *data*.

**References**

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: the case of systematicity. *Language Learning*, 33, 1-17.

Lüdeling, A. & Hirschmann, H. (2015). Error annotation systems. In: Granger, S., Gilquin, G., & Meunier, F. (Eds.). *Cambridge handbook of learner corpus research*. Cambridge: CUP, 135-157.

Tenfjord, K., Hagen, J. E., & Johansen, H. (2009). Norsk andrespråkskorpus (ASK) – design og metodiske forutsetninger. *NOA norsk som andrespråk*, 25(1), 52-81.

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C-J., Sundberg, G. & Wirén, M. (2019). The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology*, 6, 67-104.

# Task effects on phraseological complexity in learners' written and oral production: A structural equation modeling study

Stefania Spina
University for Foreigners of Perugia
stefania.spina@unistrapg.it

Studies on the use of phraseology in second language acquisition show that there is a need for more research on the spoken production of phraseological units by L2 learners (Brezina & Fox 2021; Zhang et al. 2021). In the area of phraseological complexity, in particular, the two dimensions of diversity and sophistication have been mainly investigated in learners' written use (Paquot 2019; Paquot et al. 2021). Furthermore, while the relevance of the effects of task on learners' written and oral production has been repeatedly emphasised (Alexopoulou et al. 2017; Biber & Gray 2013), little attention has so far been devoted to the ways different tasks may affect the use of phraseology in L2 learners.

This study aims to fill these gaps by investigating the effect of different tasks on phraseological complexity in written and oral productions. For this purpose, the phraseological units used within the adjectival modifier grammatical dependency in Chinese learners of Italian have been considered. These units are particularly challenging in Italian, as they can be represented by both the noun + adjective and the adjective + noun lexical combinations (Spina forthcoming).

Both combination types have been extracted from the COLI corpus (Corpus of Chinese Learners of Italian), which includes written and spoken texts produced by 30 pre-intermediate, intermediate, and upper-intermediate Chinese learners of Italian. The learner data was elicited based on four different tasks: two tasks for the spoken section (an interactive conversation with the researcher, and a monologic description of a set of pictures), and two for the written section (the answers to a fixed number of questions, and the advice given to some characters from a picture). The lexical combinations have been extracted from the pos-tagged version of the COLI corpus as noun + adjective and adjective + noun sequences.

Using Structural equation modeling (SEM), this study adopts a confirmatory approach and aims to verify the following three hypotheses on phraseological complexity, in its two dimensions of diversity and sophistication:

Hypothesis 1: phraseological diversity and sophistication are affected by the mode of production (Brezina & Fox 2021; Uchihara et al. 2022; Vandeweerd 2019; Van Vu & Peters 2022). The two measures are expected to be higher in written production, which has fewer time constraints and therefore allows a more accurate selection of phraseological units.

Hypothesis 2: this mode effect is mediated by task (Alexopoulou et al. 2017; Biber & Gray 2013; Erman et al. 2018; Zhang et al. 2021). Differences in phraseological complexity are expected particularly between monologic and interactive tasks;

Hypothesis 3: phraseological diversity and sophistication increase with proficiency (Paquot 2019; Römer & Garner 2019; Rubin et al. 2021), but this effect is mediated by task as well.

The use of SEM is particularly promising in the field of learner corpus research, as it allows for testing a priori hypotheses on the relations between multiple variables (Hancock & Schoonen 2015; Larsson et al. 2020). SEM is a flexible technique for hypothesis-driven research, and offers at least the following advantages:

- it allows multiple dependent variables to be included in the models;
- it enables the distinction between variables that directly affect other variables and variables that have an indirect effect (mediators).

The model used to test the three mentioned hypotheses includes several independent variables, which in turn can be observed (mode, topic, combination type) or latent (L2 proficiency), a mediating variable (task), and two dependent variables (phraseological diversity and sophistication). In line with previous studies (Paquot 2019; Paquot et al. 2021; Rubin et al. 2021), phraseological diversity and sophistication are operationalised using the root type-token ratio and the mutual information (MI) scores, which are computed in an Italian reference corpus.

Preliminary results confirm the effects of mode and proficiency on phraseological complexity: learners use more diversified and more sophisticated phraseological units in written than in oral productions, and the two measures tend to increase with proficiency. More importantly, the presence of a mediating effect of the task is also supported by the statistical model: this is especially evident in the answer task, which shows a higher

phraseological complexity even when compared to the other more interactive and "pragmatic" written task (giving advice to someone). By and large, then, the two measures of phraseological complexity seem to be task sensitive.

**References**


Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67, 180–208.

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL iBT test: A lexico-grammatical analysis. *ETS Research Report Series*, 2013(1), i-128.

Brezina, V., & Fox, L. (2021). Adjective + Noun Collocations in L2 and L1 Speech: Evidence from the Trinity Lancaster Corpus and the Spoken BNC2014. In S. Granger (Ed.). *Perspectives on the L2 Phrasicon: The View from Learner Corpora*. Bristol, Blue Ridge Summit: Multilingual Matters, 152-177.

Erman, B., Lundell, F., & Lewis, M. (2018). Formulaic Language in Advanced Long-Residency L2 Speakers. In K. Hyltenstam, I. Bartning & L. Fant (Eds.). *High-Level Language Proficiency in Second Language and Multilingual Contexts*. Cambridge: CUP, 96-119.

Hancock, G. R., & Schoonen, R. (2015). Structural Equation Modeling: Possibilities for Language Learning Researchers: SEM Possibilities for Language Learning. *Language Learning*, *65*(S1), 160–184.

Larsson, T., Plonsky, L., & Hancock, G. R. (2020). On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory*, 17(3), 683-714.

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

Paquot, M., Naets, H. & Gries, S. Th. (2021). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: verb + object structures in LONGDALE. In B. Le Bruyn & M. Paquot (Eds.). *Learner corpus research meets second language acquisition*, Cambridge: CUP, 122-147.

Römer, U., & Garner, J. R. (2019). The development of verb constructions in spoken learner English: Tracing effects of usage and proficiency. *International Journal of Learner Corpus Research*, 5(2), 207-230.

Rubin, R., Housen, A., & Paquot, M. (2021). Phraseological Complexity as an Index of L2 Dutch Writing Proficiency: A Partial Replication Study. In S. Granger (Ed.). *Perspectives on the L2 Phrasicon: The View from Learner Corpora*. Bristol, Blue Ridge Summit: Multilingual Matters, 101-125.

Spina, S. (forthcoming). The effect of time and dimensions of collocational relationship on phraseological accuracy. In A. Leńko-Szymańska & S. Götz (Eds). *Complexity, Accuracy & Fluency in Learner Corpus Research*, Amsterdam: John Benjamins.

Uchihara, T., Eguchi, M., Clenton, J., Kyle K., & Saito, K. (2022). To What Extent is Collocation Knowledge Associated with Oral Proficiency? A Corpus-Based Approach to Word Association. *Language and Speech*, 65(2), 311-336.

Van Vu, D., & Peters, E. (2022). The Role of Formulaic Sequences in L2 Speaking. In T.M. Derwing, M.J. Munro & R.I. Thomson (Eds.). *The Routledge Handbook of Second Language Acquisition and Speaking*. Abingdon: Routledge, 285-298.

Vandeweerd, N. (2019). *Phraseological Complexity in Oral and Written L2 French.* Presentation at Eurosla 29 - Lund, Sweden.

Zhang, X., Zhao, B., & Li, W. (2021). N-gram use in EFL learners' retelling and monologic tasks. *International Review of Applied Linguistics in Language Teaching*.

# Modality in Chinese EFL learners' academic writing: From semantic meaning to disciplinary variation

Qiuyi Sun
University of Birmingham
qxs649@student.bham.ac.uk

Modal verbs are typically used in academic writing to communicate the writer's evaluation of a proposition, or their attitude towards the content of a sentence. Important though these devices are, language learners often struggle to use them. There has been substantial research undertaken on the frequency distribution of modal verbs used by Chinese EFL learners (e.g., Yang 2018). However, little attention has been given to the semantic distribution of these modal verbs and their verb collocates in Chinese learner writing. This study seeks to add to the current understanding of learner use of modality in these two aspects with the help of semantic vector space models (Turney & Pantel 2010) and, in addition, identify potential disciplinary variations. The research questions addressed in this paper are: 1) What modal meanings can be identified in Chinese learners' academic English? 2) What verb collocates of modal verbs can be identified? 3) Do the meanings and verb collocates of modal verbs show any similarity or difference between two disciplines: English Literature, and Business and Management?

The study uses data from the Chinese Advanced English Learner Corpus of Academic Written English (CAEL-CAWE) (Zou 2018). This corpus consists of 456 dissertations and 4,193,413 tokens written in English by Chinese undergraduates and postgraduates in the disciplines of English Literature and Business and Management. To address the questions above, modal verbs were manually annotated with two meanings: epistemic modality (evaluating the probability of the truth of propositions) and deontic modality (imposing obligations or giving permission to the reader/audience). Verb collocates of each sense of modal verbs were then extracted to construct the semantic vector space. How these verbs semantically relate to each other was demonstrated on a two-dimensional map using multidimensional scaling. In other words, verbs that shared similar meanings were placed in proximity to one another based on the scaling algorithm. They were further divided into clusters for comparison.

This paper will therefore discuss the findings on disciplinary variation in the use of different meanings and verb collocates of modal verbs found within the CAEL-CAWE. My initial data analysis highlights a significant association between the disciplines and the meanings of *must*. The epistemic and deontic use of *must* are equally balanced in the discipline of English Literature but *must* is predominantly only used in the deontic sense in the Business and Management texts. As to the verb collocates, my findings are consistent with Biber et al. (1999)'s observation, that is, that modal verbs in the epistemic sense usually appear with stative verbs, whereas the deontic modality mostly occurs with dynamic verbs. In terms of discipline variation, the verb collocates of epistemic *must* seems to be more widely distributed in the English Literature texts than in the Business and Management ones. But verb collocates of deontic *must* show an opposite pattern. The verb collocates that are distinctive for English Literature mostly denote concrete actions (*dance*, *put*, etc.), whereas those in the Business and Management texts show a more balanced distribution across abstract (*provide*, *identify*, etc.) and concrete actions. A more composite picture is now provided for both the semantic distribution and verb collocates of *must* used by Chinese EFL learners across disciplines. Further investigation of other modal verbs will be reported to expand our understanding of how Chinese learners use modality, providing implications for language teaching.

**References**

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finnegan, E. (1999). *Longman Grammar of Spoken and Written English. Longman*. London: Longman.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141-188.

Yang, X. (2018). A Corpus-Based Study of Modal Verbs in Chinese Learners' Academic Writing. *English Language Teaching*, 11(2), 122-130.

Zou, Y. (2018). *First person pronouns in academic discourse by novice writers in China* (Doctoral dissertation, University of Birmingham, Birmingham, UK). Retrieved from https://etheses.bham.ac.uk/id/eprint/8522/5/Zou18PhD.pdf

## Automatic classification of Arabic learners of English based on complexity metrics

Jessica Tayeh Chamoun[1], Nicolas Ballier[2]
Université de Paris Cité, CLILLAC-ARP ERP 3967
jessica.tayeh@etu.u-paris.fr[1], nicolas.ballier@u-paris.fr[2]

This study investigates the correlation of already existing metrics to assigned CEFR levels of English productions written by Arab learners of English from the UAE. We are working with 383 essays from the ZAEBUC corpus collected at the university of Zayed University (Palfreyman 2021a, 2021b). The task for students consisted of writing (150-200 word) essays on the effect of social media. The corpus was collected for Arabic and English writings, but we only analyzed the English subcomponent. A rating of each text into one of the six CEFR bands (A1/C2) was provided for each text by three different raters. There were three different rankings but the experiment was based on the majority band

| ## | A1 | A2 | B1 | B2 | C1 |
|---|---|---|---|---|---|
| ## | 7 | 93 | 193 | 80 | 10 |

Our objective is to automatically classify these essays into the CEFR levels using metrics and compare the human and automated CEFR grading to evaluate the reliability of these classifications. Similar previous studies on complexity metrics were carried out for Spanish learners of French and French learners of English. Lexical diversity metrics were used to quantify the written productions of Spanish learners of French resulting in a 69% accuracy (Lissón & Ballier 2018). As for the French learners, lexical and syntactic complexity metrics supported best classifications, and microsystem metrics improved CEFR prediction (Gaillat et al. 2021).

The pipeline used for our data follows a four-step process (Sousa et al. 2020). First, the texts that were typed by students are annotated via the Stanford parser and analyzed with a chain of tools for complexity scores. Complexity features are extracted using LCA (Lu 2012) and TAALES (Kyle 2018) for lexical complexity, L2SCA (Lu 2010) and TAASC (Kyle et al. 2018) for syntactic complexity, TAACO (Crossley et al. 2019) for cohesion, and the PyEnchant python library for misspelled words (Kelly 2016). The Python textstat7 library was also used to compute complementary readability metrics. Third, a machine learning algorithm is applied to predict the CEFR levels. Finally, visualization and results are generated using Rstudio.

There are 768 metrics in total that are computed. Some are related to clause complexity such as cl_av_deps that computes dependents per clause. Others computed noun phrase complexity such as vmod_all_nominal_deps_struct for verbal modifiers per nominal. Other types of metrics are basic_connectives that calculate the number of basic connectives such as 'for, and, nor', and microsystems such as MD_WILL, MD_MAY, MD_CAN for modals.

After feature selection, statistical analysis used random forests (rf), linear support vector machine (lsvm), and extreme gradient boosting (xgboost) classifiers, so as to test the utility of metrics by producing confusion matrices showing the automated classifications of our dataset. Feature importance was also computed to show the strength of specific metrics relating to a specific level. By comparing A2/B1 feature importance, our study showed that more metrics correlated to A2 than B1 such as CT (shortest meaningful sentence), CP (coordinate phrases, conjunctions), and specific microsystem metrics (e.g. the rate of uses of may among the use of modal auxiliaries provided by the metric MD_MAY).

By analyzing certain metrics, our study showed that classification was more successful in classifying B1 levels with an overall accuracy of 32%. 81 out of 193 were properly classified as B1 as opposed to the other levels where fewer essays were appropriately labeled. Nevertheless, using extreme gradient boosting (XGB, Chen et al. 2015) with a 70/30 split ratio for training and testing on our data, we managed 72.17% accuracy on our test set.

This talk will describe the metrics we found relevant for the classification of learners.

**References**

Ballier, N., Gaillat, T., Simpkin, A., Stearns, B., Bouyé, M., & Zarrouk, M. (2019). A supervised learning model for the automatic assessment of language levels based on learner errors. In *European Conference on Technology Enhanced Learning*, 308-320. Springer, Cham.

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, 722-731.

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (2021). Predicting CEFR

levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 1-17. doi:10.1017/S095834402100029X

Kelly, R. (2016). PyEnchant a spellchecking library for Python. Available on: https://pythonhosted. org/pyenchant

Khushik, G. A., & Huhta, A. (2020). Investigating Syntactic Complexity in EFL Learners' Writing across Common European Framework of Reference Levels A1, A2, and B1. *Applied Linguistics*, 41(4), 506-532.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods*, 50(3), 1030-1046.

Lan, G., Lucas, K., & Sun, Y. (2019). Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essays written by Chinese students in a first-year composition course. *System*, 85, 102116.

Lissón, P., & Ballier, N. (2018) 'Investigating Lexical Progression through Lexical Diversity Metrics in a Corpus of French L3', *Discours*, 23, 2. https://doi.org/10.4000/discours.9950

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International journal of corpus linguistics, 15(4), 474-496.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190-208.

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Springer.

Palfreyman, D. (2021a). A bilingual writer corpus for research on biliteracy, Zayed University (UAE) https://www.zu.ac.ae/main/en/research/lhebc/_blog/_blog-pages/a-bilingual-writer-corpus-for-research-on-biliteracy.aspx

Palfreyman, D. (2021b). Language on the Move. Designing and using a bilingual writer corpus. https://www.youtube.com/watch?v=TKETk-zvqpI

Sousa, A., Ballier, N., Gaillat, T., Stearns, B., Zarrouk, M., Simpkin, A., & Bouyé, M. (2020). From Linguistic Research Projects to Language Technology Platforms: A Case Study in Learner Data. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 112-120

# Acquisition of Norwegian as a second language: What are the differences between the written and spoken language of the learners?

Annely Tomson
University of Oslo
annely.tomson@iln.uio.no

Most of the studies focusing on the structures of Norwegian as a second language (L2) are based on the learners' written language (see Jensen 2018 for an overview). Consequently, research on features of the L2 learners' spoken language is rare in the Norwegian context (Jensen 2018: 253). In this presentation, I will discuss the preliminary results of a study based on a learner corpus of written and spoken Norwegian called NORINT[1] Corpus (Tomson et al. 2021). The NORINT Corpus consists of spoken and written data elicited from adult learners of Norwegian with an intermediate command of the target language i.e. the B1 level or higher. It is made up of three subcorpora: (1) NORINT Speech, (2) NORINT Text, and (3) NORINT Recited. In total, 40 first languages are represented in the NORINT Corpus and it uses Glossa[2], a user-friendly and functional search tool developed by the Text Laboratory[3]. The NORINT Corpus is accessible through Feide/CLARIN or available upon request for research purposes. Additional information about how the corpus is transcribed and annotated will be given in my presentation.

The study based on the NORINT Corpus examines the usage of the traditional Norwegian three-gender system (masculine, feminine, and neuter) among L2 learners. This system is nowadays undergoing a change and is in many dialects being replaced by a two-gender system (common and neuter), known e.g. from the conservative version of the dominating official written standard, *Bokmål* (Lødrup 2011; Rodina & Westergaard 2015; Busterud et. al 2019; Opsahl 2021). However, the three-gender system is often presented in textbooks intended for L2 learners, even though most of the textbooks are written in Bokmål. There is a second official written standard, *Nynorsk*, that actively uses three gender categories but there are very few L2 learners that choose/have the possibility to learn Nynorsk. All the informants in the NORINT Corpus have learnt Bokmål. This corpus-based study examines if the L2 learners use the indefinite feminine article *ei* 'a/an', the feminine suffixed definite article *-a* 'the', and the feminine possessive *mi* 'my' in their spoken and written language, or do they use the corresponding common gender forms, *en* 'a/an', *-en* 'the', and *min* 'my' instead? A preliminary analysis of findings from the two subcorpora, NORINT Speech and NORINT Text, suggests that the definite feminine article *ei* is used rarely, the feminine suffixed definite article *-a* is almost nonexistent in the corpora, and the feminine possessive *mi* seems to be used together with nouns denoting females. Instead, the common gender forms are used.

In conclusion, I will discuss possible explanations for the above-mentioned findings. I also analyze the possibilities and limitations of the NORINT Corpus, based on the study presented above.

**References**

Busterud, G., Lohndal, T., Rodina, Y. & Westergaard, M. (2019). The loss of feminine gender in Norwegian: A dialect comparison. *Journal of Comparative Germanic Linguistics*, 22, 141–167.

Jensen, B. U. (2018). Syntaks i norsk innlærerspråk: empiriske funn. In A.-K. H. Gujord & G. T. Randen (Eds.), *Norsk som andrespråk – perspektiver på læring og utvikling*. Oslo: Cappelen Damm, 235–260.

Lødrup, H. (2011). Hvor mange genus er det i Oslo-dialekten? *Maal og minne*, 103(2), 120–136.

Opsahl, T. (2021). Dead, but Won't Lie Down? Grammatical Gender among Norwegians. *Journal of Germanic Linguistics*, 33(1), 122–146.

Rodina, Y. & Westergaard, M. (2015). Grammatical gender in Norwegian: Language acquisition and language change. *Journal of Germanic Linguistics*, 27(2), 145–187.

Tomson, A., Szymanska, O. & Hagen, K. (2021). NORINT-korpuset – et elektronisk innlærerkorpus til bruk i andrespråksforskning. *NOA norsk som andrespråk*, 37(1–2), 235-257.

[1] NORINT is an abbreviation for Norwegian for International Students, and it refers to Norwegian courses offered for international students at the Department of Linguistics and Scandinavian Studies, University of Oslo.

[2] More information about Glossa: https://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/glossa/index.html

[3] More information about the Text Laboratory: https://www.hf.uio.no/iln/english/about/organization/text-laboratory/

# The effect of phraseological complexity on ratings of oral versus written French proficiency

Nathan Vandeweerd
UCLouvain, Vrije Universiteit Brussel
nathan.vandeweerd@uclouvain.be

Research has begun to show that written L2 texts that are rated as more proficient tend to exhibit higher levels of phraseological complexity, which can be seen in the use of a more diverse set of phraseological units (Rubin et al., 2021) and in the use of phraseological units that are more strongly associated in a large L1 reference corpus (Granger & Bestgen, 2014; Paquot, 2019; Rubin et al., 2021; Vandeweerd et al., 2021). While these findings speak to the link between phraseological complexity and proficiency in L2 *writing*, studies looking at L2 speech paint a somewhat different picture. Paquot et al. (in press) for example, found that the pointwise mutual information of verb + noun collocations in L2 speech actually decreased with proficiency.

Speech and writing differ in a number of ways that may be relevant for phraseology. In particular, the pressures of real-time production mean that there is often less opportunity for online planning in speech compared to writing (Skehan, 1998). The constraints of online production may therefore lead to the use of more highly frequent but less strongly associated phraseological units (as suggested by Biber & Gray, 2013). In addition, speaking and writing often have different communicative functions and therefore require different linguistic features. Very broadly, this can be understood as a tendency for noun-based, phrasal discourse in writing and verb-based, clausal discourse in speech (Biber, 2019). These differences also extend to the level of phraseology. Biber, Conrad, and Cortes (2004), for example, found that oral corpora had a higher proportion of noun-based lexical bundles whereas written corpora had a higher proportion of noun-based lexical bundles.

Only one study so far has directly compared phraseological complexity (in terms of both diversity and sophistication) between oral and written L2 production. In a longitudinal study of 29 university-level learners of French, Vandeweerd et al. (submitted) found that although written tasks generally had higher levels of phraseological complexity overall (in line with Skehan, 1998), the difference between the two modes was larger for adjective + noun collocations than for verb + noun collocations. The generalizability of the results was limited, however, by the small and relatively homogeneous sample of learners as well as the fact that it focused on only two types of phraseological units. The current study addresses these limitations by broadening the scope of phraseological units (as defined by Gries, 2008) and by using a larger and more diverse sample of learners. Specifically, we aim to determine the extent to which the diversity and sophistication of four-word lexical bundles are predictive of proficiency scores in oral versus written components of the *Test d'évaluation de français* (TEF, Chambre de commerce et d'industrie de Paris, 2010; Noël-Jothy & Sampsonis, 2006). In order to tap into the register differences between modes (cf. Biber, 2019), we focused on two types of bundles: those starting with a noun (noun-bundles) and those starting with a verb (verb-bundles). Phraseological diversity was operationalized as the proportion of unique bundles (e.g. *faire part de mon*; 'announce my') to the number of unique POS-bundle structures (e.g. VERB + NOM + PRP + DET). Phraseological sophistication was operationalized as the pointwise mutual information of the first word in the bundle and the rest of the sequence (e.g. faire | part de mon) based on a 10-billion-word reference corpus. A mixed effects linear regression model revealed that two of the four phraseological complexity measures were positively and significantly correlated with proficiency scores: the diversity of noun bundles (in the case of the written productions) and the sophistication of verb bundles (in both the written and oral productions). These results suggest that although the communicative function of a text may promote the use of certain phraseological units (Biber et al., 2004), access to those units seems to be mediated by modality (Skehan, 1998) as well as proficiency. We discuss the implications of these findings for language testing, especially the importance of considering both mode and register in scoring rubrics descriptors of phraseological competence.

**References**

Biber, D. (2019). Text-linguistic approaches to register variation. *Register Studies*, *1*(1), 42–75.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., & Gray, B. (2013). *Discourse Characteristics of Writing and Speaking Task Types on the Toefl iBT Test: A Lexico-Grammatical Analysis*. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Chambre de commerce et d'industrie de Paris. (2010). *TEF: Le Test d'évaluation de français de la Chambre et de l'industrie de Paris*.

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 in Language Teaching*, *52*(3), 229–252. https://doi.org/10.1515/iral-2014-0011

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora: further explorations. *International Journal of Corpus Linguistics*, *13*, 403–437. https://doi.org/10.1163/9789042028012_014

Noël-Jothy, F., & Sampsonis, B. (2006). *Certifications et outils d'évaluation en FLE*. Hachette.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Paquot, M., Gablasova, D., Brezina, V., & Naets, H. (in press). Phraseological complexity in EFL learners' spoken production across proficiency levels. In A. Lénko-Szymańska & S. Götz (Eds.), *Complexity, accuracy and fluency in learner corpus research*. John Benjamins.

Rubin, R., Housen, A., & Paquot, M. (2021). Phraseological complexity as an index of L2 Dutch writing proficiency: A partial replication study. In S. Granger (Ed.), *Perspectives on the second language phrasicon: The view from learner corpora* (pp. 101–125). Multilingual Matters.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford 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. (submitted). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Manuscript Submitted for Publication*.

## How do tasks impact the different domains of L2 linguistic complexity?

Zarah Weiss[1], Detmar Meurers[2]
University of Tübingen
zarah-leonie.weiss@sfs.uni-tuebingen.de[1], dm@sfs.uni-tuebingen.de[2]

The effect of the task on the linguistic complexity of the language that it elicits is of direct relevance for the valid interpretation of L2 complexity analyses. The linguistic complexity of learner productions has been extensively studied as a dimension of language performance in the Complexity, Accuracy, Fluency (CAF) framework (Housen et al., 2012). Considerable work has been dedicated to tracking L2 development and characterizing L2 proficiency based on complexity measures (cf. Housen et al. 2019 for an overview). But despite growing evidence for the influence of task factors on the expression of linguistic complexity (Kuiken & Vedder, 2008; Yoon & Polio, 2016; Alexopoulou et al., 2017; Michel et al., 2019), so far little attention has been paid to the how tasks impact the different linguistic domains at which complexity can be expressed. Identifying which features and linguistic domains remain robust across different elicitation contexts is a crucial step towards obtaining generalizable insights from complexity research, especially because it is challenging to design tasks suited for learners at all CEFR levels. In our work, we, therefore, explore the task sensitivity of complexity features across a broad range of linguistic domains by investigating their informativeness for task prompt classification.

As an empirical basis, we use the German section of the trilingual Merlin corpus (Abel et al, 2014). It consists of 1,033 German L2 essays written by beginning to advanced German L2 learners. The data were collected as part of official standardized language certification tests at the CEFR levels A1 to C1. At each test level, approximately 200 learner texts were elicited and rated by two trained annotators on the CEFR scale from A1 to C2, independent of the test level at which they were elicited. The German section of the Merlin corpus is one of the largest German L2 corpora, offering a uniquely broad range of CEFR proficiency ratings. The learner texts were elicited using 15 different task prompts, three per CEFR test level.

We analyze the linguistic complexity of these learner texts with the complexity analysis system for German originally proposed by Weiss & Meurers (2018), which has since been applied to automatically rate German L2 writing (Weiss & Meurers, 2019, 2021) and to automatically rate reading texts for German L2 learners (Weiss, Chen & Meurers, 2021) on the CEFR scale. We extract 236 measures of clausal, phrasal, lexical, and morphological complexity as well as measures of human language processing and language use. Weiss & Meurers (2019) used these features to characterize linguistic differences between learners at different proficiency levels in Merlin in a machine learning approach. We pursue a parallel set-up for our study, which will facilitate a comparison of the impact of tasks on the linguistic complexity of L2 writing in relation to the impact of proficiency. We train a Support Vector Machine classifier with a polynomial kernel using a 70/30 train/test split. When predicting one of the 15 task prompt labels based on the 236 linguistic complexity measures, we obtain an accuracy of 89.8%, substantially outperforming the majority baseline of 8.8%. The accuracy is also substantially higher than the state-of-the-art results for proficiency classification of around 70% on this data (Weiss & Meurers, 2019). This is remarkable since distinguishing 15 classes (the different task prompts) mathematically should be harder than 5 classes (the different proficiency levels), and L2 complexity research has focused on complexity differences related to proficiency differences, ignoring the task differences that apparently are even more richly encoded in the linguistic complexity characteristics. Probing into which complexity features are indicative of what, we analyze the features' information gain. We find that the 20 most informative features for task identification contain predominantly features of lexical diversity, surface length, word frequency, and morphological features of nominalization. Linguistically informed phrase-level complexity features, such as measures of clausal, phrasal, and cognitive complexity, are less informative for the task classification.

Our results show that especially word-level features of complexity are prone to reflect task differences rather than proficiency characteristics, whereas syntactic and cognitive features seem to be more robust indicators of proficiency. Based on these findings, we argue that the task dependency of complexity domains needs to be systematically considered in the setup and interpretation of complexity analyses to obtain generalizable results.

**References**


Abel, A, Wisniewski, K., Nicolas, L.; Boyd, A.; Hana, J.; Meurers, D. (2014). A trilingual learner corpus illustrating European reference levels. *Ricognizioni – Rivista di Lingue, Letterature e Culture Moderne*, 2 (1), 111-126. https://doi.org/10.13135/2384-8987/702

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1), 180-208. https://doi.org/10.1111/lang.12232

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second language research*, 35(1), 3-21.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA. John Benjamins Publishing.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of second language writing*, 17(1), 48-60.

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: evidence from a large learner corpus of A1 to C2 writings. *Instructed Second Language Acquisition*, 3(2), 124-152. https://doi.org/10.1558/isla.38248

Weiss, Z., Chen, X., & Meurers, D. (2021). Using broad linguistic complexity modeling for cross-lingual readability assessment. In: *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning*, 38-54. https://aclanthology.org/ 2021.nlp4call-1.4.pdf

Weiss, Z., & Meurers, D. (2018). Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. *Proceedings of the 27th International Conference on Computational Linguistics*, 303-317. https://aclanthology.org/C18-1026.pdf

Weiss, Z., & Meurers, D. (2019). Broad linguistic modeling is beneficial for German L2 proficiency assessment. *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, Louvain-La-Neuve. Presses Universitaires de Louvain, 419-435. http://purl.org/dm/papers/Weiss.Meurers-19LCR.pdf

Weiss, Z., & Meurers, D. (2021). Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. *International Journal of Learner Corpus Research*, 7(1), 83-130. https://doi.org/10.1075/ijlcr.20006.wei

# Using linguistic complexity to probe into genre differences? Insights from the multilingual SWIKO learner corpus

Zarah Weiss[1], Nina Selina Hicks[2], Detmar Meurers[3], Thomas Studer[4]
Universität Tübingen[1,3], University of Fribourg[4]
zarah-leonie.weiss@sfs.uni-tuebingen.de[1], nina.hicks@unifr.ch[2], dm@sfs.uni-tuebingen.de[3], thomas.studer@unifr.ch[4]

We investigate the expression of genre differences in writings of adolescent Foreign Language (FL) learners of English, French, and German. Language acquisition involves learning how to adapt language to its intended communicative function in context. Despite considerable research on the register and task effects on language performance (Biber & Gray 2010; Biber 2012; Alexopoulou et al. 2017; Yoon & Polio 2017) for English, little is known about their cross-linguistic generalizability and differences between FL learners and native (L1) speakers. We investigate linguistic complexity as a central dimension of language performance in the Complexity, Accuracy, Fluency framework in SLA research (Housen et al. 2012) to address this gap by asking:

(1) How to model cross-lingual genre differences in FL writing using linguistic complexity?
(2) To which degree does our cross-lingual FL model extend to L1 writing?
(3) Which linguistic measures play a role in the models' characterization of genre differences?

We analyzed 1,803 form-based, local normalizations of English, French, and German texts from the task-based SWIKO corpus (Karges et al. 2019). The corpus was elicited in Swiss lower secondary schools and consists of 1,002 FL texts (EN = 497; FR = 329; GE = 176) and 585 writings in students' language of schooling (L1 approximation, henceforth 'L1') elicited in French-speaking schools in the French-speaking part of Switzerland (FR = 152) and German-speaking and bilingual (English-German) schools in the German-speaking part of Switzerland (GE = 332; EN = 101). The task prompts used for data elicitation in the SWIKO corpus systematically vary what we refer to as genre (argumentative or descriptive), following, e.g., Yoon & Polio (2017), formality, and structuredness, making the corpus ideal for analyses of task factors.

To comparably measure linguistic complexity for English, French, and German on this corpus, we used the multi-lingual analysis platform CTAP (www.ctapweb.com; Chen & Meurers 2016). While originally developed for English, CTAP has been extended to support multiple languages (including French and German) and cross-lingual analyses (Weiss, Chen & Meurers 2021). We extracted 262 established measures of clausal, phrasal, lexical, and morphological complexity, as well as language use and human processing for French, German, and English (e.g., Chen 2018; Weiss & Meurers 2019, 2021).

We measure how FL learners adapt their language across a broad range of linguistic domains by predicting text genre with a machine learning classifier using these complexity measures. We first identified all measures that were sufficiently variable on the FL data (N = 200). We define a feature as sufficiently variable if its most common value does not occur in more than 80% of the data. We then trained a random forest algorithm using 10-fold cross-validation on the FL texts. We obtained an average accuracy of 80.4% (95% confidence interval = [79.6%; 81.3%]) against a majority baseline of 51.1%. This shows that we can successfully identify how FL learners implement genre differences in their writing. We then applied the model to the 'L1' texts obtaining an accuracy of 84.3% (baseline 50.1%). Thus, our classifier learns to distinguish genre based on linguistic markers of genre differences that are shared between FL and 'L1' writing and possibly expressed more pronounced in 'L1' writing, leading to the considerably higher accuracy on the 'L1' data.

To obtain a better understanding of these findings, we investigated the importance of the individual features in our model. We ranked the 200 features based on the average drop in accuracy when removing them from the model, keeping all other parameters constant. We zoomed in on the top and bottom 35 features in our ranking and compared the feature values across text genres and languages. We observe several differences between argumentative and descriptive FL writing with respect to clausal and nominal complexity as well as language use. Measures of human processing costs and cohesion hardly factor into the distinction of genres. These distinctions are remarkably homogenous across all three languages.

Overall, our findings demonstrate that broad cross-linguistic complexity modeling makes it possible to capture interlanguage genre differences in FL and 'L1' writings. We demonstrate that cross-linguistic studies are crucial to foster generalizable insights into language learning and register differences. Naturally, more research is needed to shed further light on the interplay between task factors, language proficiency, and linguistic complexity.

**References**


Alexopoulou, T., Michel, M., Murakami, A. & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1), 180-208.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.

Biber, D. & Gray, B. (2010). Challenging stereotypes about academic writing: Com- plexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9(1), 2–20.

Chen, X. (2018). *Automatic Analysis of Linguistic Complexity and Its Application in Language Learning Research* (PhD thesis). Eberhard Karls Universität Tübingen.

Chen, X. & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity at COLING*. Osaka: The International Committee on Computational Linguistics, 113-119.

Housen, A., Kuiken, F. & Vedder, I. (Eds.) (2012). *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Language Learning & Language Teaching Vol. 32. Amsterdam/Philadelphia: John Benjamins Publishing.

Karges, K., Studer, T. & Wiedenkeller, E. (2019). On the way to a new multilingual learner corpus of foreign language learning in school: Observations about task variations. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (Eds.). *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*., Corpora and Language in Use – Proceedings 5, Louvain-La-Neuve: Presses Universitaires de Louvain, 137-165.

Weiss, Z., Chen, X. & Meurers, D. (2021). Using broad linguistic complexity modeling for cross-lingual readability assessment. In D. Alfter, E. Volodina, I. Pilan, J. Graën, & L. Borin (Eds.). *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*. Linköping: LiU Electronic Press, 38-54.

Weiss, Z. & Meurers, D. (2019). Broad linguistic modeling is beneficial for German L2 proficiency assessment. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (Eds.). *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Corpora and Language in Use – Proceedings 5, Louvain-La-Neuve: Presses Universitaires de Louvain, 419-435.

Weiss, Z. & Meurers, D. (2021). Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. *International Journal of Learner Corpus Research*, 7(1), 83-130.

Yoon, H. J. & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *Tesol Quarterly*, 51(2), 275-301.

# Investigating effects of L1 and discipline on syntactic complexity in master's theses and research articles

Niwat Wuttisrisiriporn
Victoria University of Wellington
niwat.wuttisrisiriporn@vuw.ac.nz

Over the past two decades, syntactic complexity has been researched and recognized as an important construct in the second language (L2) writing research. Previous L2 writing studies have extensively examined the relationship of syntactic complexity to language development (e.g., Lu 2011) and language proficiency (e.g., Khushik & Huhta 2020; Lan et al. 2019; Zhang et al. 2022) of L2 writers at school and university levels. Recently, L2 writing studies are starting to pay attention to syntactic complexity in research articles written by L1 English writers and (compared to) L2 writers (e.g., Wu et al. 2020; Yin et al. 2021) of different disciplines (e.g., Lu et al. 2021). Results of those studies showed differences in syntactic complexity affected by different L1s and disciplines. However, syntactic complexity studies focusing on postgraduate research genres (e.g., theses and dissertations) are very few, and most studies comparing syntactic complexity of L1 English texts to that of L2 English texts treated different groups of L2 writers as a homogeneous group, not treating L2 writers' L1 background as an independent variable. This study, therefore, seeks to address the extent to which different L1s and disciplines affect syntactic complexity in master's theses and research articles.

To address the research question, a corpus of 4 million words was built, consisting of eight 500,000-word subcorpora of master's theses and research articles written by L1 and Thai writers of English in the fields of applied linguistics and engineering. Lu's (2010) 14 syntactic complexity measures were used to extract syntactic structures in the texts of the corpus. To investigate the differences in the 14 syntactic complexity measures, factorial ANOVAs were computed to observe interaction effects between genre and L1 as well as those between genre and discipline. In addition, the main effects of L1 and discipline on syntactic complexity in master's theses and research articles were also observed. Results of the present study show interaction effects between genre and L1 as well as those between genre and discipline in most of the 14 syntactic complexity measures, covering the length of production units, subordination, coordination, phrasal complexity, and overall sentence complexity. Significant main effects of L1 and pairwise comparisons indicate that L1 English writers made significantly greater use of most of the syntactic complexity structures for both theses and research articles, i.e. longer clauses and sentences, more dependent clauses, more coordinate phrases, and more complex nominals. Significant main effects of discipline and pairwise comparisons suggest that applied linguistics texts, compared to those of engineering, used a significantly greater number of most of the syntactic structures for both theses and research articles. The results of the present study contribute to a better understanding of the syntactic complexity of L2 research writing and provide useful pedagogical implications for teaching English for master's thesis and research publication purposes.

**References**

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across common European framework of reference levels A1, A2, and B1. *Applied Linguistics*, *41*(4), 506–532.

Lan, G., Lucas, K., & Sun, Y. (2019). Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essays written by Chinese students in a first-year composition course. *System*, *85*, 102116.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62.

Lu, X., Casal, J. E., Liu, Y., Kisselev, O., & Yoon, J. (2021). The relationship between syntactic complexity and rhetorical move-steps in research article introductions: Variation among four social science and engineering disciplines. *Journal of English for Academic Purposes*, *52*, 101006.

Wu, X., Mauranen, A., & Lei, L. (2020). Syntactic complexity in English as a lingua franca academic writing. *Journal of English for Academic Purposes*, *43*, 100798. https://doi.org/10.1016/j.jeap.2019.100798

Yin, S., Gao, Y., & Lu, X. (2021). Syntactic complexity of research article part-genres: Differences between emerging and expert international publication writers. *System*, *97*, 102427. https://doi.org/10.1016/j.system.2020.102427

Zhang, X., Lu, X., & Li, W. (2022). Beyond Differences: Assessing Effects of Shared Linguistic Features on L2 Writing Quality of Two Genres. *Applied Linguistics*, *43*(1), 168–195.

# A corpus-based contrastive analysis of transition markers in L1 Arabic and L2 English argumentative writing

Abdelhamid Ahmed[1], Lameya Rezk[2], Xiao Zhang[3]
Qatar University[1,3], Hamad Bin Khalifa University[2]
aha202@qu.edu.qa[1], lrezk@hbku.edu.qa[2], zhang.xiao@qu.edu.qa[3]

**Background Literature Review**

Transition markers, as a metadiscourse marker, have been named differently such as internal conjunctions (Halliday and Hasan, 1976; Hyland & Tse, 2004), linking adverbials (Biber et al., 1999), linking adjuncts (Richards & Schmidt, 2010), cohesive ties (Al-Jarf, 2001), discourse connectives (Blakemore, 2002), linkers (Thornbury, 2006), and logical markers (Mur Dueñas, 2009). In the current study, we adopt Hyland's semantic sub-types of transition markers (2005) (i.e. addition, comparison & contrast, and consequence markers).

At the textual level, transition markers help create cohesion by showing the logical links between propositions (Cao & Hu, 2014), leading readers to interpret meaning purposefully (Blakemore, 2002), and functioning ideationally by signalling how the writer logically relates different ideas (Hyland & Tse, 2004). Also, students' appropriate use of transition markers semantically can help alleviate the reader's burden of connecting preceding and subsequent content information (Cao & Hu, 2014). Thus, transition markers are important to the quality of L1 writing in Arabic (Nasib, 2018) and L2 English writing (Hinkel, 2001).

Some previous studies researched transition markers in L1 Arabic (Rabbah, 2016; Abdullah, 2017; Nasib, 2018); other studies investigated transition markers in L2 English of Arab students (Hinkel, 2001; Mohamed-Sayidina, 2010; Al-Rubaye, 2015; Appel & Szeib, 2018; Appel, 2020). Yet, the use of these markers by Arab L2 student writers is under-researched and is still problematic (Khalil, 1989; Al-Jarf, 2001; Ahmed, 2010; Hamed, 2014; Alshahrani, 2015; Basheer, 2016; Appel & Szeib, 2018). Yet, investigating transition markers in L1 Arabic and L2 English needs further investigation (Alshahrani, 2015). Previous research highlighted the need to conduct further research to help us better understand how alternative L1 groups (other than English) make use of transition markers in their writing (Appel, 2020). Therefore, our current study bridges the gap in the literature by contrastively analysing university students' use of transition markers in their L1 Arabic and L2 English argumentative writing. Therefore, the present study contributes to knowledge by investigating the quantity and variety of transition markers in L1 Arabic and L2 English argumentative writing, written by the same students, across different language proficiency levels (high, average, and low) and gender (154 females and 41 males). It also explores students' metalinguistic understanding of these transition markers in writing.

**Research Questions**

1. What is the difference (if any) in the overall quantity of transition markers used by L1 Arabic and L2 English university students of different proficiency levels?
2. What is the difference (if any) in the variety of transition markers used by L1 Arabic and L2 English university students of different proficiency levels?
3. What is the gender difference (if any) in students' use of transition markers used by L1 Arabic and L2 English university students?
4. What is the nature of students' metalinguistic understanding of transition markers in L1 Arabic and L2 English university students?

**Methods**

The current study adopts a mixed-methods design: a corpus-based methodology enriched with students' metalinguistic understanding. Two corpora[1] were built to investigate the English and Arabic writing of Qatari university students (195 essays each). Students' essays were rated for proficiency in Arabic (9 High, 184 Average, 2 Low; and in English (23 High, 147 Average, and 25 Low. In addition, students' metalinguistic understanding of transition markers was explored through writing conversation interviews with 51 participants. All participants

[1] The corpora have been submitted to Linguistic Data Consortium (https://www.ldc.upenn.edu/). The URLs of the official publication will be added later.

were native speakers of Arabic and L2 English speakers. Each participant wrote one essay in Arabic and one in English on two different topics. All ethical issues were addressed (BERA, 2018).

**Findings**

The following findings were revealed. First, no significant statistical differences were found in the quantity of transition markers in students' L1 Arabic and L2 English argumentative writing across the three proficiency levels. Second, no significant statistical differences were found in the variety of transition markers (i.e., addition, comparison & contrast, and consequences markers) across the three proficiency levels. However, (1) two addition markers in Arabic (furthermore علاوة علي ذلك & as well as كذلك) and one addition marker in English (not to mention) were statistically significant; (2) the following comparison & contrast markers in Arabic were statistically significant (the other side الجانب الآخر , the other team الفريق الآخر, the first team الفريق الأول, on the opposite وعلى العكس, at the same time في نفس الوقت, as an alternative كبديل) and the following markers in English were statistically significant (even though, the same goes for, the first team , the other team; (3) the following consequences markers in Arabic were statistically significant (because لأن, thanks to يعود الفضل, so as to وذلك, consequences of والآثار المترتبة ) and the following consequences markers in English were statistically significant (Lead, So as to).

In reference to gender differences, the following findings were reached. First, Arabic addition markers are significant at the level of 0.05 and in favour of females compared with males. No significant differences between males and females were found on the comparison & contrast markers and consequences markers. Second, no significant differences between males and females on all types of transition markers in English were found. Third, there are no significant statistical differences for the variables of Addition and Consequences in Arabic and English as the value of the T-test is insignificant. On the other hand, there are differences in Comparison and Contrast markers as (8.508) is the value of T which is significant at the level of 0.001 in favour of English. The differences in the quantity, variety, and gender among the participants might be attributed to some educational, socio-cultural, and L1 transfer into L2 factors.

Findings revealed students' metalinguistic understanding of their reasons, importance, specific purposes, and challenges with using transition markers in L1 Arabic and L2 English argumentative writing. Pedagogical and methodological implications are provided.

**References**


Abdullah, M. (2017). The effectiveness of a program based on linguistic analysis on developing the transition markers skills in the expressive writing of non-native speakers of Arabic. *Arab Studies in Education and Psychology*. 86(2), 283-345

Ahmed, A. (2012). Students' problems with cohesion and coherence in EFL writing in Egypt, *Literacy Information and Computer Education Journal*, 1(4) (2012), 211-221.

Al-Jarf, R. (2001). Processing of cohesive ties by EFL Arab College students, *Foreign Language Annals*, 34(1), 2-23.

Al-Rubaye, M. H. K. (2015). *Metadiscourse in the academic writing of EFL and ESL Arabic-speaking Iraqi graduate students*. Missouri State University.

Alshahrani, A. (2015). A cross-linguistic analysis of interactive metadiscourse devices employment in native English and Arab ESL academic writings. *Theory and Practice in Language Studies*, *5*(8), 1535.

Alsharif, M. (2017). The Frequently Used Discourse Markers by Saudi EFL Learners. *Arab World English Journal (AWEJ).* 8(2), 384-397.

Appel, R. (2020). An exploratory analysis of linking adverbials in post-secondary texts from L1 Arabic, Chinese, and English writers. *Ampersand*, *7*, 100070.

Appel, R., & Szeib, A. (2018). Linking adverbials in L2 English academic writing: L1-related differences. *System*, *78*, 115-129.

Basheer, N. (2016). *Arabic connectives in native speaker and non-native speaker expository and argumentative writing* (Doctoral dissertation).

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). London: Longman.

Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers* (Vol. 99). Cambridge university press.

British Educational Research Association [BERA] (2018). Ethical Guidelines for Educational Research, fourth edition, London. https://www.bera.ac.uk/researchers-resources/publications/ethicalguidelines-for-educational-research-2018

Cao, F., & Hu, G. (2014). Interactive metadiscourse in research articles: A comparative study of paradigmatic and disciplinary influences. *Journal of Pragmatics*, 66, 15-31.

Chen, H., & Myhill, D. (2016). Children talking about writing: Investigating metalinguistic understanding. *Linguistics and Education*, *35*, 100-108.

Dueñas, P. M. (2009). Logical markers in L1 (Spanish and English) and L2 (English) business research articles. *English Text Construction*, 2(2), 246-264.

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hamed, M. (2014). Conjunctions in Argumentative Writing of Libyan Tertiary Students. *English Language Teaching*, *7*(3), 108-120.

Hammond, J. (2012). Hope and challenge in the Australian Curriculum: Implications for EAL students and their teachers. Australian Journal of Language and Literacy, 35(1), 223–240.

Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied language learning*, *12*(2), 111-132.

Hyland, K. & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied linguistics*, *25*(2), 156-177.

Hyland, K. (2005) *Metadiscourse*. London: Continuum.

Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics,* 113, 16 – 29.

Khalil, A. (1989). A study of cohesion and coherence in Arab EFL college students' writing, *System*, 17(3), 359-371.

Khalil, A. (1989). A study of cohesion and coherence in Arab EFL college students' writing, *System,* 17(3), 359-371.

Mahmoud, A. (2014). The Use of Logical Connectors by Arab EFL University Students: A Performance Analysis. *International Review of Social Sciences and Humanities*, 7(1).

Mohamed-Sayidina, A. (2010). Transfer of L1 cohesive devices and transition words into L2 academic texts: The case of Arab students. *RELC Journal*, *41*(3), 253-266.

Mohammed-Sayidina, A. (2010). Transfer of L1 cohesive devices and transition words into L2 academic texts: The case of Arab students, *RELC Journal*, 41(3), 253-266.

Moore, J., & Schleppegrell, M. (2014). Using a functional linguistics metalanguage to support academic language development in the English Language Arts. Linguistics and Education, 26, 92–105. http://dx.doi.org/10.1016/j.linged.2014.01. 002

Myhill, D., Jones, S., & Wilson, A. (2016). Writing conversations: fostering metalinguistic discussion about writing. *Research Papers in Education*, 31(1), 23-44.

Nasib, A. (2018). Transition Markers in Arabic and its Impact on Semantics. Unpublished PhD Thesis, University of Algeria II, College of Arabic Language, Arts and Eastern Languages, Algeria.

Rabbah, K. (2016). Transition Markers in the Arabic Language Structure, An Applied Study on Seven Long Chapters in the Holy Quran: An Analytical Grammatical Study, Unpublished PhD Thesis, Al-Aqsa University, Palestine.

Richards, J. & Schmidt, R. (2010*). Longman dictionary of language teaching and applied linguistics* (4th ed.). Harlow: Pearson.

Thornbury, S. (2006). *An A – Z of ELT*. London: Macmillan.

# Lexical complexity in L2 English speech: Exploring monologic and dialogic tasks in the Trinity Lancaster corpus

Raffaella Bottini
Lancaster University
r.bottini@lancaster.ac.uk

Different measures of vocabulary knowledge have been proposed in the field of second language research and lexical complexity plays a key role among them (Kyle 2019; Lu 2012). However, little is known about different aspects of lexical complexity in L2 spoken production and how these are influenced by task-related features. This study investigated whether task interactivity has an effect on lexical complexity in L2 speech, using data from the Trinity Lancaster Corpus (TLC; Gablasova et al. 2019). The TLC is a 4.2-million-word learner corpus based on the Graded Examination in Spoken English (GESE), a high-stakes exam of L2 English developed and administered by Trinity College London which is a large international examination board. The corpus consists of transcripts of learners' spoken performance across four tasks which differ in terms of interactivity and topic familiarity. This study compared L2 lexical production across monologic and dialogic tasks, combining quantitative and qualitative analysis. Lex Complexity Tool (Bottini 2022), which includes a wordlist from the Spoken BNC2014 (Love et al. 2017), was used to measure lexical scores. The results from repeated-measures ANOVA show that task interactivity has a statistically significant effect on lexical diversity ($\eta^2$ = .18), lexical density ($\eta^2$ = .31), and lexical sophistication values based on mean frequency scores ($\eta^2 \leq .43$). Case studies that showed opposite tendencies in the data were used to explore individual variation. The findings suggest that interactive speech is characterised by less diverse and less sophisticated vocabulary than monologic production. Among the possible explanations for these findings, factors related to task design in the GESE, learners' educational background, real-time processing, and social features of language use are discussed. This study has implications for learner corpus research, language testing, and language teaching.

**References**

Bottini, R. (2022). *Lexical complexity in L2 English speech: Evidence from the Trinity Lancaster Corpus* [Unpublished PhD thesis, Lancaster University].

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, 5(2), 126-158.

Kyle, K. (2019). Measuring lexical richness. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies*. Milton: Routledge, 454-476

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319-344.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190-208.

## Measuring syntactic complexity in L2 speech at advanced proficiency levels

Barbora Bulantová
Charles University
bbulantova@gmail.com

The past decade has seen an increase in the number of studies focusing on the complexity of learner language and thus contributing to the growing body of Complexity-Accuracy-Fluency research. The complexity of learner language is typically operationalized on the syntactic level of language performance (Bulté & Housen 2012). While much has been published on written syntactic complexity, spoken language complexity appears to have fallen somewhat by the wayside. Especially when it comes to describing how L2 syntactic complexity develops up to advanced levels of L2 proficiency, only a handful of studies are available (e.g. De Clercq & Housen 2017, Bulté & Roothooft 2020). This is perhaps due to a multitude of methodological issues connected with the analysis of spontaneous speech, which is inherently elliptical and abundant in various hesitation phenomena, which makes finding a principled way of dividing transcribed data into units very difficult.

One of the most successful attempts at solving the difficulty of segmenting spoken language is the AS-unit (Foster et al. 2000). Its authors encouraged future researchers to develop the concept by addressing some of the issues not covered in the original paper such as inter-coder agreement and establishing a clear data exclusion policy. Nevertheless, most studies to date claim to have used the guidelines on the segmenting procedure but avoid providing a detailed description of how they dealt with the many problematic issues not covered in the original manual. Such a vague delimitation of the analytical unit and avoidance of mentioning the extent of data pruning may skew the results, reducing the validity of the studies and their potential replicability.

Another issue which has received a lot of attention in CAF literature is that despite the large array of available measures, most studies analysing L2 syntactic complexity have used only crude, length- and subordination-based metrics (Norris & Ortega 2009, Bulté & Housen 2012). Only seldom have such measures been supplemented with fine-grained indices, which could reveal the source of complexification.

This study analyses the syntactic complexity of monologic tasks of ten B2 (6,247 tokens, 612 AS-units) and ten C1 (6,838 tokens, 623 AS-units) speakers of English with Czech as their L1 with the aim to determine whether the complexity of learners' spontaneous speech is higher at C1 than at B2 level and if it is, which quantitative measures of syntactic complexity show the effect of proficiency level. The data derives from LINDSEI_CZ (Gráf 2017).

The transcripts of the recordings were segmented into AS units (Foster et al. 2000). A detailed manual, much extending the original, was created to ensure a systematic approach towards data exclusion and dealing with borderline cases. Excerpts from the monologues were segmented by eight raters and compared against the original segmented text. Cohen's kappa was used to calculate inter-rater reliability ($\kappa$=0.89).

Syntactic complexity was measured using crude measures (mean length of clause, mean length of AS-unit, clauses per AS-unit) combined with fine-grained indices of structural complexity, including proportions of relative, infinitive, adverbial, complement and independent clauses, and Vercellotti's (2019) weighted complexity scale.

Mann-Whitney U test was deployed to compare the proficiency groups in terms of all the syntactic measures. It revealed that there was no significant effect of proficiency on syntactic complexity. In fact, scores of each measure tended to vary within the groups, suggesting a possible effect of inter-speaker variability among more advanced speakers. The similarity of datasets in all dimensions of syntactic complexity analysed in the research could be also linked to the sample size and the proximity of the participants in terms of their proficiency level.

This study contributes to spoken L2 complexity research by comparing two proficiency groups of speakers with Czech as their L1. Its main contribution is, however, methodological, as it aims to identify issues in analysing spontaneous speech and offer a principled way of text segmentation and data exclusion in monologic tasks.

**References**


Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.). *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA.* Amsterdam: John Benjamins, 21-46.

Bulté, B., & Roothooft, H. (2020). Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System*, 91.

De Clercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *Modern Language Journal*, 101(2), 315-334.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics*, 21(3), 354-375.

Gráf, Tomáš. (2017). LINDSEI_CZ: A Corpus of Spontaneous Spoken English of Advanced Speakers. Institute of the Czech National Corpus FF UK, https://wiki.korpus.cz/doku.php:en:cnk:linsei_cz.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555-578.

Vercellotti, M. L. (2019). Finding variation: assessing the development of syntactic complexity in ESL Speech. *International Journal of Applied Linguistics*, 29(2), 233-247.

## English in a bilingual German-Italian community: Collecting data and investigating learner variables in creating the EdiCoMC corpus

Graham Burton[1], Maria Cristina Gatti[2]
Free University of Bozen-Bolzano
grahamfrancis.burton@unibz.it[1], mariacristina.gatti@unibz.it[2]

This talk will report on the corpus-building component of the EdiCoMC ('English in the (digital) communication of multilingual communities') research project currently underway at the Faculty of Education, Free University of Bozen-Bolzano, which aims to investigate the increased presence of English and people's perception of the use of English in South Tyrol. The corpus is currently in the first stages of development and this talk will focus on i) the data collection method we adopted – through necessity – in a period of mainly online teaching and, in the case of on-site teaching, limited access to buildings and classrooms at the university; and ii) on decisions we made on how to construct the questionnaire used to collect data on learner variables.

The Province of South Tyrol, in the far north of Italy, is a territory characterised by its multilingualism, where Germanic and Romance languages have long been in contact (Dal Negro & Ciccolone, 2020). English is neither a major nor a community language in the Province, but is present in the educational system from Kindergarten onwards, is widely used in tourism and in the international operations of local businesses, and is, along with German and Italian, one of the three official languages of the Free University of Bozen-Bolzano. The EdiCoMC project contains various components, including ethnographic and sociolinguistic investigations, and corpus creation. Due to restrictions related to Covid-19, and the associated difficulties in travelling and accessing business premises and local institutions, a decision was taken early on to restrict the focus of the research project to the immediate university environment. As part of a wider investigation on experiences of and attitudes towards the use of English at the institution among university members, and wider, questionnaire-based research, we decided to build a corpus of written English produced by students and staff at the university, focusing particularly on those who grew up in South Tyrol. One intended use of the corpus is to investigate the characteristics of English produced by South Tyroleans. We expect that while it is likely to be influenced by the L1 (mainly German or Italian), it will also show influence from the L2 (again, German or Italian). The English produced by, for example, an Italian L1 South Tyrolean may differ in some respects from the English produced by speakers from the rest of Italy, just as the English produced by a German L1 South Tyrolean may differ from that produced by L1 German speakers in monolingual German contexts.

Firstly, the talk will showcase the workflow and procedure which we are using to build the corpus. We use SurveyMonkey (www.surveymonkey.com) to collect consent (from learners), learner-related variables (from learners), task-related variables (from teachers), and corpus data itself (written texts produced by learners); once this is collected, we upload the texts to Sketch Engine (Kilgarriff et al. 2014) and tag them using the data collected. Thanks to the cooperation of colleagues teaching English courses both in faculties and at the university Language Centre, who act as 'intermediates' between the research team and learners, the creation of this 'one-stop' consent/variables/data-collection process using SurveyMonkey has allowed us to gather data and consent 'at distance', without the need for face-to-face contact with learners or their teachers.

In order to collect data on learner-related variables, we decided to use, as a basis, the Learner Profile developed for the International Corpus of Learner English (Université catholique de Louvain 2022). However, as stated above South Tyrol is a multilingual community and as such many participants submitting texts to the corpus are likely to have grown up in a multilingual family or social environment and, in any case, will have certainly been exposed to or have studied a second language at school from an early age. With this in mind, we decided to reconsider the item 'Native language' from the ICLE Learner Profile, considering the difficulty or doubt that South Tyroleans might feel in choosing a single language when answering. The talk therefore also outlines the approach taken to reformulate this item, showing how the Learner Profile was modified to consider the question of 'native language' from three points of view: i) ideological/political, ii) language acquisition and iii) sociolinguistic.

**References**

Dal Negro, S. & Ciccolone, S. (2020). 'KONTATTO: A laboratory for the study of language contact in South Tyrol.' *Sociolinguistica* 34(1), pp. 241–247. https://doi.org/10.1515/soci-2020-0014

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7–36.

Université catholique de Louvain (2022) *Learner Profile*. Available at https://cdn.uclouvain.be/public/Exports%20reddot/cecl/documents/LEARNER_PROFILE.txt.

# *Do you love me*: Interrogatives in learner speech in LINDSEI and in the Trinity Lancaster corpus

Sylvie De Cock
Centre for English Corpus Linguistics, UC Louvain
sylvie.decock@uclouvain.be

The Louvain International Database of Spoken English Interlanguage (LINDSEI) contains informal interviews with intermediate to advanced level learners of English as a foreign language from a series of mother tongue backgrounds. However informal these interviews may be, they do not share two of Clark's (1996) typical features of face-to-face conversation, namely self-determination, and self-expression. While the free exchange of turns is a fundamental organizing factor of conversations, in interviews the participants do not determine for themselves what actions to take and when. Instead of being locally managed as in conversations (Lazaraton 1992), the turn-taking system is pre-specified: interviews are organized according to a question-answer format. Besides taking action as themselves (Clarke's self-expression) the participants in an interview also take action as interviewer or interviewee. As Fiksdal (1990) points out, the participants have rights and obligations as interviewer or interviewee: the interviewer has the right and obligation to ask questions and the interviewee has the obligation to answer these questions.

This paper reports on research into the use of interrogative clauses, and more specifically *Wh*-questions and *yes/no*-questions (Biber et al. 1999), by the learner interviewees in four of the subcorpora, included on the LINDSEI CD-ROM (Gilquin et al. 2010). The following research questions are addressed: (1) to what extent do the learner interviewees use interrogatives in a context that arguably does not foster the use of these structures (and how does this use compare with the use of interrogatives in spontaneous conversations reported by Biber et al. 1999), and (2) what are the discourse/pragmatic functions of the interrogatives used by the interviewees?

The four corpora investigated represent very different mother tongue backgrounds, namely LINDSEI_Chinese, LINDSEI_Dutch, LINDSEI_French, and LINDSEI_Polish. They contain between c. 60,000 and 90,000 words of interviewee speech and the interviews all follow the same set pattern. The Concord tool in WordSmith Tools 8.0 is used to retrieve the instances of *wh*-words and primary and secondary auxiliaries from the interviewee turns. The automatic retrieval is followed by careful analysis of the concordance lines to uncover the actual *Wh*-questions and *yes/no*-questions in the data.

The paper focuses more particularly on the discourse/pragmatic functions of the interrogatives uncovered in the data. The following four main discourse/pragmatic functions have been identified in the LINDSEI data: direct speech/thought reporting (e.g. *why do you have this American accent like that, do you love me* ), speech management (Allwood et al. 1990, Rühlemann 2006; e.g. *what else did we do* and direct appeals for assistance like *how do you call that in (.) in English*, Tarone *et al*. 1983), elicitation of information from the interviewer (for example to assess or establish common ground; *were you there as well*) and interview/task-oriented metadiscursive function (e.g. *is it anonymous, can I start*).

In a second part, the paper explores the extent to which interrogatives are used to a similar extent and with similar pragmatic/discourse functions by learners in the Conversation and Discussion subset of the Trinity Lancaster Corpus. The Trinity Lancaster Corpus (henceforth TLC, Gablasova et al. 2019) includes spoken data produced by learners of English from over ten different mother tongue backgrounds within the framework of the Graded Examinations of Spoken English (developed and organized by Trinity College London). The Conversation and Discussion subset feature data produced by learners taking the spoken English exam in the context of speaking tasks which have been characterized as both dialogic and jointly-led (Gablasova et al. 2019). The main focus is on a qualitative functional analysis of the interrogatives used by the learners in the TLC and LINDSEI data under study and discusses the impact of differences in the formality of the setting (semi-formal in the TLC vs. more informal in LINDSEI, Gablasova et al. 2019), in speaker roles (candidate and examiner in the TLC vs. interviewee and interviewer in LINDSEI) and in turn-taking format (jointly-led interaction - Gablasova et al. 2019 - vs. question and answer format in LINDSEI).

**References**


Allwood, J., Nivre, J. & Ahlsén, E. (1990). Speech management: on the non-written life of speech. *Nordic Journal of Linguistics*, 13, 3–48.

Biber, D., Johansson, S., Leech, G., Conrad, S. & E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Fiksdal, S. (1990). *The Right Time and Pace: A Microanalysis of Cross-cultural Gatekeeping Interviews*. New Jersey: Ablex Norwood.

Gablasova, D., Brezina, V. & McEnery, T. (2019) The Trinity Lancaster Corpus. Development, description and application. *Interntional Journal of Learner Corpus Research,* 5(2), 126-158.

Gilquin, G., De Cock, S. & S. Granger (Eds.) (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Lazaraton, A. (1992). The Structural Organization of a language Interview: A Conversation Analytic Perspective. *System*, 20(3), 373-386.

Rühlemann, C. (2006) Coming to terms with conversational grammar: 'Dislocation' and 'dysfluency'. *International Journal of Corpus Linguistics*, 11 (4), 385–409.

Tarone, E., Cohen, A., & Dumas, G. (1983). A Closer Look at Some Interlanguage Terminology: A Framework for Communication Strategies. In Faerch, C. & G. Kasper (Eds.). *Strategies in Interlanguage Communication.* London: Longman, 4-14.

# Extending experimental research on the effectiveness of an intelligent tutoring system: A corpus study systematically identifying targeted language means in authentic ESL student essays

Kordula De Kuthy[1], Detmar Meurers[2]
University of Tübingen
de-kuthy@uni-tuebingen.de[1] dm@uni-tuebingen.de[2]

According to Eurostat, there were almost 22 million school children in upper secondary schools (ISCED level 3, aged 14–18) in Europe in 2016, with 94% learning English, which makes this an important population to study. Yet, school children in their authentic learning context are hardly investigated by Second Language Acquisition (SLA) research, which typically targets readily accessible adult populations such as college students. Digital learning environments can help address this problem: When Intelligent Language Tutoring Systems (ITS) are introduced in regular foreign language classes in schools (Rudzewitz et al., 2017), randomized controlled field trials (RCT) can be set up to study instructed SLA in an authentic school context. Meurers et al. (2019) showed in the first RCT carried out with an ITS in German schools that specific feedback provided to students while they work on practice exercises effectively fosters the acquisition of the targeted language means. Following standard SLA methodology, the experimental pre-/posttest design measuring the learning gains used hand-constructed test items. At the same time, there is considerable controversy around the question of whether grammar practice generalizes and transfers to free production (Ur, 2016).

In this paper, we, therefore, explore the student performance on a more ecologically valid, free writing task collected at the end of the RCT. As the first research question, we investigate whether the language means that were practiced and tested were actually used more in the free writing of the students who scored better on the tests. Complementing this analysis specific to the targeted language means, we also analyze the free writing in terms of a more general, second research question: Is the linguistic complexity of the free writing task predictive of the overall English school grade assigned by the teacher, and, if so, which aspects of linguistic complexity? Our investigation is based on an English learner corpus consisting of essays written by students from 13 classes (N=325) at the end of a school year in which they used the FeedBook ITS (Rudzewitz et al., 2017) in place of the printed workbook. The students recruited for the study came from seventh-grade classes in four German high schools (Gymnasium), where English is taught as the first foreign language. In addition to the written essays, other variables relevant for ESL-research questions were collected, including the English grade assigned by the teacher at the end of the year. The language means covered by the FeedBook ITS are those from the official school English curriculum: tenses, progressive aspect, gerunds, comparatives, conditional clauses, relative clauses, past perfect, passive, reported speech, and reflexives. Within-class randomization was used to split the students into two groups, which differed in the language means for which the system provided the specific scaffolded feedback. The students showed significantly higher learning gains for language means they had received specific feedback (Meurers et al., 2019), with a medium effect size (Cohen's $d = 0.56$).

To collect the texts at the end of the school year, the students were asked to complete a free writing task using the following task prompt: "Write a text about your holidays. Please include the following aspects: Compare two of your holiday trips (weather, duration, …), describe your next holiday trip, and outline what you would do if you could spend 1000 € during your next holiday." The task prompt was chosen so that the language means that were practiced by the students using the FeedBook exercises could meaningfully be employed to complete the task.

The data collected in the classes resulted in 325 texts ranging from 100 to 500 words. For the current paper, the hand-written texts were transcribed by two annotators, and we base our analysis on a gold transcription produced by the first author based on the two transcripts and the original scan. For the analysis of the pedagogically targeted language means, we make use of a computational linguistic approach for identifying such curricular target constructions (Quixal et al., 2021). On this basis, we investigate whether the children that improved more on a given grammar topic for which they received specific feedback also make more frequent use of those language means in the free production activity. For the second research question, we use CTAP (Chen & Meurers, 2016) to analyze the linguistic complexity of the learner texts. The system provides a broad range of linguistic complexity characteristics that can be correlated with the overall academic achievement of the students in English as measured by the English grade assigned by the teacher. Finally, the specific and the global analyses can also be combined to address interaction effects, such as the question of whether students with good English

grades use more of the language means they got feedback on. In combining these analyses' perspectives, the study helps advance our understanding of the under-researched second language learning of school children in a real-life context.

**References**


Chen, X. & D. Meurers (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*. Osaka, Japan: COLING, 113–119. https://aclanthology.org/W16-4113.pdf

Meurers, D., K. De Kuthy, F. Nuxoll, B. Rudzewitz & R. Ziai (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics 39*, 161–188. URL https://doi.org/10.1017/S0267190519000126

Quixal, M., B. Rudzewitz, E. Bear & D. Meurers (2021). Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*. pp. 15–27. https://aclanthology.org/2021.nlp4call-1.2.pdf

Rudzewitz, B., R. Ziai, K. De Kuthy & D. Meurers (2017). Developing a web-based workbook for English supporting the interaction of students and teachers. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition.* pp. 36–46. http://aclweb.org/ anthology/W17-0305.pdf

Ur, P. (2016). Grammar practice. In Hinkel, E. (Ed.) *Teaching English grammar to speakers of other languages*, 109-27. New York, NY: Routledge.

# The use of connectors in spoken and written argumentative texts of Indonesian EFL learners: A corpus-based study

Nida Dusturia
University of Bremen
dusturia@uni-bremen.de

This paper compares the use of connectors in argumentative writing and spoken monologues by Indonesian EFL learners at different proficiency levels of A.2. and B.1.2 of the Common European Framework of Reference for Languages (Council of Europe 2001) and English native speakers.

Connectors (a.k.a. linking adverbials, see Biber et.al 1999: 875) serve a connective function and make explicit links between two units of discourse in both speaking and writing; however, many studies have found that connectors tend to be overrepresented in the writing of EFL learners; additionally, some studies have found non-target like uses in EFL writing (e.g. Granger & Tyson 1996; Altenberg and Tapper 1998; Aijmer & Strensöm 2004, Callies 2009). Similar findings have been produced in research on Asian EFL learners (Field & Yip 1992; Bolton & Nelson 2002; Chen 2006; Lei, 2012), and also Indonesian EFL learners (Swan & Smith 2001). However, there is a general lack of comparative corpus-linguistic research on the use of connectors usage in both argumentative writing and speech (Crosthwaite et.al. 2021). Hence, the present study is intended to address this research gap and examine Indonesian EFL learners' argumentative essays and spoken monologues from a learner corpus perspective.

The method used in this study is Contrastive Interlanguage Analysis (Granger 2015) which involves two types of comparison. First, a comparison between the use of connectors produced by Indonesian EFL learners at the A.2., the B.1.2 levels of the CEFR and English native speakers in written argumentative texts; and second, a comparison of connector usage in spoken monologues produced by those three groups. The data come from the International Corpus Network of Asian Learners of English (Ishikawa 2014) which provides spoken and written data produced in response to the same argumentative tasks and topics that are "It is important for college learners to have a part-time job" and "Smoking should be completely banned at all the restaurants in the country". In the compilation process, various task conditions were controlled as strictly as possible, which leads to greater reliability in varied types of contrastive analyses. The annotation is carried out by means of *UAM Corpus Tool* (O'Donnell 2015). As for the analysis, the connectors are classified into various semantic types according to their discourse function(s), such as Enumeration/Addition, Summation, Apposition, Result/Inference, Contrast/Concession, and Transition (Biber et. al. 1999). Additionally, quantitative and qualitative approaches were used to analyze the data.

The following research question is addressed:

Do Indonesian EFL learners differ from English native speakers in the use of connectors in their argumentative essays and spoken monologues in terms of:

a) the frequency of use and representation of semantic types of connectors (with a view to over-/underrepresentation)
b) potential contextual misuse of connectors, and
c) the positioning of connectors with a sentence?

The findings reveal significant differences in connectors usage between Indonesian learners and native speakers in argumentative writing and speech. Generally, connectors are more frequently used in speech than in writing. The A.2.0 level learners use more connectors in essay writing while B.1.2 learners use more connectors in speech. Additionally, Indonesian learners at both proficiency levels demonstrate misuse in the connector usage compared to the native speakers as shown in the following examples which illustrate non-target-like uses of the concessive connector *even though:*

1. We know that smoking is always banned in all of the place actually. Why? Because it has many bad risks for all of the aspects of our life, *even though* in the restaurant area in this country. (IDN SMK A.2.0 054).
2. *Even* smoking can cause a lot of disadvantages, they still won't stop smoking. (IDN SMK B.1.2. 111).

The present study also confirms previous findings in that the Indonesian EFL learners tend to use more connectors than the native speakers. As for the positioning of connectors, the learners prefer to use connectors in clause-initial position, while it is more varied for the native speakers. The study thus provides further evidence for the

assumption that there is a general tendency for learners to place connectors in the initial position irrespective of their L1 (Van Vuuren & Berns 2018).

**References**


Aijmer, K. & Strensöm, A.B. (eds.). (2004). Discourse Patterns in Spoken and Written Corpora. Amsterdam: Benjamins.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer*. Harlow: Addison Wesley Longman, 80- 93.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bolton, K. & Nelson, G. (2002). Analyzing Hong Kong English. Sample texts from the International Corpus of English. In K. Bolton (ed.): *Hong Kong English. Autonomy and Creativity*. Hong Kong: Hong Kong University Press, 241–264.

Callies, M. (2009). *Information Highlighting in Advanced Learner English: The Syntax- Pragmatics Interface in Second Language Acquisition*. John Benjamins. https://doi.org/10.1075/pbns.186

Chen, C. W. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners, *International Journal of Corpus Linguistics* 11(1), 113-130.

Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning Teaching, Assessment*. Cambridge: Cambridge University Press.

Crosthwaite, Peter, Lusiana, & Schweinberger, M. (2021). Voices from the periphery: Perceptions of Indonesian primary vs secondary pre-service teacher trainees about corpora and data driven learning in the L2 English classroom. *Applied Corpus Linguistics*, 1(1), 1-13.

Field, Y., & Yip Lee Mee, O. (1992). A comparison of internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal 23*(1), 15-28.

Granger, S. and Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English, *World Englishes*, 15(1), 17-27.

Granger S. (2015). Contrastive Interlanguage Analysis: A reappraisal, *International Journal of Learner Corpus Research*, 1(1), 7-24.

Ishikawa, S. (2014). The International Corpus Network of Asian Learners of English. http://language.sakura.ne.jp/icnale/index.html

Ishikawa, Shin'ichiro. (2015). A consideration of the difference between the spoken and written English of native speakers and Japanese learners: A corpus-based study. *Discourse and Interaction*, 8, 37-52.

Lei, L. (2012). Linking adverbials in academic writing on applied linguistics by Chinese doctoral students. Journal of English for Academic Purposes, 11(3), 267-275.

O'Donnell, M. (2015). *UAM Corpus Tool*. Version 3.1.17. Available from http://www.wagsoft.com/CorpusTool/index.html.

Swan, M. & Smith, B. (2001). *Learner English: A Teacher's Guide to Interference and Other Problems*. 2nd Edition. Cambridge: Cambridge University Press.

Van Vuuren, S. & Berns, Janine. 2018. Same difference? L1 influence in the use of initial adverbials in English novice writing. *International Review of Applied Linguistics in Language Teaching***,** 56(4): 427–461.

## A multifactorial learner corpus approach to genitive alternation in non-native English

Jane Klavan
University of Tartu
jane.klavan@ut.ee

Native speakers' choice between the use of the s-genitive and the of-genitive has been shown not to be free – there are probabilistic constraints that determine the choice (e.g. Heller et al. 2017). Only a few studies have looked at the genitive alternation in non-native English (e.g. Azaz 2020, Di Domenico & Bennati 2007, Gries & Wulff 2013, Marinis 2016); there are no studies with Finno-Ugric L1. The present study investigates the genitive alternation in Estonian EFL learners and whether (Estonian) EFL learners share a core probabilistic grammar with users of first language varieties of English. The research question is: How do Estonian EFL learners' genitive construction preferences compare with those of British English as L1? It is predicted that both native speakers and Estonian EFL learners are influenced by animacy and length when choosing between the two genitive constructions. In addition, it is predicted that the Estonian EFL learners make fewer nativelike choices in the case of the s-genitive.

A multifactorial learner corpus approach is taken to answer the research question. The data of Estonian EFL learners is comprised of short argumentative essays (233 in total, av. length = 381 words; total size = 85,173 words) written by 180 1st year BA students of English Language and Literature at the University of Tartu (Cambridge English Scale: Level C1). The comparable native speaker data is taken from the LOCNESS corpus (Granger 1998), specifically the British pupils' A-level essays (60,209 words; 114 essays) and the British university students' argumentative essays (19,019 words; 33 essays). The total size of the native speaker corpora used for the present study is 79,228 words. Both sets of data were POS-tagged using NLTK in Python. The two genitive constructions were extracted using AntConc (Anthony 2020). For native speaker data 142 s-constructions (18 uses per 10,000) and 1,104 of-construction were extracted; for Estonian EFL learner data 523 s-constructions (61 uses per 10,000) and 1,049 of-constructions.

Instead of simplistic frequency counts, a multifactorial corpus-linguistic approach is advocated (Gries 2018, Paquot & Plonsky 2017). Both sets of data were annotated manually for the variables of animacy, complexity, and length. Generalized linear mixed-effects regression (GLMM) is applied to the native speaker data to develop a model to find out when native speakers choose the s-genitive rather than the of-genitive on the basis of the variables coded. This model (C-index = 0.89, model accuracy = 90%) is used to generate a native-speaker prediction for every data point in the learner data which is then compared to the actual learner choices to see where the choices agree and disagree. A second model with the binary variable "AGREEMENT" as the dependent variable is run (C-index = 0.89, model accuracy = 82%). The results show that the Estonian EFL learners make less nativelike choices when the possessor is animate, for example choosing the s-genitive where of-genitive is more appropriate one of the funeral procession's cars, the contemporary week's event, a part of Northern Ireland's history.

The paper contributes to the discussion of probabilistic grammar (Bod et al. 2003) by extending the field to EFL learners, particularly learners of non-Germanic L1s. The study aims to start exploring whether Estonian EFL learners share a core probabilistic grammar with users of first language varieties of English. Once we have a clear descriptive picture of learner language, we can further advance the pedagogical need of compiling authentic teaching and learning materials that directly address the needs of the learners.

**References**

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Azaz M. (2020). Structural surface overlap and derivational complexity in crosslinguistic transfer: Acquisition of English genitive alternation by Egyptian Arabic-speaking learners. *Second Language Research*, 36(4), 529-556. DOI: 10.1177/0267658319834860

Bod, R., Hay, J., & S. Jannedy (eds.). (2003). *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Di Domenico, E., & E. Bennati (2007). The Alison's cat sleep in the kitchen: On the acquisition of English's Genitive Constructions by native speakers of Italian. *Studies in Linguistics* 2.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.) *Learner English on Computer*. Addison Wesley Longman: London & New York, 3-18.

Gries, Th. S., & S. Wulff (2013) The genitive alternation in Chinese and German ESL learners. Towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics*, 18(3), 327-356.

Gries, Stefan Th. (2018) On over-and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies*, 1(2), 276-308.

Heller, B., Szmrecsanyi, B. & J. Grafmiller (2017). Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics*, 45 (1), 3-27.

Marinis, T. (2016). Acquiring Possessives. In J. L. Lidz, W. Snyder, & J. Pater (eds.) *The Oxford Handbook of Developmental Linguistics*. OUP: Oxford. DOI: 10.1093/oxfordhb/9780199601264.013.19

Paquot, M., & L. Plonsky (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, 3(1), 61-94.

# Exploring the use of dependency parsing in automatic erroneous collocation extraction in learner English

Jen-Yu Li, Thomas Gaillat, Elisabeth Richard
Linguistique Ingénierie et Didactique des Langues (LIDILE), Université Rennes 2
jen-yu.li@univ-rennes2.fr

Second language learners usually encounter difficulties in collocations both for writing and for oral expression (Garner et al., 2020; Granger & Larsson, 2021; Nesselhauf, 2003; Uchihara et al., 2021). As a subset of semantic phraseme, collocation is not free: without freedom of selection of its signified and without freedom of combination of its components (Mel'čuk, 1998; Tutin, 2013). In this sense, we considered a collocation is erroneous when it is not a standard prefabricated pattern, for example, "to **create [construct] a taller building*". The correction of collocations in written essays could help learners increase their competence and thus their proficiency in English writing (Meunier & Granger, 2008). A collocation extraction module could facilitate corpus-based phraseological analysis and thus help understand the interlanguage development stages (Liu & Lu, 2020; Schneider & Smith, 2015). It could also be incorporated into a computer-assisted language learning (CALL) system to help learners write and use collocations appropriately.

A prototype of automatic Verb-Noun (VN) collocation detection was developed and reported in our previous study (Li & Gaillat, 2020). We found that there were mainly six causes that degraded the performance (Li & Gaillat, 2021). First was that some extracted pairs were frequent composites but not collocations, for example, "*see section*", "*thank president*". The second was that the noun was not the object of the verb, for example, "*give (something) (to) dog*", "*remove (something) (from) heat*". Other remaining causes were in principle due to Part-Of-Speech (POS) errors. Recently, dependency parsing was reported to improve the quality of collocation extraction (Uhrig & Proisl, 2012). Dependency parsing reveals the pairwise syntactic relations between words in the sentences, i.e. the dependence of a word on a head-word. We considered that syntactic parsing is a line of research to explore. On the previous basis, we shall use a dependency parser to further improve the performance of our module. With dependency annotation, we may restrict the word association analysis to those pairs that have a specific grammatical relation: for example, the VN pairs in accusative cases in which the noun is the direct object (*dobj*) of the verb. This may also help find long-distance pairs that are outside a window of adjacent tokens.

Uhrig et al. (2018) systematically studied various dependency parsers and schemes for the extraction of standard collocations. However, since automatic parsers are generally developed for native language data, Huang et al. (2018) demonstrated that, despite the high accuracy (>80%), parsers built from standard English are not robust to learner errors: 63% of the learner errors caused at least one parsing error. Berzak et al. (2016) proposed a Treebank of Learner English and measured the effect of grammatical errors on parsing accuracy. Yet, to the best of our knowledge, the impact of dependency parsing on erroneous collocation extraction has not been studied.

This research presents a work-in-progress report about extracting erroneous VN collocations. The research question is: can a dependency parser built from standard English data be used to extract erroneous collocations in a learner corpus? For evaluation, we used the National University of Singapore Corpus of Learner English (NUCLE) (Dahlmeier et al., 2013) where grammatical errors, including wrong collocations, are annotated by professional instructors. The main principle is, firstly, to extract all possible collocations in the learner corpus, and then identify standard collocations by comparing extractions with examples from a reference corpus (in our case collocations extracted from British National Corpus) (Li & Gaillat, 2021); the remainder of the items are considered as erroneous collocations. We will compare the performance (precision and recall rate) of the extraction of erroneous collocations on the basis of dependency relations to the previously developed window-based approach. The results will be evaluated by manual inspection to investigate the effect of learner errors on dependency parsing and erroneous collocation extraction. Future improvement is envisaged based on a large web corpus as a reference (Paquot et al., 2021) and other statistical measures (Gries & Durrant, 2020).

Note: NUCLE is a collection of 1,414 essays (in a total of 1.2 million words) written by students who are non-native English speakers. It is available by submitting a license agreement via https://www.comp.nus.edu.sg/~nlp/corpora.html

**References**


Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., & Katz, B. (2016). Universal Dependencies for Learner English. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 737–746. https://doi.org/10.18653/v1/P16-1070

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 22–31.

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

Garner, J., Crossley, S., & Kyle, K. (2020). Beginning and intermediate L2 writer's use of N-grams: An association measures study. *International Review of Applied Linguistics in Language Teaching*, 58(1), 51–74. https://doi.org/10.1515/iral-2017-0089

Granger, S., & Larsson, T. (2021). Is core vocabulary a friend or foe of academic writing? Single-word vs multi-word uses of thing. *Journal of English for Academic Purposes*, 52, 100999. https://doi.org/10.1016/j.jeap.2021.100999

Gries, S.T. & Durrant, P. (2020). Analyzing Co-occurrence Data. In: Paquot, M. & Gries, S. T. (Eds) *A Practical Handbook of Corpus Linguistics*. Springer, 111–140. https://doi.org/10.1007/978-3-030-46216-1_7

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1), 28–54. https://doi.org/10.1075/ijcl.16080.hua

Li, J.-Y. & Gaillat, T. (2020). Automatic detection of unexpected/erroneous collocations in learner corpus. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, 101–106, online. Association for Computational Linguistics.

Li, J.-Y. & Gaillat, T. (2021). Extraction of Standard collocations from British National Corpus, *Europhras 2021*, Sep 2021, online, European Association for Phraseology

Liu, Y. & Lu, X. (2020). N1 of N2 constructions in academic written discourse: A pattern grammar analysis. *Journal of English for Academic Purposes*, 47, 100893. https://doi.org/10.1016/j.jeap.2020.100893

Mel'čuk, I. (1998). Collocations and Lexical Functions. In Cowie, A. P., (Ed), *Phraseology: theory, analysis, and applications*. Oxford: OUP, 23–53.

Meunier, F., & Granger, S. (Eds.). (2008). *Phraseology in foreign language learning and teaching*. John Benjamins Pub. Co.

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, 24(2), 223–242. https://doi.org/10.1093/applin/24.2.223

Paquot, M., Naets, H. & S. Th. Gries. (2021). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: verb + object structures in LONGDALE. In Le Bruyn, B. & Paquot, M. (eds.). *Learner Corpus Research Meets Second Language Acquisition*, Cambridge University Press, 122-147.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., & Smith, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 455–461, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Schneider, N. & Smith, N. A. (2015). A corpus and model integrating multiword expressions and super- senses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1537–1547, Denver, Colorado, May–June. Association for Computational Linguistics.

Tracy-Ventura, N. & Paquot, M. (Eds.). (2021) *The Routledge handbook of second language acquisition and corpora*. New York: Routledge

Tutin, A. (2013). Les collocations lexicales : une relation essentiellement binaire définie par la relation prédicat-argument. *Langages*, 89(1), 47–63.

Uchihara, T., Eguchi, M., Clenton, J., Kyle, K., & Saito, K. (2021). To What Extent is Collocation Knowledge Associated with Oral Proficiency? A Corpus-Based Approach to Word Association. *Language and Speech*, 238309211013865. https://doi.org/10.1177/00238309211013865

Uhrig, P., Evert, S., & Proisl, T. (2018). Collocation Candidate Extraction from Dependency-Annotated Corpora: Exploring Differences across Parsers and Dependency Annotation Schemes. In P. Cantos-Gómez & M. Almela-Sánchez (Eds.), *Lexical Collocation Analysis*. Springer International Publishing, 111–140. https://doi.org/10.1007/978-3-319-92582-0_6

Uhrig, P., & Proisl, T. (2012). Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates. *Lexicographica*, 28(2012), 141–180. https://doi.org/10.1515/lexi.2012-0009

# The acquisition and use of the progressive aspect by multilingual learners of English as L3: Preliminary results from a longitudinal learner corpus-based study

Olga Lopopolo
Eurac Research
olga.lopopolo@eurac.edu

The recent interest in the analysis of Tense/Aspect (TA) in corpus-based SLA studies (Leńko-Szymańska 2007, Dose-Heidelmayer & Götz 2016, Fuchs & Werner 2018, Díez-Bedmar 2021) has helped to test the most prominent theories in the field as well as to describe patterns of use through frequency-related information. Especially at the early stages, many researchers have found that in both L1 and L2 acquisition language learners tend to communicate events by associating certain grammatical morphemes to specific *Aktionsart* categories of verbs and that this association is highly dependent on the inherent semantic features of verbs. This general claim has been postulated as the Aspect Hypothesis (AH) by Andersen and Shirai (1994) and it is now considered the most widely discussed SLA theory in the domain of TA acquisition (Fuchs & Werner, 2018). Many of these studies have been tested empirically against diverse sets of L1 backgrounds, leading to the conclusion that the AH in its 'strong form' is not "an absolute acquisitional universal" (Housen 2000), rather it was found to interact with other determinants, notably cross-linguistic influence (CLI), proficiency level of the learners, task type, the distribution of the input and individual preferences.

Within this framework, the goal of my investigation is to provide further evidence from longitudinal corpus data about the acquisition of the progressive aspect by learners of English as a third language instructed at schools. The corpus analyzed is the English subsection of LEONIDE (Glaznieks et al. 2022), a longitudinal trilingual learner corpus collected over three years of secondary school in the Italian Province of Bozen.

The choice of the progressive aspect in this ongoing PhD project is motivated by different reasons. First of all, the combination of the two typologically different languages of the environment (Italian and German) that morphologically encode progressive in different ways, could shed light on CLI phenomena in L3. Secondly, very few studies concerning progressive aspect look at the initial stages of L3 learning using longitudinal learner corpora in a foreign language context. As a final and general consideration, there are no corpus-based studies that consider the heterogeneity of multilingual environments and provide a socially embedded qualitative perspective on the learners' metadata. To the factors already tested in several studies (task type, proficiency level, age of exposure), I added new dimensions of language use considering the private and social domain, providing different language constellations and profiles for each learner.

The following research questions have been formulated:

1) Is there a relation between specific semantic classes and a progressive vs non-progressive constructional choice?
2) Is there any development towards a target like usage of progressives over three years of secondary school?
3) Which factor/s among learner L1, task type, age of exposure, and private and social use of the languages (Italian and German) have an impact on the acquisition of progressive aspects in English as L3?

In order to answer the 1st RQ, in my presentation I will test the predictions of the AH empirically on LEONIDE data, looking at learners' choices of progressive vs non- progressive constructions across the four traditional Vendlerian *Aktionsart* categories (states, activities, accomplishments, achievements) and Biber's taxonomy of semantic domains (Biber et al. 1999). A multi-layer annotation scheme was developed to manually annotate all verbal forms of the English texts considering aspect as the dependent variable with only two levels (progressive and non-progressive) and two independent variables, i.e. *Aktionsart* and semantic domain. It is also worth noting that the manual annotation procedure was necessary to also annotate instances of code-switchings, foreignizings, and innovations in the use of progressives. Annotations have been imported into ANNIS and raw frequencies of progressive vs non-progressive constructions have been extracted. To predict learners' choices of progressive vs non-progressive constructions, and their relation with semantics, a logistic regression analysis was conducted using a generalized linear model.

In order to answer the 2nd RQ, I will show the longitudinal development of target-like and non-target-like usage of progressives over three years of schooling mapping different form-function combinations. Raw frequencies of progressives have been automatically extracted and normalized by considering the total number of

predicates per text and learner. It followed a second round of annotations concerning form-function combinations. This was essential to analyse the rates of overextensions, target-like usages and avoidances over time using linear mixed-effect models.

The 3rd RQ will be answered through a qualitative analysis of the questionnaires and a clearer picture of learners' everyday language usage in different contexts and domains (family, school, friends) is in the process to be defined. This last step won't be part of this presentation but in the future will give reasons for the activation of particular languages in multilingual production as well as the relative weight and interaction of the different factors in a more holistic manner.

**References**


Andersen, R. W., & Shirai, Y. (1994). Discourse motivations for some cognitive acquisition principles. *Studies in Second Language Acquisition*, 16(2), 133-156.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson.

Díez-Bedmar, M. B. (2021). The use of the progressive in light of the AH in monolingual EFL-instructed spanish learners at university level: A longitudinal learner corpus-based SLA study. *Circulo de Linguistica Aplicada a La Comunicacion*, 87, 53–69.

Dose-Heidelmayer, S., & Götz, S. (2016). The progressive in spoken learner language: a corpus-based analysis of use and misuse. *International Review of Applied Linguistics in Language Teaching*, 54(3), 229-256.

Fuchs, R., & Werner, V. (2018). Tense and aspect in Second Language Acquisition and learner corpus research: Introduction to the special issue. *International Journal of Learner Corpus Research*, 4(2), 143-163.

Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L. & Nicolas, L. (2022). LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, 8(1), 97-120.

Housen, A. (2002). The development of tense-aspect in English as a second language and the variable influence of inherent aspect. In M. R. Salaberry & Y. Shirai, *The L2 acquisition of tense-aspect morphology* Amsterdam: John Benjamins, 155-197.

Leńko-Szymańska, A. (2007). Past progressive or simple past? The acquisition of progressive aspect by Polish advanced learners of English. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the Foreign Language Classroom: Selected Papers from the 6th International Conference on Teaching and Language Corpora*, Amsterdam: Rodopi, 253-267.

# "So I'll need English like good English" – Functions and use of discourse marker *like* in UAE English

Eliane Lorenz
Justus Liebig University Giessen
eliane.lorenz@anglistik.uni-giessen.de

The current study investigates the use of English as a Lingua Franca (ELF) in the United Arab Emirates (UAE), focusing on the functions and use of the discourse marker *like* among university students. It is set in Sharjah, one of UAE's seven sovereign emirates, a metropolitan area characterized by intense language contact due to recent, large-scale immigration (Parra-Guinaldo & Lanteigne 2021; Pacione 2005; Siemund et al. 2021). To date, there is a lack of research investigating the use of ELF in the UAE and its status as a new English variety (Siemund et al. 2021). The discourse marker *like* has received much scholarly attention (e.g., D'Arcy 2017; Diskin 2017; Fuller 2003; Schweinberger 2014). However, it has mainly been studied in native Englishes, and considerably less research focuses on non-native speakers of English (Diskin 2017; Rüdiger 2021) or ELF varieties.

The current study addresses these research gaps by employing a small-size spoken ELF corpus consisting of semi-structured interviews, approximately 30 minutes each, conducted with 58 university students in the UAE (word tokens: 139,630). The participants come from a variety of linguistic backgrounds including both Emirati as well as non-Emirati population and are advanced users of English. The interviews were conducted as part of a larger project on *Language Attitudes and Repertoires in the Emirates* (LARES 2019–2021) and targeted family background, educational history, language biographies, as well as attitudes towards the language of the students' repertoires, i.e., English, Arabic, and others. The spoken data are complemented by a comprehensive online questionnaire. With this unique data source, it is possible to investigate the use of *like* and to correlate it with different social (non-linguistic, attitudinal) variables.

The study sets out to answer three research questions:

(1) Do the UAE students show high individual variation as has been argued to be a characteristic of ELF users (e.g., Mauranen 2017)?
(2) Does this study find support for an assumed accelerated language change in ELF contexts (e.g., Laitinen 2020)?
(3) Are there functional differences in discourse maker *like* uses and if yes, can these be explained with the social background of the students?

First results show that *like* is the third most frequently used word in the interviewees' utterances (n=3,937), with 2,951 (75%) uses as a discourse marker and 986 (25%) other uses. The mean frequency per 1,000 words (ptw) across the entire corpus is 19.5 (median: 16.0). This lends support to an accelerated language change in this particular ELF setting because other studies have found considerably lower frequencies of *like*, for example, 0.49 ptw in British English and 2.23 ptw in Philippine English (Schweinberger 2014), 11.6 ptw in American English (Fuller 2003), or approximately 8 ptw in Korean English (Rüdiger 2021). The specific register, i.e., rather informal, personal interviews, may partly explain the high frequency found among the UAE students.

Yet, the relatively high standard deviation in the LARES corpus of 14.75 shows that the individual variation among the ELF speakers in Sharjah is comparably large. The lowest frequency is 0.51 ptw and the highest is 55.14 ptw. This is in line with Mauranen (2017) who argued for variability in ELF encounters. A regression analysis shows that the social background of the speakers (gender, citizenship, dominant language, year of birth, number of languages, college, self-assessed proficiency in English, and English usage score) cannot explain the variability identified in the use of the discourse maker *like*.

An additional (functional) coding (sentence position, i.e., clause-initial, medial, final, and non-clausal (see Schweinberger 2014); co-occurrence with hesitation or other stylistic markers such as *well*, *so*, *let's see*), will further assess the use of *like* among the UAE students.

**References**


D'Arcy, A. (2017). *Discourse-pragmatic variation in context. Eight hundred years of LIKE*. Amsterdam: Benjamins.

Diskin, C. (2017). The use of the discourse-pragmatic marker 'like' by native and non-native speakers of English in Ireland. *Journal of Pragmatics* 120, 144-157.

Fuller, J. M. (2003). Use of the discourse marker like in interviews. *Journal of Sociolinguistics*, 7(3), 365-377.

Laitinen, M. (2020). Empirical perspectives on English as a Lingua Franca (ELF) grammar. *World Englishes*, 39(3), 427-442.

Mauranen, A. (2017). A glimpse of ELF. In M. Filppula, J. Klemola, A. Mauranen & S. Vetchinnikova (Eds.). *Changing English. Global and local perspectives*. Berlin: De Gruyter Mouton, 223-253.

Pacione, M. (2005). Dubai. *Cities*, 22(3), 255-265.

Parra-Guinaldo, V. & Lanteigne, B. (2021). Morpho-syntactic features of transactional ELF in Du-bai/Sharjah. In P. Siemund & J. R. E. Leimgruber (Eds.). *Multilingual global cities: Singapore, Hong Kong, Dubai*. Singapore: Routledge, 303-320.

Rüdiger, S. (2021). Like in Korean English speech. *World Englishes*, 40(4), 548-561.

Schweinberger, M. (2014). *The discourse marker LIKE: A corpus-based analysis of selected varieties of English*. Doctoral dissertation. University of Hamburg.

Siemund, P. Al-Issa, A. & Leimgruber, J. R. E. (2021). Multilingualism and the role of English in the United Arab Emirates. *World Englishes*, 40(2), 191-204.

## Towards more appropriate modeling of (and with) linguistic complexity indices

Akira Murakami
University of Birmingham
a.murakami@bham.ac.uk

Linguistic complexity indices have been used widely within and beyond learner corpus research (e.g., Alexopoulou et al., 2017; Housen et al., 2019; Vyatkina & Housen, 2021). Complexity indices often take the form of ratios (e.g., number of dependent clauses per T-unit). However, currently, practically all analyses of such indices in the field are based on linear regression models (e.g., multiple regression), including their special cases (e.g., ANOVAs). This is a problem because, firstly, there is a fundamental mismatch between the statistical model and the data-generation process that it is supposed to represent. Secondly, those models assume that the dependent variable has no theoretical bound (i.e., it can theoretically take the value between negative infinity to positive infinity). Those models, therefore, could yield a prediction that is beyond the theoretical limits of target complexity indices (e.g., negative mean sentence length). Finally, the prediction interval based on Gaussian linear regression models fails to take into account the difference in the denominator of the ratio representing complexity indices. For instance, both a learner essay with two dependent clauses in 10 T-units and an essay with six dependent clauses in 30 T-units have the same number of dependent clauses per T-unit (i.e., 0.2). The latter, however, is more reliable due to the larger number of T-units. When the outcome variable is the dependent clauses per T-unit as in typical linear regression models, the difference is ignored in prediction, and they yield the same prediction interval for the observations with different numbers of T-units as long as the values of the other predictors are the same, thereby potentially over- or underestimating uncertainty in prediction.

In this work-in-progress talk, I will propose alternative modeling techniques in the analysis of linguistic complexity indices. Specifically, when we model the indices that take the form of ratios, I suggest that we use Poisson or negative binomial regression models (e.g., Winter & Bürkner, 2021) to model the nominator (e.g., the number of dependent clauses, that of words) and include the log-transformed denominator (e.g., the number of T-units, that of sentences) as an offset. Drawing learner writings from EF-Cambridge Open Language Database (Geertzen et al., 2014), I will empirically demonstrate that these models are free of the issues mentioned above. Specifically, the number of dependent clauses in each writing was modeled as a function of writing number (e.g., 1 indicates the first writing of a learner, 2 indicates the second learner), learner's L2 proficiency, by-learner random intercepts, by-learner random slopes for writing number, random intercepts by topic prompts, and the log-transformed number of sentences as an offset in a Bayesian mixed-effects negative binomial regression model. A similar model was built that models the mean number of dependent clauses per sentence and assumes a normal distribution of errors. The mean number of dependent clauses per sentence was then predicted based on those two models when the writing number and proficiency take certain values (i.e., writing number = 1, proficiency = 3). In the negative binomial model, the width of the prediction intervals decreased from 0.0-1.0 to 0.03-0.31 when the number of sentences increased from 1 to 100, while in the Gaussian model, the intervals remained the same and included theoretically impossible values (-0.41 to 0.80).

A similar issue is present when complexity indices are used as predictor variables. The uncertainty of the individual values of a complexity index should be reflected on its standard error when used as a predictor. This can be achieved with a variant of measurement error models, where the observed value of a complexity index is assumed to have been generated from the true value plus noise. The noise is assumed to follow a Poisson distribution and to be larger when the denominator (e.g., the number of T-units) is smaller. Since no ready-made model is available, a bespoke model was built with Stan (Stan Development Team, 2019), and I will illustrate its utility in predicting learner's proficiency based on their mean sentence length. Issues similar to the above apply to many other complexity indices (e.g., MTLD) and some association/collocation measures as well. I argue that we should pursue more appropriate modelling of those metrics.

**References**


Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language Processing techniques. *Language Learning*, *67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Geertzen, J., Alexopoulou, T. and Korhonen, A. (2014) Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In R.T. Millar, K.I. Martin, C.M. Eddington, A. Henery, N.M. Miguel, A. Tseng, A. Tuninetti and D. Walter (Eds.), *Selected Proceedings of the 2012 Second Language Research Forum. Building Bridges between Disciplines* (pp. 240–254). Somerville, MA: Cascadilla Proceedings Project.

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (Eds.) (2019). Linguistic complexity [Special Issue]. *Second Language Research*, *35*(1).

Stan Development Team. (2019). Stan modeling language user's guide and reference manual, Version 2.28. https://mc-stan.org

Vyatkina, N., & Housen, A. (2021). Complexity. In N. Tracy-Ventura and M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 318-331). New York, NY.: Routledge

Winter, B., & Bürkner, P. (2021). Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass*. https://doi.org/10.1111/lnc3.12439

# "Today's lesson was really interesting": Improving second-language learning and obtaining feedback through students' reflective Padlet posts

Carla Quinci
University of Padova
carla.quinci@unipd.it

Learning and reflective diaries are personal narratives whereby learners can observe and reflect upon their learning process, develop critical thinking, metacognition, and creativity, and improve their writing and communication skills (Wallin & Adawi 2018). Personal narratives have served as investigation tools in a number of disciplines, including sociolinguistics, cognitive psychology, and applied linguistics, with a special focus on the study of bilingualism and second-language acquisition (Khezrlou 2021; Pavlenko 2007). Recently, reflective diaries have transitioned from simple journals in paper or electronic form to multimodal and/or interactive formats, e.g. audiovisual products, online blogs, forums, and social media posts, which foster self-reflection and allow learners to share their experiences in more interactive settings (Mann & Walsh 2017: 150). This study explores precisely one of such new formats, i.e. social media posts, by adopting a still underexploited perspective allowing to (a) observe and investigate the second-language acquisition and (b) obtain feedback to monitor the effectiveness of teaching activities and material. To these ends, a corpus has been collected which includes 900 Padlet posts (61,577 tokens) written in English by over 100 Italian second-language learners throughout the second-year BA-level course in Translation Strategies. The analysis makes use of corpus linguistics tools to observe common patterns in the students' posts at the lexical and phrasal levels. The linguistic trends observed in the corpus include the use of recurrent (incorrect) lexical and phrasal patterns, especially to describe past events (e.g. "we talked about", "we learned/learnt") or express personal opinions (e.g. "I found really/very interesting", "it is (very) important to", "how important it is to/that", "it is useful/important to"). Interestingly, low-frequency items were found to include also synonyms or alternative phrasings of high-frequency words and structures referring to past events (e.g. "we touched some main points"), personal ideas and feelings (e.g. "vital", "worthwhile", "it came to my surprise", "I was stunned", "capture my interest"), as well as idioms (e.g. "the tip of the iceberg", "*stucked in my mind"). This shows how most learners still rely on limited vocabulary and syntactic structures, while others are able to craft language in more personal and/or idiomatic ways. If shared and discussed with students, these results and analysis – as long as the corpus itself – can be used to encourage then to reflect upon and self-assess their competence in L2 and further explore language. Purposely developed teaching (corpus-based) activities might also be implemented so that students widen their linguistic repertoire and learn alternative and idiomatic ways to express similar concepts. Additionally, the analysis showed how reflective posts could also serve to obtain indirect feedback from learners, monitor their interest and acquisition of new contents, and (correct) metalanguage. Drawing on sentiment analysis, words expressing positive and negative feelings towards any teaching activities, material, or contents (e.g. "captivating"/"capture", "interesting", "intriguing", "useful"/"useless", "impressed") were analysed in context to explore the students' global feedback on the course and highlight any material or activity turning out to be particularly (dis)favoured and/or (in)effective. Finally, language- and translation-specific terms and phrases were also analysed to investigate whether and how students master subject-field constructs and metalanguage. This revealed some inaccuracies in both the students' understanding of specific theoretical concepts and their use of metalanguage (e.g. "*tag readers" instead of "target readers"), which were thus further discussed and explained in later classes. In this perspective, Padlet posts also proved to be a valuable tool for formative assessment and the identification of any gaps in the teaching strategies and/or the learning process (Wallin & Adawi 2018).

**References**

Khezrlou, S. (2021). Learners' reflective practice between the repeated performances of tasks: effects on second language development. *Dutch Journal of Applied Linguistics*, *10*.

Mann, S., & Walsh, S. (2017). *Reflective Practice in English Language Teaching: Research-Based Principles*. Routledge.

Pavlenko, A. (2007). Autobiographic narratives as data in applied linguistics. *Applied Linguistics*, *28*(2), 163–188.

Wallin, P., & Adawi, T. (2018). The reflective diary as a method for the formative assessment of self-regulated learning. *European Journal of Engineering Education*, *43*(4), 507–521.

# Lexical similarity in L1 and L2 German as evidence for the structure and dynamics of the lexicon

Anna Shadrova
Humboldt-Universität zu Berlin
anna.shadrova@hu-berlin.de

A written text is not only the result of a communicative intention but also a crystallization of the procedural ongoings in the mind of a speaker during composition. While the choice of lexemes is intentional to some degree, it is also influenced by priming and self-priming (Szmrecsanyi 2005, Gries & Kootstra 2017). Priming and self-priming activation patterns are mediated by the structure of the underlying lexicon and other linguistic systems. Activated nodes may be more likely to be uttered, resulting in different lexical choices depending on the underlying structure. Diversity in lexical choice through the course of text composition can then at least partially be interpreted as evidence for an underlying structural diversity, for example at different stages of acquisition in L2, which in turn can shed light on the process of acquisition.

This contribution asks how similar L1 and L2 speakers are in their lexical choice in a task-specific corpus of German, Kobalt (Zinsmeister et al. 2012, Shadrova 2021), that is compiled from essays written by native speakers and learners of German from Belarus and China at different levels of acquisition (approx. A2-C1 of the CEFR as derived from c-test scores (onDaF, now onSET, Eckes 2010)). The corpus contains a total of 151 L2 texts (87 BEL, 64 CH, 300-1200 tokens) and 20 L1 texts (400-600 tokens).

Lexical similarity is measured as the overlap of verb and noun lexemes respectively for sets of three and four speakers. This measure may appear trivial, but it possesses conceptual validity relative to phraseological assumptions such as the idiom principle (Sinclair 1991) and the primacy of formulaic or coselectionally constrained elements over free combinatorics (Wray 2002): if speakers indeed choose from the same inventory and formulaic elements are complex signs, i.e. conceptually holistic and inseparable units, then they should be shared in identical ways by groups of speakers. This cannot be determined through cumulative corpus counts regardless of statistical techniques.

However, results indicate that this is not the case:

a) Lexical overlap is extremely low in production in L1, with an average overlap of less than four non-prompt-related nouns in sets of four speakers each (less than 6% of noun lexemes on average per speaker) and even lower numbers for verbs. These results are further corroborated in a comparison with two other task-specific corpora (Falko, Reznicek et al 2010; RUEG, Wiese, et al. 2019) and are in line with previous observations from same-situation verbalizations (Chafe 1980).

b) Lexical overlap is similarly low and surprisingly stable by learner groups at lower (A2, B1) and higher (C1) levels of proficiency (5.9-6.7 in BEL, 4.28-4.55 in CH). As in L1, lexical variability is surprisingly high in spite of identical prompt and learner background.

c) There is a significant increase ($p<0.0001$) in lexical overlap at B2 levels in both learner groups, up by 1.5-2 lexemes with a clearly right-skewed distribution towards higher values compared to the other groups. Preliminary results also suggest that B2 learners further differ in their lexical use more generally from the other groups in what appears to be an effect of lexical semantics, namely divergent use of abstraction and specificity.

Results are interpreted from a structure-mediates-process perspective, namely as the result of a hyperconnected lexicon at late-intermediate proficiency. It is hypothesized that learners in the early stages of L2 acquisition learn words from more or less isolated semantic fields. As their lexicon grows, more and more connections are established. At late-intermediate stages, this may become inefficient to navigate due to too many contextually unhelpful activations that place high demands on inhibition. At this stage, learners may converge most in their lexical use because their production is most influenced by the global effects of the target language. A reorganization is then assumed to take place, resulting in lower global and higher local connectivity, eventually giving rise to the ability to memorize coselectional constraints such as collocations via local probabilities (as opposed to global probabilities spanning the entire lexicon).

This hypothesis is coined lexical connectivity pruning in analogy to synaptic pruning, a similar process in childhood brain development, whereby the already large number of synapses at birth grows further until about age two and is then gradually cut back until late adolescence. Simulations suggest that the emergent differentiation, namely the emergence of tightly interconnected groups (communities), as opposed to one big

network where everything is tangled (sometimes referred to as a hairball graph), allows for higher functionality through structural advantages (small-world effect, cf. Lindenberger & Lövdén 2019, Millán et al. 2018, Calvo Tapia et al. 2020).

Lexical connectivity pruning integrates existing evidence for collostructional entrenchment in L1 and L2 (Gries & Wulff 2009, Siyanova-Chanturia & Martinez 2015) with the results discussed in this paper demonstrating a lack of similarity in the lexical choice between speakers. It further provides an explanation for the apparently limited learnability of collocations or coselectional constraints well into advanced levels of L2 (Paquot 2019 among others).

**References**


de Bot, K. (2006). The plastic bilingual brain: Synaptic pruning or growth? Commentary on Green et al. *Language Learning*, *56*, 127-132.

Calvo Tapia, C., Makarov, V. A., & van Leeuwen, C. (2020). Basic principles drive self-organization of brain-like connectivity structure. *Communications in Nonlinear Science and Numerical Simulation*, *82*, 105065.

Chafe, W. L. (1980). The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. Advances in Discourse Processes, vol. III. Norwood.

Croft, W. (2010). The origins of grammaticalization in the verbalization of experience. *Linguistics*, *48*(1), 1-48. https://doi.org/10.1515/ling.2010.001

Eckes, T. (2010). Fremdsprache (onDaF): Theoretische Grundlagen. *Der C-Test: Beiträge aus der aktuellen Forschung: The C-Test: contributions from current research*, *18*, 125-192.

Gries, S. T., & Kootstra, G. J. (2017). Structural priming within and across languages: A corpus-based perspective. *Bilingualism: Language and Cognition*, *20*(2), 235-250.

Gries, S. T., & Wulff, S. (2009). Psycholinguistic and corpus-linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, *7*(1), 163-186.

Lindenberger, U., & Lövdén, M. (2019). Brain plasticity in human lifespan development: The exploration-selection-refinement model. *Annual Review of Developmental Psychology*, *1*, 197–222.

Millán, A. P., Torres, J. J., Johnson, S., & Marro, J. (2018). Concurrence of form and function in developing networks and its role in synaptic pruning. *Nature communications*, *9*(1), 1-10.

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

Reznicek, M., Walter, M., Schmidt, K., Lüdeling, A., Hirschmann, H., Krummes, C., & Andreas, T. (2010). Das Falko-Handbuch: Korpusaufbau und Annotationen. *Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin*.

Sakai, J. (2020). Core Concept: How synaptic pruning shapes neural wiring during development and, possibly, in disease. *Proceedings of the National Academy of Sciences*, *117*(28), 16096-16099.

Shadrova, Anna. (2021). Kobalt: Extension Corpus and Annotation Guidelines for Verb Classification and Dependency Adjustments (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5730224

Shadrova, A., Linscheid, P., Lukassek, J., Lüdeling, A., & Schneider, S. (2021). A Challenge for Contrastive L1/L2 Corpus Studies: Large Inter-and Intra-Individual Variation Across Morphological, but Not Global Syntactic Categories in Task-Based Corpus Data of a Homogeneous L1 German Group. *Frontiers in psychology*, *12*.

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, *36*(5), 549-569.

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

Szmrecsanyi, Benedikt (2005). "Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English". Corpus Linguistics and Linguistic Theory 1(1): 113-150.

Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., Jahns, E., Klotz, M., Krause,T., Labrenz, A., Lüdeling, A., Martynova, M., Neuhaus, K., Pashkova, T., Rizou,V., Tracy, R., Schroeder, C., Szucsich, L., Tsehaye, W., Zerbian, S., & Zuban, Y. (2019). *RUEG Corpus (Version 0.3.0) [Data set]*. Zenodo. http://doi.org/10.5281/zenodo.3236069

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press, 110 Midland Ave., Port Chester, NY 10573-4930 (45 British pounds).

## Using IPSyn to measure early L2 syntactic development

Masatoshi Sugiura[1], Akiko Eguchi[2], Mariko Abe[3], Remi Murao[4], Takashi Koizumi[5], Daisuke Abe[6]
Nagoya University[1,4,5], Nagoya Women's University[2], Chuo University[3], Chubu University[6]
{sugiura, eguchi.akiko, murao, koizumi}@nagoya-u.jp[1,2,4,5], abe.127@g.chuo-u.ac.jp[3], abe.gsid@gmail.com[6]

For the elucidation of second language acquisition (SLA), measurement of grammatical development is essential. In the field of SLA, complexity, accuracy, and fluency (CAF) measures are used (e.g., Ortega 2009). Some traditional measures use the mean length of linguistic units such as sentences (MLS), T-units (MLT), or clauses (MLC) (e.g., Lu 2010). The length-based measures can be used as indirect indexes of complexity or fluency, but they do not represent complexity or fluency themselves. Lately, however, the index of productive syntax (IPSyn) was invented in the field of L1 research. It was originally created by Scarborough (1990) as a form of checking sheet of typical grammatical constructions expressed by children of particular stages, based on Miller's (1981) Assigning Structural Stage and Lee's (1974) Developmental Sentence Score. This index is based not on length-based but on grammatical complexity features of spoken utterances. Now, the IPSyn scores can be automatically calculated using the tools based on the Child Language Data Exchange System (CHILDES) (MacWhinney 2000). Altenberg et al. (2018) revised the index, and the tools were modified based on Roberts et al.'s (2020) evaluation of the program. If IPSyn is applicable to the measurement of L2 development, it can facilitate SLA research using learner corpora.

The aim of the current study is to explore if IPSyn can measure early L2 syntactic development. It also examines whether there is any order of importance among the four subscales of syntactic development: the noun phrase, the verb phrase, questions/negations, and sentence structures. Unlike L1, learning L2 can start at a much later stage, especially in the case of English as a foreign language (EFL). The formal introduction of L2 English in Japan starts at around age ten. We collected a cross-sectional data set from junior high school students, aged 12 to 15, in 2020. A total of 223 students, 79 first year, 73 second year, and 69 third year performed a series of communicative tasks: five picture descriptions designed to produce interrogative sentences and two short narrative tasks. The proficiency scores of the TOEFL Primary Speaking Test indicated that they could be estimated as Basic Users (A1–A2) according to the CEFR benchmark. All the spoken data were transcribed manually in the CHILDES CHAT format and then automatically parsed by the CHILDES CLAN program. The corpus size is 39,109 words, after excluding fillers, repetitions, and self-corrections based on the CHILDES protocol. According to previous studies, IPSyn requires at least 50 utterances to measure the full syntactic ability of the speaker. However, due to the smaller number of utterances produced by our participants, the present study computed IPSyn scores from 25 utterances. Although this smaller sample size risks higher variability in the data, our data set comes from responses to the same set of prompts, and the smaller sample size should have less of an impact on the results compared to data coming from sessions of natural speech.

After automatically calculating IPSyn scores, we analyzed the development of IPSyn scores among the three-year groups of learners using Generalized Linear Models (GLM) with the IPSyn scores as the response variable and school year as the predictor variable. The result of the GLM showed that there were significant differences among the school years. The post-hoc multi comparisons showed that all differences among the years were significant. In addition, an ordinal logistic regression analysis revealed that verb phrases, sentence structures, and questions/negations were significant in that order. The noun phrase was not significant. Our investigation suggests that IPSyn can show L2 syntactic developmental differences and that the verb phrase is the most significant variable for syntactic development in our data.

Further research is necessary to re-examine the results of this study using longitudinal data. Additionally, the linguistic items referenced by IPSyn need further examination, since they are based on the language development of L1. It is worth considering features that more accurately represent the language development of L2, which would lead the discussion to the issue of L1/L2 parallelism. Another issue to discuss is the role of task types. Rather than the naturalistic context of L1 child data collection, L2 data are collected through tasks, which may affect the quality/quantity of the data.

**References**


Altenberg, E. P., Roberts, J. A., & Scarborough, H. S. (2018). Young children's structure production: A revision of the Index of Productive Syntax. *Language, speech, and hearing services in schools*, *49*(4), 995-1008.

Garbarino, J., Ratner, N. B., & MacWhinney, B. (2020). Use of computerized language analysis to assess child language. *Language, speech, and hearing services in schools*, *51*(2), 504-506.

Lee, L. (1974). *Developmental Sentence Analysis.* Evanston, IL: Northwestern University Press.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B., Roberts, J. A., Altenberg, E. P., & Hunter, M. (2020). Improving automatic IPSyn coding. *Language, Speech, and Hearing Services in Schools*, *51*(4), 1187-1189.

Miller, J. F. (1981). *Assessing Language Production in Children.* Baltimore, MD: University Park Press.

Ortega, L. (2009). Sequences and processes in language learning. In M. H. Long & C. J. Doughty (Eds.). *Handbook of Language Teaching*. Malden, MA: Wiley-Blackwell, 81-105.

Roberts, J. A., Altenberg, E. P., & Hunter, M. (2020). Machine-scored syntax: Comparison of the CLAN automatic scoring program to manual scoring. *Language, speech, and hearing services in schools*, *51*(2), 479-493.

Scarborough, H. S. (1990). Index of productive syntax. *Applied psycholinguistics*, *11*(1), 1-22. doi:10.1017/S0142716400008262

## Corpora as input and output: A fragile link in classroom research

Anita Thomas[1], France Rousset[2]
University of Fribourg
anita.thomas@unifr.ch[1], france.rousset@unifr.ch[2]

The integration of corpora of language use in L2 language teaching has gained in importance during the last decade. Corpora are appreciated especially for L2 writing development (Boulton & Cobb 2017). In the francophone world, several projects have produced corpus-based teaching material to support the development of L2 French spoken and interactional competence. However, most of them are within a qualitative and one-time design (e.g. André 2019; Etienne & David 2020).

In this contribution, we present an ongoing longitudinal project over two years in which we use corpora of French spoken language as teaching material during classroom interventions and in exercises available on an educational platform. The corpus-based materials focus on spoken and interactional features. This authentic material has been integrated into carefully designed pedagogical sequences.

The aim of the project is to examine the influence of this material on the L2 learners' development of interactional competence. For this purpose, we have developed a series of tools ranging from simple questionnaires to evaluation grids targeting specific linguistic phenomena in spoken interaction.

The project is conducted in ten classes of mixed L1 (n=±40) and L2 (n=±40; B1/B2) speakers of French attending vocational training for manual professions. The interventions take place during ordinary classes twice a term. After each intervention, the participants are asked to complete two exercises that are available on an educational platform. Each exercise targets a specific interactional marker in an interactive manner and is completed autonomously. At the end of each intervention, we ask the participants for feedback orally and in anonymous questionnaires. These feedbacks are generally positive, the learners appreciate this way of discovering French. However, this does not mean that they learn from this input.

In order to assess the development of L2 French, we are collecting a corpus of short free peer interactions at the very beginning of each intervention. Based on research about assessing L2 interaction, we have developed an evaluation grid including a range of criteria addressing turn organization, topic management, and the use of conversational markers (Salaberry & Kunitz 2019; Salaberry & Rue Burch 2021). The grid also includes key features of our corpus-based teaching material. While such criteria are crucially needed to assess interactional competence, the construction and application of an evaluation grid is a challenge given factors such as rater subjectivity.

We will present a case study with the results from eleven speakers of the corpus having Tigrinya as L1. The audio files have been transcribed with the EXMARaLDA software.

Preliminary results from the first recordings show that despite difficulties with pronunciation, grammar and vocabulary most of the interactions develop smoothly between the interlocutors. However, most of the topic shifts are rather abrupt and disagreements are ignored. The markers introduced in the input activities are rare in the learners' first recordings but appear in the following recordings.

While the feedbacks and the evaluation grid are useful to establish a link between input and L2 output, this relation remains fragile. The production of interactional features is strongly related to the topics chosen by the interlocutors. However, comments from the participants suggest that the taught features have become salient and might develop over time.

**References**

André, V. (2019). Pourquoi faire de la sociolinguistique des interactions verbales avec des enseignants et des apprenants de Français Langue Étrangère ? *Linx. Revue des linguistes de l'université Paris X Nanterre*, 79.

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning*, 67(2), 348-393.

Etienne, C., & David, C. (2020). L'enseignement du français avec les interactions: Approche méthodologique et mise en œuvre en classe depuis le niveau débutant. *SHS Web of Conferences*, 78 (07004). https://doi.org/10.1051/shsconf/20207807004

Salaberry, M. R., & Kunitz, S. (Eds.). (2019). *Teaching and Testing L2 Interactional Competence: Bridging Theory and Practice*. New York: Routledge.

Salaberry, M. R., & Rue Burch, A. (Eds.). (2021). *Assessing Speaking in Context. Expanding the Construct and its Applications*. Bristol : Multilingual Matters.

# Investigating connective use in L2 German: A corpus study

Helena Wedig[1], Carola Strobl[2], Jim Ureel[3]
University of Antwerp
helena.wedig@uantwerpen.be[1], carola.strobl@uantwerpen.be[2], jim.ureel@uantwerpen.be[3]

Writing a cohesive text, that is, the ability to connect sentences, paragraphs, and ideas via the use of a range of grammatical and lexical devices, must be taught explicitly in foreign language (L2) writing classes because the preferred devices used to express cohesion differ between languages (Kunz et al. 2017). L2 writers struggle with cohesion since they tend to rely on native language (L1) strategies to create cohesive texts. This has been shown in studies on L2 English (e.g., Appel & Szeib 2020; Hinkel 2001; Johnson 2017; Stemmer 1991). To date, the most comprehensive monolingual study of cohesion is Halliday and Hasan's (1976) *Cohesion in English*, which has served as a point of departure for other languages, including German. The authors list five categories of cohesion: (1) co-reference, (2) substitution, (3) ellipsis, (4) conjunction, and (5) lexical cohesion. While English cohesive devices have garnered considerable attention in second language acquisition (SLA) research (e.g., Das et al. 2017 (connectives); Tanskanen, 2006 (lexical cohesion)), less attention has been paid to the use of cohesive devices in German (e.g., Belz 2005; Strobl 2020; Walter 2007). Conjunction is the cohesive category that has received the most attention from scholars investigating L1 German. For example, Stede (2016) analysed connectives in the Potsdam Commentary Corpus, using a self-developed tool (ConAno) for semi-automated connective extraction and analysis. In addition, Walter (2016) investigated aspects of academic writing in the Korpus Akademisches Deutsch, including the distribution of subordinating conjunctions. A contrastive study, which also includes an intra-language comparison between genres and focuses on connectives, the overarching category of cohesive devices which includes conjunctions, performed by Kunz et al. (2021) revealed that German academic texts contain more temporal (e.g., *bevor*) and expansion (e.g., *anhand*) connectives compared with contingency (e.g., *aufgrund*) and comparison (e.g., *dagegen*) connectives. In stark contrast to the growing research interest in connectives in L1 German, research into cohesion in L2 German to date has been scarce, with a handful of available studies focusing on texts produced by writers with heterogeneous L1 backgrounds (e.g., Strobl 2020; Walter 2007). Given the impact of the L1 on L2 learners' cohesion-building patterns, there is a dire need to investigate cohesion in texts produced by L2 German writers with a homogeneous L1 background.

The present study aims to close this research gap, by investigating cohesion in L2 German texts written by learners with L1 Dutch. The first category we will focus on is the conjunction, since it is the category that has received the most attention in research on cohesion in German to date. This will allow us to compare our results with previous studies, shedding light on L1-specific aspects of conjunction in learner writing. This analysis will be based on the Belgisches Deutschkorpus (Beldeko) (Strobl 2020), which has recently been built for this specific purpose. Beldeko consists of 301 summaries (70774 tokens) written by advanced learners of L2 German in an academic writing course. The corpus has been pre-processed and automatically annotated with PoS-tags and lemmata. Furthermore, connectives have been pre-annotated automatically according to guidelines based on PDTB3 (Webber et al. 2019) in combination with DimLex, a database containing German connectives and their corresponding PDTB3 tags (Scheffler & Stede 2016; Stede 2002).

A preliminary descriptive analysis of the automatically pre-annotated data via R shows higher use of temporal and expansion connectives compared with contingency and comparison connectives. This ties in with Kunz et al.'s (2021) results for L1 German. In another study, Konjevod (2012) stated that L2 German learners do not use concessive, conditional, and disjunctive connectives and, furthermore, restrict their use of additive connectives to *und*. In addition, Walter and Schmidt (2008) concluded that *und* is mostly used in sentence-initial position by learners. Concerning L2 English research, Martinez (2002) showed that L2 English learners use only a restricted set of connectives, disregarding others. The preliminary result, as well as the hypotheses from earlier research on connectives in learner language, will be further investigated with manual annotation, using the online annotation platform Inception (Klie et al. 2018). In conclusion, we will analyse whether (1) L2 German learners use connectives from all categories, (2) they use a restricted set of connectives per semantic category, (3) they restrict specific connectives to certain positions in sentences, and (4) they use more temporal and expansion connectives than contingency and comparison connectives.

**References**


Appel, R., & Szeib, A. (2018). Linking adverbials in L2 English academic writing: L1-related differences. *System*, *78*, 115–129. https://doi.org/10.1016/j.system.2018.08.008

Belz, J. A. (2005). Corpus-driven characterizations of pronominal da-compound use by learners and native speakers of German. *Die Unterrichtspraxis/Teaching German*, *38*(1), 44–60. https://doi.org/10.1111/j.1756-1221.2005.tb00041.x

Das, D., Scheffler, T., Bourgonje, P., & Stede, M. (2018). Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 360–365). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5042

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning*, *12*(2), 111–132.

Johnson, M. (2017). Improving cohesion in L2 writing: A three-strand approach to building lexical cohesion. *English Teaching Forum*, *55*, 2–13.

Klie, J. C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrationsm* (pp. 5–9). Association for Computational Linguistics.

Konjevod, A. (2012). Connectives in student writing: A learner corpus study. *Strani Jezici*, *41*, 1.

Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K., & Steiner, E. (2017). English–German contrasts in cohesion and implications for translation. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (pp. 265–312). De Gruyter Mouton. https://doi.org/10.1515/9783110459586-010

Kunz, K., Lapshinova-Koltunski, E., Martínez Martínez, J., Menzel, K., & Steiner, E. (2021). *GECCo - German–English Contrasts in Cohesion: Insights from corpus-based studies of languages, registers and modes*. De Gruyter Mouton. https://doi.org/10.1515/9783110711073

Martínez, A. C. L. (2002). The use of discourse markers in EFL learners' writing. *Revista alicantina de estudios ingleses*, *15*(4), 123–132. https://doi.org/10.14198/raei.2002.15.08

Scheffler, T., & Stede, M. (2016). Adding semantic relations to a large-coverage connective lexicon of German. In *Proceedings of LREC* (pp. 1008–1013). *ELRA.*

Stede, M. (2002). DiMLex: A lexical approach to discourse markers. In A. Lenci & V. Di Tomaso (Eds.), *Exploring the lexicon: Theory and computation* (pp. 1–15). Edizioni dell'Orso.

Stede, M. (2016). Konnektoren und Argumente. In M. Stede (Ed.), *Handbuch Textannotation: Postdamer Kommentarkorpus 2.0* (pp. 111–131). Universitätsverlag Potsdam.

Stemmer, B. (1991). *Kohäsion im gesprochenen Diskurs deutscher Lerner des Englischen*. J. Groos.

Strobl, C. (2020). Darum sind Pronominaladverbien eine Herausforderung für Deutschlerner: Eine korpusbasierte kontrastive Interimssprachenanalyse hierzu. *Germanistische Mitteilungen*, *45* (1), 89–111.

Tanskanen, S. K. (2006). *Collaborating towards coherence: Lexical cohesion in English discourse*. John Benjamins. https://doi.org/10.1075/pbns.146

Walter, M. (2007). Hier wird die Wahl schwer, aber entscheidend: Konnektorenkontraste im Deutschen. In H.-J. Krumm (Ed.), *Theorie und Praxis - Österreichische Beiträge zu Deutsch als Fremdsprache* (pp. 145–161). StudienVerlag.

Walter, M., & Schmidt, K. (2008): "Und das ist auch gut so": Der Gebrauch des satzinitialen und bei fortgeschrittenen Lernern des Deutschen als Fremdsprache. In B. Ahrenholz, U. Bredel, W. Klein, M. Rost-Roth, & R. Skiba (Eds.), *Empirische Forschung und Theoriebildung. Beiträge aus Soziolinguistik, Gesprochene-Sprache- und Zweitspracherwerbsforschung: Festschrift für Norbert Dittmar zum 65. Geburtstag* (pp. 331–342). Peter Lang.

Walter, M. (2016). In der Kürze liegt die Würze: Lexikalisch-grammatische Strukturen im akademischen Schreiben. In H. Schweiger, V. Ahamer, C. Tonsern, T. Welke, & N. Zuzok (Eds.), *In die Welt hinaus: Festschrift für Renate Faistauer* (pp. 201– 217). Praesens.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). The penn discourse treebank 3.0 annotation manual. *University of Pennsylvania*.

# An exploratory corpus-based study of Arab learners' usage of English phrasal verbs

Sadeem Ibn Alameer[1], Dagmar Divjak[2], Paul Thompson[3]
University of Birmingham
sxi939@student.bham.ac.uk[1], d.divjak@bham.ac.uk[2], p.thompson@bham.acuk[3]

Phrasal verbs are widely recognized as among the most confusing and complex forms for learners of English as a foreign language. They are notoriously difficult for Arab EFL learners to master as Arabic is a Semitic language which has no designated category of phrasal verbs. To determine Arab EFL learners' weaknesses in using phrasal verbs and to draw Arab EFL instructors' attention to those weaknesses, an exploratory corpus-based study was conducted. The main aim of this study is to understand the actual use of phrasal verbs by Arab EFL learners, to explore how Arab learners of English use phrasal verbs, and to identify the types of mistakes learners make when using them. This poster will present the results of this analysis. For this exploratory corpus-based study, the ten most and ten less frequently used phrasal verbs of British and American English listed in Liu (2011) were selected for investigation. The EF-Cambridge Open Language Database (EFCAMDAT) was used as a resource of data produced by pre-intermediate Arab EFL learners. The data of 1640 Arab learners from 17 different nationalities at levels seven to nine according to EFCAMDAT were selected. To retrieve the 20 phrasal verbs selected for investigation, the Query Pattern function in the EFCAMDAT was used by formulating two strings to search for different syntactic structures of phrasal verbs. The inseparable phrasal verbs where the object must follow the particle, or 'construction$_0$' of the separable phrasal verbs, in which the verb is followed by the particle and the other string was used to search for 'construction$_1$' of the separable phrasal verbs, in which the object is placed between the verb and the particle. The outcomes of the used strings were saved as Excel sheets for annotation and further investigation. First of all, all the retrieved concordance lines were visually inspected to identify whether the combinations were phrasal verbs or other free combinations by following the criteria for identification listed in (Biber et al. 1999) and (Quirk et al. 1985). Second, if the concordance lines involved phrasal verbs, the learners' use of such PVs was examined from two different angles, namely semantics, and syntax. Third, to confirm the identification of errors, all the data were re-examined three weeks after the initial examination, and the concordance lines were checked by two native speakers of English. The results of the descriptive statistics for the pre-intermediate Arab learners' use of the 20 English phrasal verbs revealed that learners used the ten most frequent phrasal verbs more frequently and more correctly than the ten less frequent phrasal verbs. This result could be linked to the frequent nature of these high frequently used phrasal verbs by native speakers of English which may influence learners' usage. Moreover, learners' preferred choice when using the separable phrasal verbs is to use construction$_0$ in which the verb was followed by the particle instead of construction$_1$ where the object is placed between the verb and the particle. Pre-intermediate Arab EFL learners may refrain from using construction$_1$ due to the complexity of the structure which may lead to different errors in production. The study's results also revealed that learners did not make any mistakes with object placement in phrasal verbs, but they did misuse some phrasal verbs in terms of other semantic and syntactic aspects. The main semantic error was the learners' use of inappropriate phrasal verbs given the context, either by choosing an incorrect phrasal verb to deliver the intended meaning or selecting a wrong particle as in *put off the fire*. In addition, learners used phrasal verbs in contexts in which single verbs were more appropriate. In terms of syntactic errors, the learners made many grammatical errors, such as the lack of tense consistency with the phrasal verbs and incorrect subject-verb agreement. This finding may indicate that Arab EFL learners treated phrasal verbs as a vocabulary aspect of the language regardless of their need to apply grammatical rules to produce complete grammatical sentences.

**References**

Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *Tesol Quarterly, 45*(4), 661–688.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English Language*. London: Longman.

## Using ICALL to collect spoken learner data in real-life conversation tasks

Elizabeth Bear[1], Bronson Hui[2], Haemanth Santhi Ponnusamy[3],
Björn Rudzewitz[4], Xiaobin Chen[5], Detmar Meurers[6]
University of Tübingen
{elizabeth.bear[1], bronson.hui[2], haemanth.santhi-ponnusamy[3], bjoern.rudzewitz[4], xiaobin.chen[5],
detmar.meurers[6]}@uni-tuebingen.de

There have been calls for more attention to learners' spoken production in corpus research because oral communication is as important as its written counterpart (e.g., Yoon 2020). However, spoken learner corpora are much less common, partly because speech samples are more difficult to collect and process. Commonly used spoken corpora often rely on recordings made during the learner's production in various experimental or assessment contexts. This approach is not only time-consuming and costly (e.g., Andersen 2010; Love et al. 2017), but also has implications for sampling, thus affecting the generalization of the research findings based on these corpora. For example, students from a certain socio-economic background (e.g., from richer countries or families) may be more likely to participate in research or take a standardized test. Hence, corpora built upon these situations may or may not reflect the language produced by other second language learner populations. Such bias can have a negative impact on the field's ability to empirically explore the language acquisition process. In this light, we argue that intelligent computer-assisted language learning (ICALL) systems lend themselves as a valuable tool for data collection due to their potential implementation scale (Meurers et al. 2019; Alexopoulou et al. 2022), natural language processing capability (Meurers 2020), and system logs providing longitudinal records of the learner's interaction with the system (Hui et al. in press).

In this talk, we will first present an ongoing project which aims at developing an ICALL system for training spoken English in real-life contexts. The system is capable of conversing with users in natural language to help them learn how to realize real-life tasks (e.g., booking a table at a restaurant, comparing universities, etc.) with the target language. The system follows Task-Based Language Teaching (TBLT) design principles and integrates the latest technologies in conversation agents, natural language and speech processing, as well as mobile platforms. Besides allowing for speech interaction between the learner and the system, the ICALL system under development will also provide corrective and scaffolding feedback. Currently, the system framework has been set up and a number of tasks have been implemented, so we will be able to demonstrate how the system works at the LCR conference.

In the second part of the talk, we will present the unique opportunities our ICALL system offers in terms of the types of data and corpus it can help collect, including, for example, speech data under natural communication tasks, student reaction to different types of feedback, longitudinal changes/development of learner production and so on. This extends previous work (Strik et al. 2012; Cucchiarini et al. 2014) which has highlighted the different types of resources that are made possible by language learning applications targeting spoken language. The corpus and interaction data collection framework and technical details of the ICALL system will also be presented. We believe that these data will help inform Second Language Acquisition (SLA) research and practical system development. Although the actual data collection is planned at a later stage of the project, we believe that the LCR community would be interested in seeing what state-of-the-art technology has to offer in terms of spoken corpus construction and the unique opportunities ICALL systems can offer by integrating spoken corpora with interaction logging data.

In sum, we highlight the need for large-scale spoken corpora collection with naturalistic real-life language use tasks. We exemplify how an ICALL system under construction will offer the possibility to collect such corpora and corresponding data on spoken interaction. We propose how such corpora and data can help corpus linguists answer SLA and language learning questions.

**References**

Alexopoulou, T., Meurers, D., & Murakami, A. (2021). Big Data in SLA: advances in methodology and analysis. In N. Ziegler & M. González-Lloret (Eds.). *Routledge Handbook of Second Language Acquisition and Technology*. https://doi.org/10.4324/9781351117586-9

Andersen, G. (2010). How to use corpus linguistics in sociolinguistics. In A. O'Keeffe & M. McCarthy (Eds.). *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 547-562

Cucchiarini, C., Bodnar, S., de Vries, B. P., van Hout, R., & Strik, H. (2014). ASR-based CALL systems and learner speech data: New resources and opportunities for research and development in second language learning. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* 2708-2714.

Hui, B., Rudzewitz, B., & Meurers, D. (in press). Learning Processes in Interactive CALL Systems: Linking Automatic Feedback, System Logs, and Learning Outcomes. *Language Learning and Technology*. Preprint available at https://osf.io/gzs9r/

Love, R. Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319-344.

Meurers, D. (2020). Natural Language Processing and Language Learning. In C. A. Chapelle (Ed.). *The Concise Encyclopedia of Applied Linguistics*. Wiley, 817-831.

Meurers, D., De Kuthy, K., Nuxoll, F., Rudzewitz, B., & Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, 39, 161-188. https://doi.org/10.1017/S0267190519000126

Strik, H., Colpaert, J., van Doremalen, J., & Cucchiarini, C. (2012). The DISCO ASR-based CALL system: Practicing L2 oral skills and beyond. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2702-2707.

Yoon, S. (2020). The learner corpora of spoken English: What has been done and what should be done? *Language Research*, 56(1), 29-51.

## Sarramanka: An online tool for learner corpora analysis

El Ayari Sarra
Université Paris Lumières, CNRS, Université Paris 8
sarra.elayari@cnrs.fr

We present an online tool developed to facilitate corpus-based research in Second Language Acquisition (SLA). It aims to provide solutions for researchers working on learner corpora, which are difficult to analyze automatically with Natural Language Processing methods. The approach we propose is to provide ready-to-use functionalities correlated to a user-friendly platform for qualitative analysis purposes as well as to offer some quantitative data based on the manually annotated corpus. It is developed in collaboration with researchers and intends to be a tool as accessible as possible. Researchers do not need to take care of the installation of the tool, updates, settings, or formats of data and annotations. The platform can be used free of charge, and no other requirements are needed from the users.

This platform is part of a larger research project named Sarramanka which has two main goals. First, to develop tools that are easy to use and dedicated to the needs of researchers who wish to collaborate on the project by exploring and analyzing corpora with it. Secondly, to provide more accessibility and interoperability to facilitate the exchanges between tools and projects corpora as a part of Open Science. Researches in SLA have produced (and still do) a multitude of different types of learner corpora (Granger 2004) which are transcribed (for oral corpora), annotated, and analyzed. Many of these corpora are accessible today in different databases (Ortolang, Cocoon, Talkbank for example) and ready to be used. Different levels of availability can be provided for the corpora depending on their licences.

Different functionalities have been created in our tool so far, following the process of linguistic analysis once the corpus has been collected. It has been designed for written corpora as well as for spoken ones. The transcription functionality allows to listen to an audio file and to transcribe the speech. It is possible to segment the transcription freely, by utterances for example. The transcription can always be modified at any stage of the work. The annotation functionality allows annotating the corpus once transcribed in the case of spoken corpora. At this stage, it is possible to create an annotation schema that will be presented to the user as a web form - where the users can indicate all linguistic information they want to fill in during the analysis.

During this annotation process, the users can navigate through all the files of the corpus and annotate the phenomena that interest them. A string to be annotated - which can be composed of letter(s), word(s), or utterance(s) - is highlighted with the mouse, which brings up the web form containing the elements corresponding to the annotation schema. Once the elements have been filled in, the user can continue to annotate the rest of the corpus. The annotated elements are then displayed in different colors so that their repartition within the corpus is accessible visually. The users can also add comments on the go during the transcription phase as well as during the annotation phase. These comments can help the annotation schema to evolve, a dynamic process that can be refined during the data analysis. A search engine allows searching within the corpus for words as well as annotations or comments. This function allows users to create sub-corpora based on specific phenomena.

All the annotations are available for quantitative purposes, and statistics and graphs can be generated automatically. The annotated phenomena can be quantified according to filters based on the annotation schema as well as on the corpus metadata (information about the learners for example). Moreover, it is possible to export the annotated corpus (or sub-corpus) in different interoperable formats (CSV, XML, CHAT for CLAN) in order to use it with other tools.

The diversity of projects that have used and will use the platform allows us to continue its development and to consider more levels of annotations, such as phonetic transcriptions for example. We think that it could also be very interesting to incorporate more multi-modality and to add video annotation with image alignment as well as an export compatible with the ELAN software. The platform has already been used for two researches in SLA (Watorek et al. 2021; El Ayari & Watorek 2021) and is currently used for research on speech-language pathology. We aim at creating more customized features for researchers interested to contribute to this collaborative project.

**References**

El Ayari, S. & Watorek, M. (2021). Exploration outillée pour un corpus de productions orales des Apprenants débutants en L2. In *Colloque Influence translinguistique: où en est-on aujourd'hui ?* Toulouse, France.

Granger, S. (2004). Computer learner corpus research: current status and future prospects. In T. Connor U. & Upton (Ed.), *Applied corpus linguistics: a multidimensional perspective* (pp.123-145). Amsterdam/Atlanta: Rodopi.

Watorek M., Trévisiol P. & Rast R. (2021). The Emergence of Determiners in French L2 from the Point of View of L1/L2 Comparison. *Languages*. 6(2):73.

# Mexican learner corpus: Designing and collecting a longitudinal spoken corpus of Mexican university learners of English

Ana Abigahil Flores Hernández[1], Pauline Moore[2]
Universidad Autónoma del Estado de México
aby.flores.hernandez@gmail.com[1], paulinelenguas@gmail.com[2]

In the Mexican context, the LCR field is still characterized by few studies which make use of small collections, cross-sectional, and written data mainly focus on higher levels of proficiency (Flores 2019, García 2012, Fuentes 2012). This situation points to the need to develop a national learner database that can be used in the development of learner-centred tools and materials for ELT and, at the same time, a collection that can be used as an empirical base and complementary methodology in Mexican SLA research (Meunier 2021, Guilquin 2015). The present study reports the process of designing and collecting a spoken longitudinal corpus of Mexican university learners as well as a brief description of the data obtained during the first year of work as part of a Postdoctoral research project.

The design criteria were selected with the aim of collecting a large developmental and representative sample of the spoken interlanguage of university students learning English as part of their bachelor's degree programmes. To elicit spoken production, several tasks have been selected to obtain different degrees of "naturalness". The interview design focuses on the use of spontaneous monologic and non-interactive productions to capture learner interlanguage in extended turns. It is also hoped that the task design will elicit a wide range of text types including information-centred, stance-focus, and narrative texts (Biber 2004).

The tasks are applied through a multi-level from ten to 16 minutes semi-guided interview which is intended to be applied every year, following the participants in their English acquisition process during four to five years of their time in the university, resulting in a developmental collection of data (Meunier 2021, Guilquin & Meunier 2015, Callies & Paquot 2015). The parameters for the four tasks designed for the MexLeC interview were the descriptors for three "sustained-monologue productive spoken activities" as included in the communicative activities section of the CEFR general descriptive scheme (Council of Europe 2018), an analysis of tasks and materials used in currently available spoken corpora from the "Learner corpora around the world*"* list (Université Catholique de Louvain 2020), and Biber´s (2004) multi-dimensional analysis of conversation text-types.

The project is one-year-old and current participants are university students from the Bachelors in Languages of the *Universidad Autónoma del Estado de México (UAEMex)* in their second, fourth and sixth semesters. To determine language proficiency, we are using study time, class materials, and the results from internal mock examinations, which allow us to identify levels from A1 to C1. Additionally, a learner profile is applied concurrently with the interview to obtain information on student L1, background, and exposure to the target language and other foreign languages. Interviewers are Mexican English teachers with no relation to the interviewees holding at least a B2 level of proficiency. Interviews were video recorded using the Zoom video meeting app and the transcription guidelines have been adapted from Gablasova et al. (2019) and Granger et al. (2009). Currently, the size of the collection after the first year of work is 120,000 tokens. Some of the most interesting preliminary findings on the resulting data are the (expected) low scores of the type/token ratio; the wide use of fillers and pauses followed by elaborated chunks, and the dissimilar features produced in the narrative task of the ones expected to be distinctive of this text-type.

Currently, a second interview round is in progress including an estimated 350 participants in their second, fourth and sixth semester at *UAEMex* and *Universidad Autónoma del Estado de Hidalgo (UAEH),* for whom the learner profile and learner materials data is collected using the same methodology to that employed in the *UAEMex*. The collection of this corpus makes an important contribution not only to the Mexican research network on ELT and SLA but also contributes to the general expansion of the LCR field by the inclusion of some of the most underrepresented variables in the available learner corpora, longitudinal, spoken and beginner learners' productions (Guilquin 2015, Granger 2008). This collection is available at Flores and Moore (2021) and can be freely downloaded for research purposes.

**References**


Biber, D. (2004). Conversation text types: A multi-dimensional analysis. *7es Journées internationales d'Analyse statistique des Données Textuelles*

Callies, M., & Paquot M. (2015). Learner corpus research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research*, 1(1), 1–6.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, Teaching & Assessment. Companion Volume with New Descriptors*. Strasbourg, Language Policy Division: CUP.

Flores, A. (2019) *Adquisición de sufijos derivativos en inglés como L2*. Upublished doctoral thesis. Mexico: Universidad Autónoma del Estado de México.

Flores, A., & Moore, P. (2022) *Mexican Learner Corpus.* Viewed May 1st. 2018, https://sites.google.com/view/mexlec/intro.

Fuentes, A. (2012) *Análisis de la estrucctura genérica propia de ensayos argumentativos académicos en inglés producidos por aprendices mexicanos de inglés como lengua extranjera*. Unpublished master thesis. México: Universidad Autónoma del Estado de México.

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: development, description and application. *International Journal of Learner Corpus Research*, 5(2), 126-158.

García, E. (2012) *Análisis del contenido argumentativo en un corpus de ensayos en inglés.* Unpublished master thesis. Mexico: Universidad Autónoma del Estado de México.

Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (Eds.). *Corpus Linguistics. An International Handbook. Volume 1*. Berlin: Walter de Gruyter, 259-275.

Granger, S., Gilquin, G. & Meunier, F., (2015) *The Cambridge Handbook of learner corpus research*. Cambridge: Cambridge University Press.

Guilquin, G. (2015) From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.). *The Cambridge Handbook of learner corpus research*. Cambridge: CUP.

Meunier, F. (2021) Introduction to learner corpus research. In N. Tracy-Ventura & M. Paquot (Eds.). *The Routledge Handbook of Language Acquisition and Corpora*. London: Routledge.

# Lexical complexity across proficiency levels in L2 Italian: Some preliminary findings

Luciana Forti[1], Irene Fioravanti[2], Fabio Zanda[3]
Università per Stranieri di Perugia
{luciana.forti[1], irene.fioravanti[2], fabio.zanda[3]}@unistrapg.it

In this poster, we present the preliminary findings of a single-word lexical analysis based on the CELI Corpus (Spina et al. under review). The CELI corpus is a 600,000-word learner corpus of Italian, containing written texts produced by over 3,000 learners, and balanced with respect to the four CEFR levels B1, B2, C1, and C2. The texts were produced under examination conditions in the context of the CELI (Certificati di Lingua Italiana) language certification exams, which are developed and administered around the world by the CVCL (Centro per la Valutazione e le Certificazioni Linguistiche), based at the University for Foreigners of Perugia. Only the texts produced by the candidates who passed the certification exam were included in the corpus.

Vocabulary knowledge has been shown to be a key component in language competence development and a reliable predictor of overall language proficiency (Milton 2013; Kim et al. 2018). Lexical complexity, in particular, is one of the most popular constructs used to analyse vocabulary knowledge in the corpus-based analyses of language produced by learners of a second/foreign language. A number of studies on L2 English and other L2s (e.g. Lu 2012; Treffers-Daller 2013; Zhang & Daller 2019), in fact, have been dedicated to analysing lexical complexity and the different facets of its multidimensional construct (Bulté & Housen 2012), by using measures of lexical richness. In most studies, the three main dimensions of lexical richness taken into consideration are: lexical diversity, which is the number of different words in a text; lexical sophistication, that is the proportion of difficult words in a text; and lexical density, which is the ratio between content words and total words in a text. As for Italian, despite the scarcity of automatised tools to calculate lexical richness – with the notable exception of the LOPP (Bardel & Lindqvist 2011) and its further elaborations – some studies have been carried out both on spoken (Lindqvist et al. 2013; Gallina 2015) and written learner texts (Vedder & Benigno 2016; Zanda 2019). Nevertheless, corpus-based analyses examining the development of lexical complexity, with reference to proficiency level progression, are still under-represented, especially when it comes to L2 Italian (Corino & Onesti 2017; Giacalone Ramat 2003). This is largely due to the limited availability of pseudo-longitudinal learner corpora of Italian, balanced at the level of proficiency.

Our study seeks to fill this gap by addressing the following research questions:

1) How can lexis in learner Italian written texts be characterised with reference to specific indices of lexical complexity (diversity, sophistication, density), at different levels of proficiency?
2) How well do the different lexical complexity indices predict proficiency level?
3) To what extent do the specific indices of lexical complexity differ throughout the four proficiency levels considered?

On the basis of the research questions formulated above, we will investigate whether any non-linear patterns are observable along the cline of proficiency levels, as complexity theory applied to second language acquisition suggests (Larsen-Freeman & Cameron 2009; Ortega & Han 2017). Furthermore, we will explore the degree of predictability exhibited by both the individual and the entire set of lexical complexity indices considered with respect to proficiency level. Finally, we will examine the degree of variation among the different proficiency levels, with respect to the lexical indices considered.

The preliminary findings of our analysis are presented in the context of their possible implications for second language acquisition theory, language testing and assessment, and second language pedagogy. Further directions for research are also identified, with specific reference to the possible contribution of this study to the analysis of the relationship between single-word and multiword units in learner texts, across different proficiency levels.

**References**

Bardel, C. & Lindqvist, C. (2011). Developing a lexical profiler for spoken French L2 and Italian L2: The role of frequency, thematic vocabulary and cognates. In L. Roberts, G. Pallotti & C. Bettoni (Eds.). *EUROSLA Yearbook 11*. Amsterdam-Philadelphia: John Benjamins, 75-93.

Bulté, B. & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken, I. & Vedder (Eds.). *Dimensions of L2 performance and proficiency, complexity, accuracy and fluency in SLA*. Amsterdam-Philadelphia: John Benjamins, 21-46.

Corino, E. & Onesti, C. (2017). *Italiano di apprendenti: studi a partire da VALICO e VINCA*. Perugia: Guerra.
Gallina, F. (2015). *Le parole degli stranieri. Il Lessico Italiano Parlato da Stranieri*. Perugia: Guerra.
Giacalone Ramat, A. (Ed.) (2003). *Verso l'italiano. Percorsi e strategie di acquisizione*. Roma: Carocci.
Kim, M., Crossley, S., & Kyle, K. (2018), Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, 102(1), 120-141.
Larsen-Freeman, D. & Cameron, L. (2009). *Complex systems and applied linguistics*. Oxford: Oxford University Press.
Lindqvist, C., Gudmundson, A., Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production". In C. Bardel, C. Lindqvist, B. Laufer (Eds.). *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*. Amsterdam: Eurosla Monographs Series, 109-126.
Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, 96(2), 190-208.
Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis*. Amsterdam: Eurosla Monographs Series, 57-78.
Ortega, L. & Han, Z. (Eds.) (2017). *Complexity theory and language development: in celebration of Diane Larsen-Freeman*. Amsterdam-Philadelphia: Benjamins.
Spina S., Fioravanti, I., Forti, L., Santucci, V., Scerra, A., Zanda, F. (under review). Il corpus CELI: una nuova risorsa per lo studio dell'italiano L2.
Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: human ratings and automated measures*. Amsterdam-Philadelphia: Benjamins, 79-104.
Vedder, I. & Benigno, V. (2016). Lexical richness and collocational competence in second-language writing. *International Review of Applied Linguistics in Language Teaching*, 54, 23-42.
Zanda, F. (2019). *CELIC beta: proposta di un learner corpus di composizioni scritte di candidati dei livelli avanzati degli esami CELI*. MA thesis, University for Foreigners of Perugia.
Zhang, J. & Daller, M. (2020). Lexical richness of Chinese candidates in the graded oral English examinations. *Applied Linguistics Review*, 11(3), 511-533.

# Is planning time beneficial for L2 production? A corpus-based study of anaphora resolution in L1 Spanish – L2 English learners

Elena García-Guerrero[1], Cristóbal Lozano[2]
University of Granada
egarciaguerrero@ugr.es[1], cristoballozano@ugr.es[2]

It has been argued that having planning time (PT) before conducting a linguistic task in a second language (L2) may improve L2 performance. Extensive research has investigated this in relation to vocabulary, which generally reports a positive effect (Abdi Tabari 2020; Li et al. 2015), and grammar, where some studies evidence a positive effect, while others show a negative effect or, at best, lack of it (Ahangari & Abdi 2011; Asgarikia 2014; Kabiri 2015; Rostamian et al. 2018). The influence of PT has been overlooked at the syntax-discourse interface despite it being an area of persistent problems for L2 learners (Sorace 2011). This investigation aims at filling this gap by examining whether PT has an effect on the syntax-discourse interface with a focus on anaphora resolution (AR), i.e., how different referring expressions (REs) such as null pronouns, overt pronouns, and noun phrases (NPs) corefer with their antecedent in prior discourse. For instance, in (1) produced by an L1 Spanish-L2 English learner, the overt pronoun *he* refers to *Chaplin* and *she* to *the woman*.

(1) A mother$_i$ with a baby$_j$ pass by Chaplin$_k$ and he$_k$ tries to leave her$_i$ the baby$_l$ but she$_i$ refuses [ES_WR_B2_40_17_14_EDL][1]

The literature reports deficits in learners' use of REs in discourse. Corpus-based studies show evidence of overexplicitation/redundancy in contexts where there is a continuation of the topic (topic continuity), i.e., learners tend to produce fuller forms than pragmatically required (Kang 2004; Leclercq & Lenart 2013; Lozano 2009, 2016; Martín-Villena & Lozano 2020; Quesada & Lozano 2020; Ryan 2015). Assuming the benefits of PT in preceding studies (Kabiri 2015; Mehnert 1998; Tavakoli & Skehan 2005), we predict a positive effect on AR: PT is expected to help learners reduce their cognitive load and construct more coherent discourse via pragmatically adequate REs.

Our general RQ is: *Does PT benefit learners' felicitous use of REs and help them be less overexplicit?* For this, we collected a small set of controlled data for the COREFL corpus (http://corefl.learnercorpora.com/), including two narratives per participants to explore the differences between unplanned vs. planned discourse.

Participants were intermediate L1 Spanish-L2 English learners (N=46) and native English speakers (N=18). Both groups were further divided into two subgroups: (1) planning vs. (2) no-planning conditions. The materials were two written film-retelling tasks: participants did a first task (Task 1) without PT and after a 10-minute pause, they completed a second task (Task 2). The difference was that the planning groups did Task 2 after 10-minute PT and the no-planning groups did it spontaneously. Natives' data were collected for comparative purposes, but only learners' results will be reported due to time limitations. Below, there is a visual representation of the design of the study.

Learners/Natives (no-planning condition): Task 1 (unplanned) à 10-minute pause à Task 2 (unplanned)
Learners/Natives (planning condition): Task 1 (unplanned) à 10-minute PT à Task 2 (planned)

Although AR is constrained by multiple factors (Lozano 2016), this investigation analysed only the information status of the anaphor (topic continuity vs. topic shift). A tagset was designed and implemented in UAM Corpus Tool (O'Donnell 2009), and all 3rd person singular subject REs (N=2,522) were manually annotated and compared via chi-squared analyses.

With this design, we formulated two specific RQs:

**RQ1:** Is there a planning effect on the overall REs produced by no-planning vs. planning learners?

**RQ2:** Is there a planning effect on the distribution of REs by no-planning vs. planning learners when information status is considered?

To answer RQ1, no-planning and planning learners' narratives in Task 2 were compared to test for the effect of planning. Surprisingly, results evidenced a lack of PT effect as the distribution of REs was similar in planners and non-planners: mostly NPs, followed by overt and null pronouns. However, for RQ2, a significantly positive effect of PT was found only in a specific scenario (topic continuity) when comparing the planning-condition learners'

[1] The code corresponds to the ones assigned in the COREFL corpus

Task 1 (done without PT) vs. their Task 2 (with PT): when PT was allowed, learners produced less redundant REs.

Overall, although PT is not beneficial for L2 AR across the board, it seems to be beneficial for a specific scenario (topic continuity), which happens to be the most problematic scenario for learners according to the literature.

**References**


Abdi Tabari, M. (2020). Differential Effects of Strategic Planning and Task Structure on L2 Writing Outcomes. *Reading & Writing Quarterly*, 36(4), 320-338.

Ahangari, S., & Abdi, M. (2011). The effect of pre-task planning on the accuracy and complexity of Iranian EFL learners' oral performance. *Procedia - Social and Behavioral Sciences*, 29, 1950-1959.

Asgarikia, P. (2014). The effects of task type, strategic planning and no planning on written performance of Iranian intermediate EFL learners. *Procedia - Social and Behavioral Sciences*, 98, 276-285.

Kabiri, M. (2015). Guided task-based planning and writing accuracy: The case of Iranian lower-intermediate EFL learners. *Theory and Practice in Language Studies*, 5(3), 518.

Kang, J. Y. (2004). Telling a coherent story in a foreign language: Analysis of Korean EFL learners' referential strategies in oral narrative discourse. *Journal of Pragmatics*, 36(11), 1975-1990.

Leclercq, P., & Lenart, E. (2013). Discourse cohesion and accessibility of referents in oral narratives: A comparison of L1 and L2 acquisition of French and English. *Discours*, 12.

Li, L., Chen, J., & Sun, L. (2015). The Effects of Different Lengths of Pretask Planning Time on L2 Learners' Oral Test Performance. *TESOL Quarterly*, 49(1), 38-66.

Lozano, C. (2009). Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus. In N. Snape, Y. I. Leung & M. Sharwood Smith (Eds.). *Representational Deficits in SLA: Studies in honor of Roger Hawkins*. John Benjamins Publishing Company: 127-166.

Lozano, C. (2016). Pragmatic principles in anaphora resolution at the syntax-discourse interface: Advanced English learners of Spanish in the CEDEL2 corpus. In M. Alonso-Ramos (Ed.). *Spanish Learner Corpus Research: Current trends and future perspectives*. John Benjamins Publishing Company: 235-265.

Martín-Villena, F., & Lozano, C. (2020). Anaphora resolution in topic continuity: Evidence from L1 English–L2 Spanish data in the CEDEL2 corpus. In J. Ryan & P. Crosthwaite (Eds.). *Referring in a Second Language: Studies on Reference to Person in a Multilingual World*. Routledge: 119-141.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20(1), 83-108.

O'Donnell, M. (2009). The UAM Corpus Tool: Software for corpus annotation and exporation. In *Applied Linguistics Now: Understanding Language and Mind/La lingüística aplicada actual: Comprendiendo el lenguaje y la mente*. Universidad de Almería: 1433-1447.

Quesada, T., & Lozano, C. (2020). Which factors determine the choice of referential expressions in L2 English discourse?: New evidence from the COREFL corpus. *Studies in Second Language Acquisition*, 42(5), 959-986.

Rostamian, M., Fazilatfar, A. M., & Jabbari, A. A. (2018). The effect of planning time on cognitive processes, monitoring behavior, and quality of L2 writing. *Language Teaching Research*, 22(4), 418-438.

Ryan, J. (2015). Overexplicit referent tracking in L2 English: Strategy, avoidance, or myth? *Language Learning*, 65(4), 824-859.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism*, 1(1), 1-33.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.). *Language Learning & Language Teaching*. John Benjamins Publishing Company: 239-273.

# From production short-cuts to syntactic development? Analysing the production of fixed expressions (FEs) with the development of the L2 computational component

Thomas A. Hammond
The University of Sheffield
tahammond1@sheffield.ac.uk

Usage-based models of second language acquisition (SLA) posit that general cognitive abilities allow learners to acquire an L2 through the analysis of prototypical and functional fixed expressions (FEs) derived from their early input (e.g., Ellis, 2012). Generative theories of SLA place relatively less importance on a learner's early L2 input and usage and maintain that the implicit abstract L2 computational system is a language-specific mechanism and therefore develops independently of learners' exposure to or use of prototypical FEs (e.g., Krashen and Scarcella 1978). This study combines these traditionally opposed views of L2 development as an alternative way of investigating the impact of prototypical FEs on the development of L2 syntax. Specifically, it seeks to examine whether learners who make more frequent use of FEs at the early stages of the acquisition have better acquired the L2 computational properties of the FEs at later stages of acquisition.

The data used for this analysis are spoken transcripts of Spanish child EFL classroom-learners of L2 English, who participated in naturalistic L2 interview tasks at early ages (10 & 12) and later ages (16 & 17). The analysis of four representative beginner EFL textbooks (both global and local) derived 4 functional FEs that were presented most frequently in the first half of all spoken tasks. These were the four conventional expressions *what's/is your name*, *how old are you*, *where are you from* and *where do you live*, which were searched for in learners' production data. The L2 computational properties assumed in the generation of these FEs are taken to be wh-movement, T-C movement, DO support, and A-movement. A learner's accuracy of these properties was measured via a learner's accurate production of an L2 surface form that is assumed to be a product of the computational property. In the case of DO support, for example, this would be a learner's accurate use of L2 negation and question formation surface structures that require the 'dummy' auxiliary DO i.e., *she doesn't like*, *do you want to go?* Accurate L2 surface forms are measured as a relative percentage out of L1, code-switched, and inaccurate L2 productions in all contexts where an L2 computational property is required to appear in an L2 surface form.

The analysis of all transcripts shows that all learners produce the FEs over multiple periods of data collection. For all learners, when the FEs are produced for the first time, they are done so fluently in absence of any other surface structure evidence of the FEs' related computational properties. At these beginner stages, the FEs are analysed as recalls from learners' phonological memory, rather than products of online generation through computational procedures. Pearson product-moment correlation coefficients show a strong correlation between earlier age of first FE production ($r$ = -.590) and higher frequency of FE productions at the early ages (10 & 12) ($r$ = .577) with learners' L2 accuracy of the FEs' related computational properties in the transcripts of the later ages (16 & 17). An independent-samples t-test also reported a significant difference in L2 computational accuracy for those learners who produced the FEs at early ages ($M$ = 61.2%, $SD$ = 24.8%) and those who produced them for the first time at the later ages ($M$ = 22.3%, $SD$ = 14.2%; $t$ (7) = 2.95, $p$ = <0.05).

The results suggest that as well as bootstrapping beginner learners into L2 production, the more frequent and earlier practice of FEs can increase the likelihood, or quicker the acquisition of, the FEs' underlying computational procedures (see also Paradis 2004). If the increased and earlier production of FEs is dependent on a learner's general working memory (WM) abilities, this places WM capacity as a general cognitive apparatus that can indirectly impact the development of the L2 language-specific computational mechanism.

**References**

Ellis, N. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, 32, 17-44

Krashen, S., & Scarcella, R. (1978). On routines and patterns in language acquisition and performance. *Language learning*, 28(2), 283-300

Paradis, M. (2004). Neurolinguistic Theory of Bilingualism. John Benjamins, Amsterdam-Philadelphia

# A corpus-based study of derivational morphology in written L2 Swedish

Kristoffer Holmquist[1], Therese Lindström Tiedemann[2]
University of Gothenburg[1], University of Helsinki[2]
kristoffer.holmquist@gmail.com[1], therese.lindstromtiedemann@helsinki.fi[2]

Learner morphology studies have usually focused on inflectional morphology (cf. Housen et al 2019; Brezina & Pallotti 2019). Only rarely have researchers tried to include derivational affixes (e.g., Horst & Collins 2006; De Clerq & Housen 2019). Similarly, in the field of L2 Swedish, inflection has been studied (see e.g., Philipsson 2013), but learners' use of derivational morphology remains relatively unexplored. We hypothesise that frequency of exposure to affixes can facilitate vocabulary development (cf. e.g., Ellis et al 2015) and in this paper, we use corpora to study how the use of derivational morphology develops and to compare learner language to reference corpora.

Our aim is to describe how often lemmas and tokens containing three nominal derivative suffixes are used in learner texts and textbooks for L2 Swedish and to compare this to L1 reference corpora. Based on this we explore the chance that learners at a certain level could have developed morphological awareness of these suffixes and relate this to the concept of *word families* (see e.g. Bauer & Nation 1993). At the group-level there are signs of development both in frequency and in lemmas per derivational morpheme. *Hapax legomena* are a definite sign of productivity on the individual level and indicate awareness. Many of the derivations occur rarely in the L2 essays, which may be due to vocabulary knowledge, individual preferences but also essay topics (cf. Caines & Buttery 2017). This dispersion needs to be born in mind and should be related, e.g., to textbook occurrences, topic, and genre, as well as frequency in reference corpora, to better grasp the development and how derivational morphology can be related to proficiency.

L2 learner essays (n=337, the SweLL-pilot corpus, Volodina et al 2016) were analysed for occurrences of the derivational suffixes '-het' (skönhet 'beauty'), '-skap' (vetenskap 'science') and '-(n)ing' (packning 'luggage'). Occurrences were compared with those in textbooks for L2 Swedish (the Coctaill corpus, Volodina et al 2014) regarding the relative frequency of the suffixes in learner input and output as well as receptive expectancies as evident in textbooks. In addition, we analysed the spread of these three suffixes to more stems as learners became more proficient. The results were then compared to a balanced corpus of texts written primarily by L1 speakers of Swedish (SUC 3.0, cf. Gustafsson-Capková & Hartmann 2006). All data were categorised and analysed manually, and care was taken to minimise the impact of factors such as misspellings and non-idiomatic word forms.

Results show that the derivational suffixes increase in relative use in the L2 essays as learners become more proficient, and a similar pattern is found in the L2 textbooks. The relative frequency and the distinct nominalizations in the L1 texts are considerably higher than in the learner essays and in the textbooks. This could be due to differences in topic and genre and requires further analysis. Some commonly used derivations can be traced to archaic stems which complicates the recognition of derivational morphemes (e.g., *drottning* 'queen' contains an archaic root *drott*), but others are quite clearly related to common lemmas (e.g., *forskning* 'research' – *forska* 'to research').

The study is limited in scope and has flaws pertaining (1) to task and topic, (2) learners' confidence to use new vocabulary, and (3) the risk of overlooked misspellings, which are discussed. We acknowledge the difficulty of confidently seeing the increasing occurrence of morphologically complex words as the result of increasing morphological awareness while emphasising the need to investigate this further. We argue that studies such as this one are needed as a first step towards gaining a better grasp of the morphological awareness among L2 learners and how this affects second language proficiency and development more generally and that it could be interesting to follow up with experiments.

**References**

Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography*, 6, 253–279.

Brezina, V. & Pallotti, G. (2019). Morphological complexity in written l2 texts. *Second language research*, 35(1), 99–119.

Caines, A. & Buttery, P. (2017). The Effect of Task and Topic on Opportunity of Use in Learner Corpora. In V. Brezina & L. Flowerdew (Eds.). *Learner Corpus Research: New Perspectives and Applications*. London: Bloomsbury Academic, 5–27.

De Clerq, B. & Housen, A. (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second language research*, 35(1), 71–97.

Ellis, N. C., Brook O'Donnell, M. & Römer, U. (2015). Usage-based language learning. In B. MacWhinney & W. O'Grady (Eds.) The handbook of language emergence. Hoboken, New Jersey: John Wiley & Sons. 163–180.

Gustafsson-Capková, S. & Hartmann, B. (2006). Manual of the Stockholm Umeå corpus version 2.0. https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf

Horst, M. & Collins, L. (2006). From faible to strong: How does their vocabulary grow?. *Canadian Modern Language Review*, 63(1), 83-106.

Housen, A., De Clercq, B., Kuiken, F. & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second language research*, 35(1), 3–21.

Philipsson, A. (2013). Svenskans morfologi och syntax i ett andraspråksperspektiv. In K. Hyltenstam & I. Lindberg (Eds.). *Svenska som andraspråk – i forskning, undervisning och samhälle*. Lund: Studentlitteratur, 121–150.

Volodina, E., Pilán, I., Rødven-Eide, S. & Heidarsson, H. (2014). You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. *Proceedings of the third workshop on NLP for computer-assisted language learning*. Linköping Electronic Conference Proceedings 107(10), NEALT Proceedings series 22 (10), 128–144. https://ep.liu.se/ecp/107/010/ecp14107010.pdf

Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G. & Sandell, M. (2016). SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M, Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.). *Proceedings of LREC 2016*, Slovenia. European Language Resources Association (ELRA), 206–212. https://arxiv.org/pdf/1604.06583v1.pdf

# The relevance of inter and intra-rater reliability in multi-layer annotation procedures

Olga Lopopolo[1], Fabio Zanda[2]
Eurac Research[1], University for Foreigners of Perugia[2]
olga.lopopolo@eurac.edu[1], fabio.zanda@unistrapg.it[2]

Manual annotation of linguistic features in learner data may be susceptible to human coding errors. To analyse potential coding discrepancies, it is possible to assess the degree of consistency of annotation for a single coder or between different coders by measuring coefficients of intra-rater reliability and of inter-rater reliability. As pointed out by Paquot & Plonsky (2017), though, only 11% of learner corpus studies measure and report inter-rater reliability coefficients, so it can be argued that "there is no established tradition of reporting or interpreting reliability coefficients in LCR" (Larsson et al. 2020: 239). This issue becomes more significant if we consider that some LCR studies involve the investigation of even more challenging linguistic categories inclined to subjective interpretation.

In this poster, we present the results of a series of reliability tests to enhance manual annotation procedures. In particular, we report on the methodological steps outlined to assess reliability within the ongoing PhD Project CROSSLIN3. The annotation process involved the manual annotation of all verbal forms contained in the English sub-section of LEONIDE (Glaznieks et al. 2022). Since the focus of the project is to investigate learner use of the progressive aspect, a multi-layer annotation scheme was developed considering different characteristics such as tense, aspect, semantics, type of deviation, and transfer. The manual annotation process involved the main researcher (Coder 1), who defined the architecture and the annotation guidelines, and three other coders (Coder 2, Coder 3, and Coder 4) who received the guidelines and were instructed by Coder 1.
Our contribution will illustrate how the following methodological questions have been addressed:

Q1: How can the reliability of coders' annotation be assessed and used to enhance internal validity?

Q2: How can the reliability of particularly challenging annotation layers be performed and interpreted?

In order to answer Q1, a series of tests have been carried out aiming at assessing the consistency of the architecture scheme and the guidelines as well as reducing the number of potential coding errors.

A first test aimed to assess the consistency of the architecture scheme and guidelines. In this first phase, a random sample of 15 texts was annotated by Coder 1 and Coder 2. Initially, raw agreement rate only was considered, and coders reached an overall 63% of cases in which both agreed. Cohen's Kappa (κ) was successively employed instead of percentages, reaching a κ average score of 0.74, suggesting "substantial agreement" (cf. Landis & Koch 1977). Thanks to these more precise results, it was also possible to solve ambiguities in the guidelines, leading to a more fine-grained version of the final documentation. Layers showing the lowest agreement concerned - not surprisingly - the most challenging categories to annotate, which were transfer phenomena ($\kappa = 0.43$) and semantics ($\kappa = 0.77$).

A second test was conducted to assess intra-rater reliability on a random sample of 20 texts annotated by Coder 3. Coder 3 annotated the same sample within two different periods of annotations (July-September), revealing κ score of 0.91. This result showed an "almost perfect agreement" in the consistency of Coder 3 annotations over time. We then proceeded with a comparison between Coder 1 and Coder 3 to assess inter-rater reliability on the same sample. The lowest κ score concerned transfer phenomena ($\kappa = 0.72$) and, unexpectedly, infinitive verbal forms ($\kappa = 0.37$). This is surprising as infinitives, in our view, were not supposed to be a problematic level of annotation. A confusion matrix was therefore computed (GitHub - olopopolo/exb_tools) considering labels for each level annotations: it revealed that 78% of infinitives in the sample were mistagged by Coder 3 as a systematic error. This led to a complete check of all the annotations, which were then corrected manually. The different results of intra- and inter-rater reliability suggested that both procedures were necessary and complementary to shed light on the source of annotators' disagreement and therefore to enhance the internal validity of the study.

To answer Q2, we will test the reliability of a layer of annotation which proved to be a challenging category of the architecture scheme, i.e. verb aspectual semantics. This layer of annotation is considered crucial to test predictions of the Aspect Hypothesis about progressive constructions (Andersen & Shirai 1994), which is the main focus of the CROSSLIN3 project. The test is still ongoing and involves the annotation of a random sample of texts by Coder 1 and Coder 4 using Biber et al.'s (1999) seven-class taxonomy of semantic domains and Vendler's (1957) four-class *Aktionsart* categories. We expect Coder 1 and Coder 4 to reach a higher

coefficient of agreement when adopting Vendler's four-class taxonomy compared to Biber et al.'s seven-class taxonomy. Nevertheless, an overall higher coefficient of agreement does not ensure taxonomies to be flawless, as they all have limitations and are inclined to subjective interpretations. A more thorough analysis of each semantic category agreement may lead to a clearer awareness of the interpretation of both taxonomies by the coders.

**References**


Andersen, R. W., & Shirai, Y. (1994). Discourse motivations for some cognitive acquisition principles. *Studies in Second Language Acquisition*, 16, 133-156.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of spoken and written English*. Harlow, Essex: Pearson Education Ltd.

Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L. & Nicolas, L. (2022). LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, 8(1), 97-120.

Landis, J.R., & Koch, G.G. (1977). *The measurement of observer agreement for categorical data*. Biometrics, 33, 159-174.

Larsson, T., Paquot, M., & Plonsky, L. (2020). Inter-rater reliability in learner corpus research: Insights from a collaborative study on adverb placement. *International Journal of Learner Corpus Research* 6(2), 237-251.

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, 3(1), 61-94.

Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, 66(2), 143-160.

# Tracking the development of written language competence in L2 Italian: A NLP-based approach

Sara Maso
Università degli Studi di Padova
sara.maso@studenti.unipd.it

Linguists have long wondered about the relationship between writing competence and language proficiency in L2 research. The traditional approach to investigate this relationship, especially for languages other than English, was typically based on small corpora and monitored only a few linguistic features, often manually annotated from texts. In the last fifteen years, a heterogeneous variety of methods and tools deriving from computational linguistics, Natural Language Processing (NLP), and corpus linguistics research have been implemented to support large-scale text profiling analyses.

The present contribution moves in this framework and presents a corpus-based study, based on an innovative linguistic profiling methodology, to track the development of written language competence in Italian L2 in relation to three different CEFR (Council of Europe 2001) levels: A1, A2, B1. Specifically, two main questions are addressed, i.e.:

- Does the interlanguage complexity of texts vary across proficiency levels?
- Which elements of the learners' language system vary significantly between proficiency levels?

The study is performed on a collection of texts written by Italian L2 learners, which are extracted from the multilingual corpus MERLIN (Boyd et al. 2014). The analysed corpus is composed of approximately 800 Italian texts assessed at A1 (29 texts), A2 (381 texts), and B1 (394 texts) levels. The tasks of the prompts, the learners' L1, and their age (above 16, mean 30 y.o.) are mixed.

The adopted approach is inspired by the multidimensional analysis originally developed in the context of corpus linguistics (Biber 1995) and it consists of three fundamental steps. The first two are performed by Profiling-UD (Brunato et al. 2018), an NLP-based tool devised to carry out linguistic profiling of texts for multiple languages. The tool implements a two-stage process: linguistic annotation and linguistic profiling. Linguistic annotation is carried out by UDPipe (Straka et al. 2016) according to the Universal Dependencies (UD) annotation formalism (Nivre et al. 2016). Although the used parser is trained on texts representative of the standard language, it has already been used effectively in similar works, for example, to investigate the evolution of linguistic competence in Italian L1 learners (Miaschi et al. 2021). In the second step, Profiling-UD extracts from the parsed texts about 130 features representative of the underlying profile at different levels of linguistic description. For the purpose of this study, Profiling-UD is applied to the collection of A1, A2, and B1 texts considered as three separate sub-corpora. The third step is the pairwise comparison of the profiles representative of each proficiency level. To this end, a statistical significance test is applied to detect the linguistic features that vary significantly between texts at different proficiency levels. In particular, the analysis focused on multileveled features that model aspects of linguistic complexity, which is viewed as a structural property of the interlanguage determined by the variety of elements and the relationships between them (Pallotti 2015). Among these features, we considered lexical- (e.g. Type-Token Ratio), morphological- (e.g. distribution of functional and content parts of speech, distribution of lexical verbs and auxiliaries according to tense, mood, person), and syntactic-related ones (e.g. distribution of coordinated and subordinated clauses, the average length of dependency links, the average depth of parse trees).

The results are then compared to previous works that link linguistic complexity to CEFR levels (Gyllstad et al. 2014, Bernardini & Granfeldt 2019, Kuiken & Vedder 2019) and with the results obtained by the 'Progetto di Pavia' group on the development of spoken L2 Italian (Giacalone Ramat 2003). Among the main findings, it is observed that in the A1-A2 comparison only 34 out of 130 features vary significantly, while in the A2-B1 comparison 98 features show variation, suggesting a more consistent improvement of language competence over these levels. Focusing on the typology of linguistic features, texts belonging to A1 and A2 levels appear to be more homogeneous with respect to the lexical and syntactic dimensions, whereas the A2-level and B1-level texts show significant variation at all levels of linguistic description. In particular, the most frequently used measures at both lexical and syntactic levels (such as lexical richness, distribution of present and past verb forms, and clause complexity) confirm the results obtained in the previously cited works. Conversely, in contrast to what was

expected, the percentage of coordinated structures, i.e. a measure widely used to assess complexity, especially at the most basic levels of proficiency, does not appear to vary significantly.

In summary, this work shows that NLP-based technologies are mature for tracking the evolution of language competence and modelling fine-grained properties of interlanguage. In particular, the use of language-independent scales for evaluation (like CEFR) and of a tool like Profiling-UD, which adopts a common linguistic annotation framework, could be a booster for cross-linguistic studies in the field of complexity and proficiency level. The study also highlights the need for further studies addressing, for example, the development of dedicated treebanks of Italian L2 texts (as in Di Nuovo et al. 2019), which can improve the robustness of NLP tools in analyzing the peculiarities of these texts with respect to more standard ones. The creation of these dedicated resources and tools could be an important stimulus to promote L2 research and favouring the interaction between computational linguistics, CEFR-based descriptors, and linguistic complexity studies.

**References**


Bernardini, P., & Granfeldt, J. (2019). On cross-linguistic variation and measures of linguistic complexity in learner texts: Italian, French and English. *International Journal of Applied Linguistics 29.2*, 211–232.

Biber, D. (1995) *Dimensions of register variation: A cross-linguistic comparison*. Cambridge & New York, Cambridge University Press.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., & Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland 1281-1288.

Brunato, D., Cimino, A., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2020). Profiling-ud: a tool for linguistic profiling of texts. *Proceedings of the 12th Language Resources and Evaluation Conference*, 7145-7151.

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assment*. Cambridge: Cambridge University Press.

Di Nuovo, E., Bosco, C., Mazzei, A., & Sanguinetti, M. (2019). Towards an Italian learner treebank in Universal Dependencies. *6th Italian Conference on Computational Linguistics, CLiC-it 2019* (Vol. 2481, 1-6. CEUR-WS.

Giacalone Ramat, A. (2003). *Verso l'italiano: Percorsi e strategie di acquisizione*. Carocci.

Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *Eurosla yearbook*, 14(1), 1-30.

Kuiken, F., & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied Linguistics 29.2*, 192–210.

Miaschi, A., Brunato, D., & Dell'Orletta, F. (2021). An NLP-based stylometric approach for tracking the evolution of L1 written langage competence. *Journal of Writing Research*, 13(1).

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016, May). Universal dependencies v1: A multilingual treebank collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), 1659-1666.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research 31.1*, 117–134.

Straka, M., Hajic, J., & Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), 4290–4297.

# Variants and varieties of learning preserved in the historical archives of the University for Foreigners of Perugia: Toward the building of a digital learning corpus

Alice Migliorelli
University for Foreigners of Perugia, Tor Vergata University of Rome
alicemiglio@virgilio.it, alice.migliorelli@unistrapg.it

The research project involves the study and valorization of a corpus of written exams made by students of ITAL2 included in the chronological space that goes from 1926 to 1931; then the linguistic-textual investigation will include the collection of the Registers kept by the teachers and the personal data forms of the students compiled by the Administration Office during the same five-year period.

Just by virtue of a relational and integrated graphic-textual interface of the learning corpus stored in the Historical Archive of the University for Foreigners of Perugia, it will be possible to make the students' personal forms and their respective written linguistic performances dialogue. These performances are analyzed mainly according to a criterion of marking errors, revealing the transitional competence, that is interlanguage, of learners with different L1 backgrounds.

Indeed the errors, thanks also to the decisive contribution of the cognitive theories of the end of the 1950s, have regained a central, active, and creative role in the linguistic learning path, in which they are to be considered precious evidence of the way in which the student reflects and formulates hypotheses on the functioning of language, continually reworking and renegotiating cognitive strategies and processes of signification (D'Annunzio & Serraggiotto 2007).

The objectives are therefore summarized as follows.

1) critical observation of the linguistic behavior of the learners through the textual analysis of the written tests of those enrolled at the University for Foreigners of Perugia from 1926 to 1931;
2) analysis of the registers and diaries of the lessons, compiled by the teachers of ITAL2 in the corresponding academic years, in order to evaluate, contextually to the progress of the varieties of learning, the methodologies, and glottodidactics techniques characterizing, at the same time, the teaching of Italian as a second language;
3) predisposition of a textual bank of written production of Italian L2 of diachronic type with the first sample of texts of the years 1926-1931, that, thanks to a web interface, allows different types of textual interrogation, with particular attention to the phenomenology of errors.

The application of the interpretative categories of textual linguistics is aimed at the reconstruction of the "grammar of the text", which can be rebuilt starting from a critical, systematic, and organic observation of the single linguistic behaviors carried out by the learners during the evaluation of their ability to produce writing in L2. The available corpus, offering a copious, heterogeneous, and chronologically extended sample of texts, also lends itself to a sociolinguistic analysis of the varieties of learning, transversal to the different classificatory ranks, to which, conventionally, the (superficially) so-called "deviations from the norm" are ascribable: phonetics/phonology, orthography, morphosyntax, vocabulary, punctuation, construction of the discourse (Andorno 2012).

The written productions in ITAL2 of the candidates constitute a domain of observation and an area of investigation that is potentially rich in circumstantial signals about the physiognomy of the varieties of learning, in the different phases of the path of linguistic acquisition, characterized, by the most part, by strategies of the pragmatic, semantic and morphosyntactic organization of utterances within discourse (Klein & Perdue 1997).
The learners' works, as a result of their role in the language acquisition process, have been presented in the form of a series of papers.

The students' works, as authentic linguistic data, will be analyzed taking into account a series of synchronic intra- and extra-textual factors, trying to hypothesize possible interferences from L1 or other languages known to the learner on the respective linguistic-textual performance.

Recently in the classroom, the students of the three-year degree course in Digital Humanities for Italian transcribed and analyzed, under supervision, various exam papers taken from the Historical Archive. This makes it possible to use the Archives for teaching purposes, bringing students closer to the practice of analyzing errors and digitizing them. It is important to remember that these students are mostly foreigners and therefore learners

of Italian L2: for a didactic activity of this kind, this is a strength and not a weakness because they are also multilingual speakers and able to recognize more spontaneously the systematic nature of the interlanguage.

**References**


Andorno, C. (2012). "Varietà di esiti dell'apprendimento dell'italiano nella varietà dei contesti di apprendimento: possibilità e limiti dell'acquisizione naturale". In Grassi, R. (Eds.). *Nuovi contesti d'acquisizione e insegnamento: l'italiano nelle realtà plurilingui, atti del convegno-seminario CIS 2012*. Perugia: Guerra, 157-173.

Andorno, C. (2009). "Grammatica e acquisizione dell'italiano L2", in *Italiano LinguaDue*, 1, 1-15.

Andorno, C. & Rastelli, M. (2009). *Corpora di italiano L2: tecnologie, metodi, spunti teorici*. Perugia: Guerra.

Balboni, P. E. (2018). *Fare educazione linguistica. Insegnare italiano, lingue straniere e lingue classiche*.
UTET Università.

Berruto, G. & Cerruti, M. (2015). *Manuale di sociolinguistica*. Torino: UTET.

Bettoni, C. (2001) *Imparare un'altra lingua*. Laterza.

Capaccioni, A. (2009). *Guida dell'archivio storico dell'Università per Stranieri di Perugia*. Perugia.

D'Annunzio, B. & Serraggiotto, G. (2007). *La valutazione e l'analisi dell'errore*. Venezia: Università Ca' Foscari.

Gheda, P. (2004). *La promozione dell'Italia nel mondo. L' Università per stranieri di Perugia dalle origini alla statizzazione*. Bologna: Il Mulino.

Grandi, N. (2015). "Le lingue naturali tra regole, eccezioni ed errori". In Grandi, N. (Eds.). *La grammatica e l'errore. Le lingue naturali tra regole, loro violazioni ed eccezioni.* Bologna: Bononia University Press, 7-33.

Grassi, R. (2018). *Il trattamento dell'errore nella classe di italiano L2: teorie e pratiche a confronto.* Firenze: Franco Casati Editore.

Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. London: Arnold.

Klein, W. & Perdue, C. (1997) "The basic variety (or: Couldn't natural languages be much simpler?)". In *Second language research*, 13, 4, 301-347.

Stramaccioni, A. (2005). *Un'istituzione per la lingua e la cultura italiana. L'Università per stranieri di Perugia (1925-2005)*. Perugia: Edimond.

Vetrugno, R. (2021). "Il muro della lingua. Stranieri residenti e apprendimento dell'italiano avanzato a Perugia", in Fiorentino, C. & Ceci, E. & Citraro, C. & Montinaro, A. (Eds.). *Alfabetizzazione come pratica di cittadinanza: teorie, modelli e didattica inclusiva.* «Lingue e Linguaggi». Università del Salento: 159-179.

# The Beldeko corpus: A new resource for investigating L2 German texts written by L1 Dutch students

Helena Wedig, Carola Strobl, Jim Ureel
University of Antwerp
{helena.wedig, carola.strobl, jim.ureel}@uantwerpen.be

In this poster, we present a new learner corpus for investigating German as a foreign language (L2): Beldeko (*Belgisches Deutschkorpus*). The corpus was created to investigate academic writing in L2 German by advanced learners with Dutch as their first language (L1). It contains summaries produced by L1 Dutch writers. Although there are several learner L2 German corpora available, most of them are heterogeneous with regard to the learners' L1s. This means that L1-specific characteristics of L2 German have not received due attention. One exception is the ALeSKo learner corpus (Zinsmeister & Breckle 2012), which consists of two subcorpora: L2 German essays written by L1 Chinese writers and comparable L1 German essays. The largest and best-known German learner corpus to date is the Falko corpus (Reznicek et al. 2012), which was compiled at the Humboldt−Universität zu Berlin. Other corpora with various L1s are the *Kommentiertes Lernendenkorpus akademisches Schreiben* (KOLAS; Knorr & Andresen 2017), which contains 854 academic texts produced by 233 students in the context of writing consultation given by peer tutors, and the MERLIN corpus (Abel et al. 2014), which contains 2,286 texts produced by learners of Italian, German and Czech taken from written exams of CEFR testing institutions. To date, no L2 German corpus produced by L1 Dutch students is available. The corpus being presented aims to close this gap.

The 301 summaries included in the Beldeko corpus (70,774 tokens) were written by 115 students with L1 Dutch. The texts were collected at Ghent University (in 2013 and 2014) and University College of Ghent (in 2013) as pretests, immediate posttests, and delayed posttests in an intervention study on collaborative writing. 82 students produced three summaries each and 33 students produced two summaries each. The tasks at hand were to write summaries of two popular scientific texts (newspaper articles, interviews, or websites) about a topic related to language variation in contemporary German (*Kiezdeutsch*, *Mundartdebatte in der Schweiz*, *Viadrinisch*, *Varianten-Wörterbuch des Deutschen*).

For a research project aimed at investigating cohesive strategies deployed by L2 German writers with L1 Dutch, the corpus was pre-processed and several linguistic annotation layers were added automatically: PoS tags, morphological information, lexemes, and universal dependencies. Moreover, a target hypothesis was added manually. In the course of the project, the corpus will be annotated with information about cohesive devices targeting several of the categories described by Halliday and Hasan (1976), starting with co-reference, conjunction, and lexical cohesion. These categories are especially interesting in the context of Dutch–German influences, as studies into Dutch–German translations have found that co-reference (in German) shifts to lexical cohesion (in Dutch) (Van de Velde 2011), which Dendooven (2018) explains as the result of language-specific grammatical restrictions on the one hand (e.g., *der Stuhl, auf dem er sitzt* vs *de stoel waarop hij zit*) and of language-specific preferences on the other (e.g., *man* vs *je*).

An automatic pre-annotation of these categories has been performed with the help of CorZu (Tuggener, 2016: coreference), DimLex (Scheffler & Stede 2016; Stede 2002: connectives) and GermaNet (Hamp & Feldweg 1997; Henrich & Hinrichs 2010: synonyms, hyponyms und hypernyms). Based on the automated pre-annotation, manual annotations will be conducted, using the annotation platform Inception (Klie et al. 2018) and guidelines based on PTDB3 scheme (Webber et al. 2019: connectives), the co-reference guidelines developed by Reznicek et al. (2012) and lexical cohesive devices as presented in Tanskanen (2006). These guidelines will be put to the test and possibly revised after a pilot phase, depending on the inter-annotator agreement. The poster will introduce the corpus to the research community and show preliminary results of the analysis of cohesion retrieved from the automatic annotation. This includes an analysis of the homogeneity of the corpus to investigate learner-specific use of cohesive devices.

**References**


Abel, A., Wisniewski, K., Nicolas, L., Boyd, A., Hana, J., Meurers, D. (2014). *A trilingual learner corpus illustrating European Reference Levels. Ricognizioni – Rivista di Lingue, Letterature e Culture Moderne, 2(1)*, 111–126.

Dendooven, F. (2018). *Die Übersetzung von Koreferenzmitteln: Eine Studie auf Basis eines deutsch–niederländischen Übersetzungskorpus von Museumstexten* [Unpublished master's thesis]. Ghent University.

Hamp, B., & Feldweg, H. (1997). GermaNet: A lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (pp. 9–15). Association for Computational Linguistics.

Henrich, V., & Hinrichs, E. (2010). GernEdiT: The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC 2010) (pp. 2228–2235). ELRA.

Knorr, D., & Andresen, M. (2017). Commented Learner Corpus Academic Writing (KoLaS). Archived in *Hamburger Zentrum für Sprachkorpora*. Version 2.0. http://hdl.handle.net/11022/0000-0001-B732-8.

Klie, J. C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrationsm* (pp. 5–9). Association for Computational Linguistics.

Reznicek, M., Lüdeling, A., & Schwantuschke, F. (2012). *Das Falko-Handbuch: Korpusaufbau und Annotationen: Version 2.01*. Humboldt-Universität zu Berlin. Institut für deutsche Sprache und Linguistik - Korpuslinguistik. Retrieved from https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v2.01

Scheffler, T., & Stede, M. (2016). Adding semantic relations to a large-coverage connective lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 16)* (pp. 1008–1013). ELRA.

Stede, M. (2002). DiMLex: A lexical approach to discourse markers. In A. Lenci & V. Di Tomaso (Eds.), *Exploring the lexicon: Theory and computation* (pp. 1–15). Edizioni dell'Orso.

Tanskanen, S. K. (2006). *Collaborating towards coherence: Lexical cohesion in English discourse*. John Benjamins. https://doi.org/10.1075/pbns.146

Tuggener, D. (2016). *Incremental coreference resolution for German* [Doctoral dissertation]. University of Zürich.

Van de Velde, M. (2011). Explizierung und Implizierung im Übersetzungspaar Deutsch Niederländisch: Eine quantitative Untersuchung. In P. A. Schmitt, S. Herold, & A. Weilandt (Eds.), *Translationsforschung* (pp. 865–884). Peter Lang.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). *The penn discourse treebank 3.0 annotation manual.* University of Pennsylvania.

Zinsmeister, H., & Breckle, M. (2012). The ALeSKo learner corpus. *Multilingual Corpora and Multilingual Corpus Analysis*, *14*, 71–96. https://doi.org/10.1075/hsm.14.06zin

# Investigating spoken classroom interactions in linguistically heterogeneous learning groups – An interdisciplinary approach to compile multi-modal corpora in second language classrooms

Zarah Weiss[1], Moritz Sahlender[2], Inga ten Hagen[3], Anastasia Knaus[4], Stefanie Helbig[5]
University of Tübingen[1], German Institute for Adult Education – Leibniz Centre for Lifelong Learning[2], TU Dortmund University[3], Mercator-Institute for Literacy and Language Education[4,5]
zweiss@sfs.uni-tuebingen.de[1], sahlender@die-bonn.de[2], inga.tenhagen@tu-dortmund.de[3], anastasia.knaus@mercator.uni-koeln.de[4], stefanie.helbert@mercator.uni-koeln.de[5]

There still are relatively few spoken learner corpora (Gilquin, 2015; Paquot & Plonsky, 2017), which is an issue given the considerable linguistic differences between spoken and written language (see e.g., Biber et al., 1999; Goulart et al., 2020; Koch & Oesterreicher, 1985). Spoken language also is crucial in models of communicative L2 competence (Canale & Swain, 1980; Salaberry et al., 2019) and in the Common European Framework of Reference (Council of Europe, 2001). Spoken learner corpora of authentic teacher-learner classroom interactions are especially underrepresented although they not only facilitate the study of L2 learners' oral and interactive proficiency but also provide insights into the role of adaptive input in L2 teaching practice.

Increasingly elaborate and variable input in the learner's Zone of Proximal Development (Vygotsky, 1979) fosters L2 acquisition (Cummins, 2000; Krashen, 1985; Swain, 1985). Yet, it remains unclear to which extent teachers succeed in adapting their language in practice, especially given the challenging heterogeneity of learners in many SLA classrooms. Classroom recordings are also of interest for empirical educational research where video studies on classroom interactions support insights into pedagogical-psychological processes such as instructional quality, teaching practices, or learner-teacher-interactions (e.g., Lotz, 2016; Seidel et al., 2017). Yet, these data are rarely compiled into machine-readable, re-usable, well-documented, and accessibly licensed multi-modal corpora. We see untapped potential for collaboration between empirical educational and learner corpus research (LCR).

We present an interdisciplinary multi-modal L2 corpus of authentic teacher-learner interactions in German as a Second Language (GSL) classrooms at the interface of LCR and empirical educational research. The corpus currently consists of video recordings of 59 45-minute GSL lessons for beginning to low intermediate GSL learners that took place in equal parts in preparatory classes in early secondary schools and integration courses in adult education institutions in Germany. We used questionnaires and standardized tests to elicit rich meta information on teachers' and learners' demographics (gender, age, country of birth, languages spoken, educational background), learners' motivation to learn German, and their German proficiency using a C-Test. For teachers, we elicited information regarding their teaching qualifications and expertise, their motivation and beliefs (self-efficacy, enthusiasm, work satisfaction, stress, burnout, perspective on multilingualism), and their professional competencies (pedagogical knowledge, GSL teaching competence, professional vision). We also elicited teachers' assessments of each of their students (language proficiency, motivation, classroom behavior).

We used the EXMARaLDA transcription editor (www.exmaralda.org; Schmidt & Wörner, 2014) to create time-aligned transcriptions of all teacher and learner utterances in a normalized orthographic transcription layer. In our systematic piloting of the annotation process, we also considered a narrower transcription following the cGAT standard (Selting et al. 2009) which however turned out to not be sufficiently reliable given the available annotators and not relevant to the research questions pursued in our project. Only study participants who gave informed consent were included in the transcription. We encoded the addressees for each utterance to support the analysis of targeted adaptation processes. The corpus also contains several time-aligned annotation layers encoding gestures, social forms, and the initiation of classroom participation to support multi-modal analyses of classroom interactions. The entire transcription and annotation procedure was documented in our extensive annotation manual which was designed to also support the annotation of future corpora of multi-modal classroom interactions. We evaluated the robustness of our transcription and addressee annotation on two full-lesson recordings and the remaining annotation layers on our full pilot corpus comprised of ten full-lesson recordings (we increased the number of recordings due to the lower frequency of these non-verbal cures within individual lessons). We used seven annotators paired into eleven unique annotator pairings. For all annotation layers, we observe high to a near-perfect agreement as measured by Cohen's kappa (Landis & Koch, 1977). We evaluate the robustness of the transcription using turn-wise Levenshtein distance (Levenshtein, 1966) normalized by maximal turn length, again finding little disagreement between annotators.

The pilot corpus and the annotation guidelines will be made available under a CC-BY-SA 4.0 license by January 1, 2023. The full corpus will be made available later. Our work directly fosters the interdisciplinary study of teacher-student interactions, teacher competencies, and language acquisition. It allows to pursue several research avenues at the intersection of empirical educational research, SLA, and LCR.

**References**


Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman Publications Group.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire. Multilingual Matters (Bilingual education and bilingualism)*, 23.

Gilquin, G. (2015). From design to collection of learner corpora. S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*, 9-34. https://doi.org/10.1017/CBO9781139649414.002

Goulart, L., Gray, B., Staples, S., Black, A., Shelton, A., Biber, D., Egbert, J., & Wizner, S. (2020). Linguistic perspectives on register. *Annual Review of Linguistics*, 6, 435-455. https://doi.org/10.1146/ANNUREV-LINGUISTICS-011718-012644

Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte [Language of Proximity – Language of Distance. Orality and writing in the field of tension between language theory and language history]. *Romanisches Jahrbuch*, 36, 15-43. de Gruyter.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Longman.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. https://doi.org/10.2307/2529310

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707-710.

Lotz, M. (2016). *Kognitive Aktivierung im Leseunterricht der Grundschule: Eine Videostudie zur Gestaltung und Qualität von Leseübungen im ersten Schuljahr* [Cognitive activation in primary school reading lessons: A video study on the design and quality of reading exercises in the first school year]. Springer-Verlag.

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, 3, 61-94. https://doi.org/10.1075/ijlcr.3.1.03paq

Salaberry, M. R., Kunitz, S., Sandlund, E., & Sundqvist, P. (2019). Doing versus assessing interactional competence. M. R. Salaberry & S. Kunitz (Eds.), *Teaching and Testing L2 Interactional Competence: Bridging Theory and Practice*. Routledge.

Schmidt, T., & Wörner, K. (2014). EXMARaLDA. J. Durand, U. Gut, & G. Kristoffersen (Eds.), *Handbook on Corpus Phonology*, 402-419. Oxford University Press.

Seidel, T., & Thiel, F. (2017). Standards und Trends der videobasierten Lehr-Lernforschung [Standards and trends in video-based teaching-learning research]. *Zeitschrift für Erziehungswissenschaft*, 20(1), 1-21. https://doi.org/10.1007/s11618-017-0726-6

Selting, M., P. Auer, D. Barth-Weingarten, J. Bergmann, P. Bergmann, K. Birkner, E. Couper-Kuhlen, A. Deppermann, P. Gilles, S. Günthner, M. Hartung, F. Kern, C. Mertzlufft, C. Meyer, M. Morek, F. Oberzaucher, J. Peters, U. Quasthoff, W. Schütte, A. Stukenbrock, and S. Uhmann. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung*, 10, 353-402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf

Swain, M. (1985): Communicative competence: Some roles of comprehensible input and comprehensible output in its development. Susan M. Gass and Carolyn G. Madden, editors, *Input in second language acquisition*, 235-253. Newbury House, Rowley, MA.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

# Czech errors in writings based on the Polish learner corpus PoLKo: A pilot study

Adrian Jan Zasina[1], Elżbieta Kaczmarska[2]
Charles University[1], University of Warsaw[2]
adrian.zasina@ff.cuni.cz[1], e.h.kaczmarska@uw.edu.pl[2]

As learners acquire foreign languages, they tend to apply the structures of their mother tongue to a target language. This phenomenon, called language transfer, became the focus of second language acquisition more than fifty decades ago (Selinker 1969) and it is still crucial in language teaching and learning (Gass and Swlinker 1992). Knowing transfer errors, teachers are able to adjust classroom materials and be aware of features that may be easy or difficult for learners (Odlin 1989: 4). Moreover, similarities between languages can be helpful in language acquisition (1989: 27). In particular, it concerns the genetically close languages such as Polish and Czech, in which there is a high level of mutual interference. Czech learners benefit from a positive transfer while using common structures for both languages. They acquire the Polish language system much faster than non-Slavic learners. However, they also get into a trap using structures seemingly common for both languages caused by a negative transfer. As a result, the production of Czech learners in the Polish language is characterised by a unique set of language errors, which is typical for this group of learners.

Since there is a large interest in language errors of foreigners studying Polish (Dąbrowska 2004; Dąbrowska & Pasieka 2008; Kita et al 2008; Krawczuk 2009; Skura 2013; Dąbrowska & Pasieka 2014; Górska 2015; Kowalewski 2018; Skura 2018; Kaczmarska & Zasina 2020) only few studies (Pösingerová 2001) pay attention to errors of Czechs. Our research (Kaczmarska & Zasina 2021) is the first attempt to explore this area using a learner corpus.

**Data**

The pilot study uses a part of the PoLKo learner corpus that consists of 28 texts written in Polish by Czech native speakers. It makes in total 8,721 tokens. The examined sample covers essays written as a homework or an exam task. As a corpus manager, the TEITOK tool (Janssen 2016) was used.

**Analysis**

The pilot analysis was provided manually as the corpus does not yet have any error annotation. It investigates the most prominent errors in the areas of syntax, word order, spelling, and lexis. These areas are discussed sequentially with examples. Each example has its learner's and revised version with an English translation in brackets. The given instances also have symbols that indicate, respectively, gender, age, and language level of Polish.

The syntax errors represent most often distortions of valency patterns, i.e. failures to the rules of syntactic connections that are also classified as grammatical errors. One of the most prominent issues in this area is the use of the accusative case instead of the genitive after a verb with negation. In the contemporary Czech language, negation does not require changing the case into genitive. A typical example presents the following sentence:

(1) **Wczoraj* ***nie zobaczyłam Ewę****, było za ciemno*. (F, 27, A2)
*Wczoraj* ***nie zobaczyłam Ewy****, było za ciemno*.
(I didn't see Ewa yesterday because it was dark outside.)

The right word order is the next area that causes problems for Czech learners. For instance, it is visible in the example of the reflexive pronoun *się*:

(2) **Dzień wcześniej długo uczyłem* ***się*** *i ranem zaspałem budzik.* (M, 22, A1)
*Dzień wcześniej długo* ***się*** *uczyłem i rano zaspałem*.
(The day before, I studied for a long time and I overslept.)

Lexical errors in this language group are mainly related to interlingual interference.

(3) *(...) w Czechach* ***wolimy*** *między matematyką albo językiem obcym.* (M, 22, A1)
*(...) w Czechach* ***wybieramy*** *pomiędzy matematyką albo językiem obcym*.
((…) in the Czech Republic, we are choosing between mathematics and foreign language.)

The Czech verb *volit* means 'to vote; to choose'; therefore, this is the most likely cause of this error in the Polish language.

The last area presents spelling errors that are primarily due to the insufficient language competence of the learner. Knowledge of Polish orthography can be divided into three issues in terms of the difficulties of Czech

native speakers: 1) writing palatal and retroflex consonants (e.g. *ś* vs. *sz*), 2) confusing the letters *ł* and *l*, 3) distinguishing phonemes (letters) *i* and *y*.

**Conclusion**

Our preliminary observations provide an output for learners and teachers that could be used to create customized learning materials. This approach undoubtedly has the opportunity to improve the teaching methods of Polish as a foreign language among Czech-speaking learners. Precisely knowing the weakness of the learners' group makes it possible to focus on difficulties rather than general issues that are not related to Czech speakers. However, further analysis based on a larger number of texts is still needed to shed light on unobvious language errors.

**References**


Dąbrowska, A. (2004). Najczęstsze błędy popełniane przez cudzoziemców uczących się języka o*rozwijanie i testowanie znajomości języka polskiego jako obcego* (105-136). Kraków: Universitas.

Dąbrowska, A., & Pasieka, M. (2008). Nowa typologia błędów popełnianych przez cudzoziemców w języku polskim. In M. Kita, M. Czempka-Wewióra & M. Ślawska (Eds.). *Błąd językowy w perspektywie komunikacyjne*. Katowice: Wyższa Szkoła Zarządzania Marketingowego i Języków Obcych, 73-102.

Dąbrowska, A., & Pasieka, M. (2014). *Badania błędów cudzoziemców prowadzone w Szkole Języka Polskiego i Kultury dla Cudzoziemców UWr*. In A. Dąbrowska & U. Dobesz (Eds.). *40 lat wrocławskiej glottodydaktyki polonistycznej: Teoria i praktyka*. Wrocława: Oficyna Wydawnicza ATUT, 331-342.

Gass, S. M., & Selinker, L. (Eds.). (1992). *Language transfer in language learning: Revised edition* (Vol. 5). John Benjamins Publishing.

Górska, A. (2015). Błędy studentów z Ukrainy – zapobieganie i eliminacja w grupach o zróżnicowanych możliwościach (na podstawie doświadczeń Centrum Partnerstwa Wschodniego Uniwersytetu Opolskiego). *ACTA UNIVERSITATIS LODZIENSIS*, *Kształcenie Polonistyczne Cudzoziemców*, 22, 357-370. https://doi.org/10.18778/0860-6587.22.24

Janssen, M. (2016). TEITOK: Text-faithful annotated corpora. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož: ELRA, 4037-4043.

Kaczmarska, E., & Zasina, A. J. (2020). Błędy walencyjne w tekstach obcokrajowców uczących się języka polskiego w świetle korpusu PoLKo. *Prace Filologiczne*, 75(1), 197-213. https://doi.org/10.32798/pf.657

Kaczmarska, E., & Zasina, A. J. (2021). Język polski w tekstach osób czeskojęzycznych na podstawie korpusu uczniowskiego PoLKo. *ROSSICA OLOMUCENSIA*, LX (2), 5-17.

Kita M., Czempka-Wewióra M., & Ślawska M. (Eds.). (2008). *Błąd językowy w perspektywie komunikacyjnej*. Katowice: Wyższa Szkoła Zarządzania Marketingowego i Języków Obcych.

Kowalewski, J. (2018). Badania statystyczne nad błędami językowymi uczących się języka polskiego na Ukrainie. In M. Maciołek (Ed.), *Polonistyka na początku XXI wieku. Diagnozy. Koncepcje. Perspektywy. Tom III: Współczesne aspekty badań nad językiem polskim – teoria i praktyka*. Katowice: Wydawnictwo Uniwersytetu Śląskiego, 277-298.

Krawczuk, A. (2009). Błędy leksykalne i leksykalno-stylistyczne w polszczyźnie Ukraińców, *Postscriptum Polonistyczne*, 1(3), 167-183.

Odlin, T. (1989). *Language transfer*. Cambridge: Cambridge University Press.

Pösingerová, K. (2001). *Problematika negativních transferů při výuce polského jazyka v českém jazykovém prostředí*. Univerzita Karlova v Praze, nakladatelství Karolinum.

Selinker, L. (1969). Language transfer. *General linguistics*, 9(2), 67.

Skura, M. (2013). Błędy wynikające z interferencji kulturowej popełniane przez Niemców uczących się języka polskiego jako obcego, *ACTA UNIVERSITATIS LODZIENSIS*, *Kształcenie Polonistyczne Cudzoziemców*, 20, 149-158.

Skura, M. (2018). *Błędy popełniane przez Niemców uczących się języka polskiego jako obcego – implikacje glottodydaktyczne*, (Doctoral thesis). Wydział Polonistyki Uniwersytetu Warszawskiego, Warszawa.

Zasina, A. J., & Kaczmarska, E. (2020). *Infrastructure of the Polish Learner Corpus PoLKo*. Retrieved from https://www.researchgate.net/publication/342888260_Infrastructure_of_the_Polish_Learner_Corpus_PoLKo. https://doi.org/10.13140/RG.2.2.23874.40648

## Towards crowdsourcing research for learner keylogging data

Nicolas Ballier[1], Helen Yannakoudakis[2]
Université de Paris[1], King's College London[2]
nicolas.ballier@u-paris.fr[1], helen.yannakoudakis@kcl.ac.uk[2]

This paper presents the methodology and preliminary results of an ongoing research project funded by a joint research grant shared between King's College London and Université de Paris Cité.

Keyloggers are devices that record typing events and that can be used to analyse writing processed with time stamps. Even though the tools have been available for some fifteen years (Sullivan and Lindgren 2006) and have been used as a microscope for activity mining in writing (Leijten and Van Waes 2013) and translation (Tirkkonen-Condit 2005), it is only recently that the learner corpus research community has begun investigating learner data with these tools (Ballier et al. 2018, Gilquin 2020, Gilquin & Laporte, 2021), especially with the advent of the PROCEED corpus (the Process Corpus of English in Education) at the University of Louvain (Gilquin 2022). Several systems have been designed recently to facilitate a linguistic analysis of keylogs (Goodkind 2021, Mahlow et al 2022). We introduce a JavaScript keylogger that allows anonymous data collection; its possible integration to websites; and an R package designed to analyse the data. We will discuss the metadata collection in compliance with GDPR and in relation to other existing online systems. On top of data collection organised with research assistants and colleagues, we have also set up a web-based interface that integrates keylog data collection into an automatic grading system of essays. We collected data for several tasks including essay writing, translation, and picture description. For our picture description task, we replicated Berman & Slobin's (1987) seminal study to investigate 'how to talk about events', using a visual prompt which has also been chosen as part of the COREFL (Lozano et al. 2020) protocol; therefore, potentially increasing interoperability across learner corpora.

We will demo the web-based interface and some of the scripts already available in the R package we designed. We will present preliminary results from our data collection and demo how we process data to visualise "textual bursts" to investigate the writing processes and the preliminary results for the investigation of phraseological units and "islands of reliability" (Dechert 1983) in learner productions. Scripts calculating typing speeds and other metrics (Conijn 2019) will be used for supervised learning of CEFR levels as the essays will be graded by experts. Our data modelling intends to build on previous research for cross-task investigation (Conijn et al., 2019). The project would like to address millennials' misconceptions of textual structure (short paragraph structures) biased by constant short messages on mobile phones (SMS) and social media. We are designing paragraph-based metrics to assess the role of paragraph structure in writing competence. At the end of the project, we hope to improve student writing competency at the academic level by guiding them with (experimental) automated visual feedback displaying their performance and comparing it with native performance.

**References**

Granger, S., Gilquin, G., & F. Meunier (2015). *Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP.

Li, P., Eskildsen, S., & Cadierno, T. (2014). Tracing an L2 learner's motion constructions over time: a usage-based classroom investigation. *The Modern Language Journal*, 98(2), 612-628.

Van Hout, R. & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton & J. Treffers-Daller (Eds.). *Modelling and Assessing Vocabulary Knowledge*. Cambridge: CUP, 93-115.

Ballier, N., Pacquetet, E., & Arnold, T. (2018). Investigating Keylogs as time-stamped graphemics. In *Graphemics in the 21st Century*, 353-365.

Berman, R. A., & Slobin, D. I. (1987). Five ways of learning how to talk about events: A crosslinguistic study of children's narratives. Cognitive Science Program, Institute of Cognitive Studies, University of California at Berkeley.

Conijn, R., Roeser, J., and Van Zaanen, M. (2019). Understanding the keystroke log: the effect of writing task on keystroke features. *Reading and Writing*, 32(9), 2353–2374.

Dechert, H.-W. (1983) How a story is done in a second language. In C. Farch & G. Kasper, eds., *Strategies in interlanguage communication*, London; New York: Longman, 175–195.

Gilquin, G., & Laporte, S. (2021). The use of online writing tools by learners of English: Evidence from a process corpus. *International Journal of Lexicography*, 34(4), 472-492.

Gilquin, G. (2020). In search of constructions in writing process data. *Belgian Journal of Linguistics*, 34(1), 99-109.
Gilquin, G. (2022). The Process Corpus of English in Education: Going beyond the written text. *Research in Corpus Linguistics*, 10(1).
Goodkind, A. (2021). TypeShift: A User Interface for Visualizing the Typing Production Process. arXiv preprint arXiv:2103.04222.
Leijten, M. and Van Waes, L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3), 358–392.
Lozano, C., Diaz-Negrillo, A. & M. Callies (2020). Designing and compiling a learner corpus of written and spoken narratives: The Corpus of English as a Foreign Language (COREFL), in C. Bongartz & J. Torregrossa (eds.), *What's in a Narrative? Variation in Story-Telling at the Interface Between Language and Literacy*. Frankfurt/Main: Peter Lang, 21-45.
Mahlow, C., Ulasik, M. A., & Tuggener, D. (2022). Extraction of transforming sequences and sentence histories from writing process data: a first step towards linguistic modeling of writing. *Reading and Writing*, 1-40.
Sullivan, K. and Lindgren, E. (2006). *Computer keystroke logging and writing*. Brill.
Tirkkonen-Condit, S. (2005). The monitor model revisited: Evidence from process research. *Meta: journal des traducteurs/Meta: Translators' Journal*, 50(2), 405–414.

# Meeting ROGER: An open-access bilingual corpus search platform

Mădălina Chitez[1], Cosmin Striletchi[2], Karla Csürös[3]
West University of Timisoara[1,3], Technical University of Cluj-Napoca[2]
madalina.chitez@e-uvt.ro[1], cosmin.striletchi@com.utcluj.ro[2], karla.csuros98@e-uvt.ro[3]

This software demonstration proposal presents the open access bilingual corpus search platform ROGER. The platform supports searches within the recently released ROGER – Corpus of Romanian Academic Genres (Chitez et al. 2021), compiled by the research group at CODHUS (Centre for Corpus Related Digital Humanities) from the West University of Timisoara. ROGER consists of novice academic writing genres, in Romanian (L1) and English (L2), collected from Romanian universities, with the purpose of investigating student writing practices. The corpus was compiled between 2018 and 2021 and it amounts to 3.11 million words.

Processing the ROGER corpus data was performed in several stages. Original texts were submitted in various printed or handwritten formats. The digital variants were converted to .txt files. The handwritten texts were transcribed manually by our research assistants. The corpus databank consists of processed .txt files with UTF-8 encoding. Learners' metadata is stored in .xlxs files which are automatically parsed by the ROGER online search platform. The corpus data is anonymized and follows European GDPR laws.

The ROGER corpus support platform has been designed and implemented as a cross-platform distributed web application. Its main purpose is to offer access to learner academic writing corpus data that can be consulted and assessed in a multidimensional contrastive framework. ROGER features texts from eight disciplines: (i) Humanities; (ii) Economics; (iii) Political Sciences; (iv) Engineering; (v) Computer Science; (vi) Law; (vii) Mathematics; (viii) Social Sciences. In each discipline, the students labeled the genre of their own writings with: (i) Essay; (ii) Scientific paper; (iii) Thesis; (iv) Literary analysis; (v) Others (to be elaborated further).

The platform frontend capabilities are offered to registered users, allowing them to search for specific keywords and to refine the obtained results by applying a series of filters. Creating a regular user account on the ROGER platform is entirely free of charge and is based on an email login system. Current platform features for regular users include search terms and phrases, n-gram distributions, and statistical visualizations for performed queries. After inputting a search term/ phrase, regular users may filter texts by: (i) language; (ii) genre; (iii) study year; (iv) level; (v) discipline and (vi) gender. Any registered regular user can contact the platform administrators via the Contact page on the ROGER website to become an enhanced user, i.e. have less restrictions on accessing and downloading ROGER data. The platform also includes instructional tutorials and a section dedicated to research results based on ROGER corpus data. All data made available on the ROGER platform is protected by the CC BY-NC-ND license. The backend interface of the ROGER platform is available to authenticated administrators and it provides the digital tools for managing the database's stored texts and associated metadata, while also offering an extensive statistics mechanism that covers the data composition and usage. All functionalities will be illustrated in the demo.

The ROGER platform is comparable to other corpus support search platforms such as MICUSP – Michigan corpus of Upper-Level Student Papers (2009), CROW – Corpus and Repository of Writing (Staples & Dilger 2018), BAWE – British Academic Written English Corpus (Nesi & Gardner 2012) and ICLE – International Corpus of Learner English (Granger et al. 2020). What sets ROGER apart is the fact that it is the first bilingual learner academic writing corpus with a dedicated freeware corpus query platform. What is more, ROGER represents the first corpus offering information about L2 English academic writing in the Romanian context, while also offering discipline-specific information about learner academic writing from a contrastive perspective.

Through the ROGER corpus search platform, researchers, teachers, students, and general users are offered free access to the ROGER corpus data. ROGER can inform two main types of research topics: academic writing contrastive studies and corpus-based genre analyses. Case studies conducted by CODHUS researchers (Chitez & Bercuci 2019, Chitez et al. 2020, Bercuci 2020, Dincă & Chitez 2021) have used the ROGER corpus to investigate rhetorical features (author roles, hedging, metadiscourse) and linguistic features (lexico-grammatical profiles, academic phrases) of novice student writing. As such, we believe that the ROGER corpus search platform, especially due to its free-access and bilingual features, will be of interest to experts in areas such as Corpus Linguistics, Academic Writing, and Contrastive (Interlanguage) Analysis, Language for Specific Purpose studies, and Computer-Assisted Language Learning.

**References**


Bercuci, L. (2020). Discipline-specific Metadiscourse Markers in ESP Expert Writing in Political Science. In R.M. Nistor, & C. Teglaș (Eds.). *Limbajele specializate în contextul noilor medii de învățare: Provocări și oportunități*. Cluj-Napoca: Presa Universitară Clujeană, 331-345.

Chitez, M. & Bercuci, L. (2019). Data-driven learning in ESP university settings in Romania: multiple corpus consultation approaches for academic writing support. In F. Meunier, J. Van de Vyver, L. Bradley, L. & S. Thouësny (Eds). *CALL and complexity – short papers from EUROCALL 2019*. *Research-publishing.net*, 75-81.

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

Chitez, M., Rogobete, R. & Foitoș, A. (2020). Digital Humanities as an Incentive for Digitalisation Strategies in Eastern European HEIs: A Case Study of Romania. In A. Curaj, L. Deca & R. Pricopie (Eds.). *European Higher Education Area: Challenges for a New Decade*, Cham, Springer, 545-564.

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

Granger, S., Dupont, M., Meunier, F., Naets, H. & Paquot, M. (2020) *The International Corpus of Learner English. Version 3*. Louvainla-Neuve: Presses universitaires de Louvain. Available at https://dial.uclouvain.be/pr/boreal/object/boreal:229877.

*Michigan Corpus of Upper-level Student Paper*s. (2009). Ann Arbor, MI: The Regents of the University of Michigan. Available at https://micusp.elicorpora.info/main.

Nesi, H. & Gardner, S (2012) *Genres across the Disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

Staples, S., & Dilger, B. (2018-). *Corpus and repository of writing [Learner corpus articulated with repository]*. Available at https://crow.corporaproject.org.

# LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English

Aivars Glaznieks[1], Jennifer-Carmen Frey[2], Maria Stopfner[3], Lorenzo Zanasi[4], Lionel Nicolas[5]
Institute for Applied Linguistics, Eurac Research
{aivars.glaznieks[1], jennifer.frey[2], maria.stopfner[3], lorenzo.zanasi[4], lionel.nicolas[5]}@eurac.edu

This corpus demonstration introduces the recently created longitudinal corpus of young learners of Italian, German and English, called LEONIDE (Glaznieks et al. 2022). The corpus contains 2,512 texts from 163 pupils, who participated in the project "One school, many languages" conducted in eight schools in the multilingual Italian province of South Tyrol (Engel & Stopfner 2018). The aim of the project was to document the development of the pupils' plurilingual competences by collecting oral and written language samples in three languages, to capture a holistic view of their individual linguistic repertoire.

LEONIDE is a collection of the written texts collected in the project over the span of 3 consecutive years (2015-2018) in public lower secondary schools (grade 6 to 8). The pupils were 11 years old at the beginning of the data collection and 13 years in the end. In each grade, two written tasks with different genres were given: the first was a picture story re-telling task; the second elicited an opinion text on aspects related to the pupils' life and public discourse. For each genre and each grade, the corpus provides texts in three languages German, Italian and English. In order to reflect the school system of the Province of South Tyrol, about half of the texts were collected in four schools in which German is the main language of instruction and Italian is taught as L2. The other half of the texts were collected in four schools in which Italian is the main language of instruction and German is taught as L2. In all schools, English is taught as L3 (i.e., as a foreign language at school). The overall size of the corpus amounts to ca. 237,000 tokens with the three sub-sections of 844 Italian texts (93,300 tokens), 833 German texts (73,900 tokens), and 835 English texts (69,700 tokens). Furthermore, a series of relevant learner-related data was collected for each learner, providing information about, e.g., age, gender, and first language(s). Learner-related metadata comes along with text-related metadata (e.g., task-type, the language of the text, year of text production), administrative information (e.g., version, license information), and information about corpus design (e.g., target languages, corpus size, study level, place of data collection). Each text has been manually annotated to reflect structural features (e.g., lines and paragraphs), orthographic errors (adding the correct spelling as the target hypothesis), choice of linguistic means (e.g., foreign words that do not belong to the target language), legibility of handwriting, pupils' self-corrections (i.e., deletion and insertions of letters or words), use of stylistic means (e.g., fully capitalized words, symbols) and anonymized text parts. In addition, each text has been automatically enriched with lemma and part-of-speech information using a UD tagger to facilitate comparisons over the three languages.

LEONIDE is unique in that it compiles texts by the same writers in three languages collected over a period of three years (true longitudinal data) while for instance, the ICCI corpus (Tono & Díez-Bedmar 2014) assembles only English texts from different writers (and regions) of various grades (cross-sectional data). Compared to other multilingual corpora, e.g., TRAWL (Dirdal et al. 2017), a longitudinal learner corpus of L1 Norwegian learners of English, French, German and Spanish, or SWIKO (Karges et al. 2019), a trilingual corpus of young Swiss learners of German, French, and English, LEONIDE represents monolingual as well as plurilingual learners who live in a multilingual region.

As LEONIDE documents the development of plurilingual competences of individual learners, it allows for contrastive longitudinal research on the development of young learners' writing skills in different languages, also considering person-related metadata. Moreover, the corpus is a valuable resource for language teachers to create and improve their teaching material and language courses as a large amount of authentic and longitudinal data reflects their difficulties and progress of language skills over three consecutive years in three languages. The corpus demonstration will give an overview of the main features of the corpus, show sample queries on the openly accessible ANNIS search interface, and guide the audience through the Eurac Research Clarin Centre repository (http://hdl.handle.net/20.500.12124/25) on which all relevant data for further use of the corpus can be downloaded for free and used for research purposes (ACA-BY-NC-NORED 1.0).

**References**

Dirdal, H., Danbolt Drange, E.-M., Graedler, A.-L., Guldal, T.M., Hasund, I.K., Nacey, S.L. & Rørvik, S. (2017). Tracking Written Learner Language (TRAWL): A longitudinal corpus of Norwegian pupils' written texts in second/foreign languages. In *Book of Abstracts of the 4th Learner Corpus Research Conference – LCR 2017 (Bolzano/Bozen, 5-7 October 2017)*, 182-183.

Engel, D. & Stopfner, M. (2019). Communicative competence in the context of increasing diversity in South Tyrolean schools. In E. Vetter & U. Jessner (Eds.). *International research on multilingualism: Breaking with the monolingual perspective*. Cham: Springer Nature, 59-80.

Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L. & Nicolas, L. (2022). Leonide. A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, 8(1), 97-120.

Karges, K., Studer, T. & Wiedenkeller, E. (2019). On the way to a new multilingual learner corpus of foreign language learning in school: Observations about task variations. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (Eds.). *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Louvain: Presses universitaires de Louvain, 137-165.

Tono, Y. & Díez-Bedmar, M.B. (2014). Focus on Learner writing at the beginning and intermediate stages: The ICCI corpus. *International Journal of Corpus Linguistics*, 19(2), 163-177.

## Demonstration of the CEDEL2 (version 2) interface: A multi-L1 corpus of L2 Spanish

Cristóbal Lozano[1], Nobuo Ignacio López-Sako[2]
University of Granada[1], University of Granada[2]
cristoballozano@ugr.es[1], nilsako@ugr.es[2]

Learner corpora (LC) are large, systematic databases of authentic language produced naturalistically by learners of a second language (L2) (Callies & Paquot 2015; Granger et al. 2015; Le Bruyn & Paquot 2021; Tracy-Ventura & Paquot 2021). LC are designed to cater to the specific needs of Second Language Acquisition (SLA) researchers, natural language processing scientists, foreign language learners/teachers, as well as materials designers (Díaz-Negrillo & Thompson 2013). Traditionally, most LC have targeted L2 English, but over the past years, L2 Spanish research has seen an increase, thus triggering the creation of large written and spoken L2 Spanish corpora (Lozano 2021b), such as CAES (Rojo & Palacios 2016), SPLLOC (Mitchell et al. 2008) and LANGSNAP (Tracy-Ventura et al. 2016).

We will showcase CEDEL2 (version 2): Corpus de Español como L2 (Lozano 2021a), a state-of-the-art L2 Spanish corpus that has been specially designed following Sinclair's (2005) ten corpus-design principles and the latest LC recommendations (Tracy-Ventura et al. 2021). CEDEL2 is a multi-L1 corpus of L2 Spanish with learners from typologically (un)related languages (English, German, Dutch, Portuguese, Italian, French, Greek, Russian, Arabic, Chinese, and Japanese), coming from all proficiency levels, diverse learning environments (instructed/naturalistic) and different countries. It currently holds language data from 4,399 participants (1,105,936 words) and data collection for a future version is still ongoing. It is mainly a written corpus though there are samples of spoken language (audios & transcriptions) as well. CEDEL2 also contains several native-control subcorpora for comparative purposes. All the learner and native subcorpora have been designed following the same principles and criteria so that full Contrastive Interlanguage Analysis (Granger 2015) can be carried out. Finally, CEDEL2 contains large amounts of SLA-motivated metadata (i.e., detailed information about the variables belonging to each speaker and each text) that allow to test key aspects in SLA, e.g.: L1 (cross-linguistic influence); proficiency level via a placement test (developmental effects); the age of onset to L2 Spanish (critical period and age effects); length of exposure to the L2 (exposure effects); length of residence in a Spanish-speaking country (effects of immersion in naturalistic settings); knowledge and proficiency in other foreign languages (other possible cross-linguistic influence); type of task and task conditions (task effects); etc.

We will do a software demonstration of CEDEL2's web-based search engine which, following the latest trends in Open Science, is freely available and downloadable at http://cedel2.learnercorpora.com. In particular, the following functionalities will be shown:

(i) The corpus can be either interrogated directly via the web interface (concordances, frequencies) or downloaded (texts with(out) metadata in several formats (txt and csv)) for additional analysis.

(ii) The search and download engine includes simple searches (strings of characters with(out) wildcards) and complex searches, whose results (output) can be of four types: typical concordances (KWIC), simple frequencies (words per million), complex frequencies (breakdown according to metadata), and texts (i.e., a tabulated list of files and corresponding variables). Several sorting options are available.

(iii) The searches can also produce different result subtypes: words, word categories, proximal words, and proximal word categories. In particular, the corpus has been automatically tagged with FreeLing, and the interface allows to search for lemmas and also POS (parts-of-speech) via an intuitive, drop-down menu containing word (sub)categories.

(iv) Filtering criteria can be applied to the searches/downloads. Filters are based on a set of key SLA-motivated metadata: L1, medium (spoken/written/spoken&written by the same person), sex, proficiency level category (lower/upper beginner, lower/upper intermediate, lower/upper advanced), placement test numeric score, learner's self-assessed proficiency on the four linguistic skills, task title (14 titles), filename, age of exposure to Spanish, years studying Spanish, and stay in Spanish speaking country. Filters permit targeting those elements (learners, concordances, texts, lemmas, etc.) that meet the user's desired criteria.

These features of CEDEL2 corpus and its web interface are ultimately meant to meet the needs of a wide range of users (SLA/LCR researchers, natural language processing scientists, language-teaching practitioners, and materials designers).

**References**


Callies, M., & Paquot, M. (2015). Learner Corpus Research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research*, 1(1), 1-6. https://doi.org/10.1075/ijlcr.1.1.00edi

CEDEL2 (Corpus Escrito del Español como L2), version 2: http://cedel2.learnercorpora.com

Díaz-Negrillo, A., & Thompson, P. (2013). Learner corpora: Looking towards the future. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.). *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, 9-29. https://doi.org/10.1075/scl.59.03dia

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP. https://doi.org/10.1017/CBO9781139649414

Le Bruyn, B., & Paquot, M. (Eds.). (2021). *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: CUP. https://doi.org/10.1017/9781108674577

Lozano, C. (2021a). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, 02676583211050522. https://doi.org/10.1177/02676583211050522

Lozano, C. (2021b). Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2. In M. Cruz Piñol (Ed.). *E-Research y español LE/L2: Investigar en la era digital*. New York: Routledge, 138-163. http://doi.org/10.4324/9780429433528-9

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.). *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, 65-100. https://doi.org/10.1075/scl.59.06loz

Mitchell, R., Domínguez, L., Arche, M., Myles, F., & Marsden, E. (2008). SPLLOC: A new database for Spanish second language acquisition research. In L. Roberts, F. Myles, & A. David (Eds.). *EUROSLA Yearbook 8*. Amsterdam: John Benjamins, 287-304. https://doi.org/10.1075/eurosla.8.15smit

Rojo, G., & Palacios Martínez, I. (2016). Learner Spanish on computer: The CAES 'Corpus de Aprendices de Español' project. In M. Alonso Ramos (Ed.). *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam: John Benjamins, 55-87. https://doi.org/10.1075/scl.78.03roj

Sinclair, J. (2005). How to build a corpus. In M. Wynne (Ed.). *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow books, 79-83.

Tracy-Ventura, N., & Paquot, M. (Eds.). (2021). *The Routledge Handbook of SLA and Corpora*. New York: Routledge. https://doi.org/10.4324/9781351137904

Tracy-Ventura, N., Mitchell, R., & McManus, K. (2016). The LANGSNAP longitudinal learner corpus: Design and use. In M. Alonso-Ramos (Ed.). *Spanish Learner Corpus Research: State of the Art and Perspectives*. Amsterdam: John Benjamins, 117-142. https://doi.org/10.1075/scl.78.05tra

Tracy-Ventura, N., Paquot, M., & Myles, F. (2021). The future of corpora in SLA. In N. Tracy-Ventura & M. Paquot (Eds.). *The Routledge Handbook of Second Language Acquisition and Corpora*. New York: Routledge, 409-424. https://doi.org/10.4324/9781351137904

# The CELI corpus: A new resource to analyse Italian L2

Stefania Spina, Irene Fioravanti, Luciana Forti, Francesca Malagnini, Angela Scerra, Valentino Santucci, Fabio Zanda
Università per Stranieri di Perugia
{stefania.spina, irene.fioravanti, luciana.forti, francesca.malagnini, angela.scerra, valentino.santucci, fabio.zanda}@unistrapg.it

In this demonstration, we present the main features of the CELI corpus, a new learner corpus designed to analyse Italian L2. The main novelty of this corpus is reflected in at least two aspects:

1. it is based on a balanced pseudo-longitudinal design;
2. the CEFR-level attribution of the texts derives from obtained language certification exams.

While the corpus-based analysis of Italian L2 can benefit from a number of learner corpora which have been built over the years (Gallina 2015; Corino et al. 2017; Bratankova 2015; Spina & Siyanova Chanturia 2018; Bailini & Frigerio 2018; Wisniewski et al. 2013; Glaznieks et al. 2022), pseudo-longitudinal analyses reflecting progression between proficiency levels, using largely written corpora of Italian language certification tests, is still an underexplored area.

The CELI corpus derives its name from the CELI (*Certificati di Lingua Italiana)* exams, Italian language certification exams administered by the CVCL – *Centro per la Valutazione e le Certificazioni Linguistiche* at the University for Foreigners of Perugia. The corpus contains the written texts produced under examination conditions by candidates of proficiency levels B1, B2, C1, and C2. Only texts written by candidates who passed the certification exam were included in the corpus. Overall, the corpus consists of 3,041 texts, reaching a total of 608,614 tokens, which are evenly distributed among the four proficiency levels, and 24,698 types. The texts were produced on paper and were thus manually transcribed in digital form. Each text contains, on average, 200 tokens. The metadata inserted in the corpus includes those that are systematically collected by CVCL. These are gender, age, nationality, exam score related to the overall test, exam score related to the written test, the type of writing task on the basis of which the text was produced, exam score related to the writing task, and analytic scores assigned to lexical, grammatical sociolinguistic and textual competences. The most frequent nationalities represented in the corpus are Greek, Spanish, Swiss, Romanian, and Albanian. The text genres are letter, e-mail, blog entry, story, article, and essay, whereas the text types are classified into argumentative, descriptive, and narrative, as well as mixed typologies (descriptive-narrative, argumentative-narrative, argumentative-descriptive, argumentative-narrative-descriptive).

The transcribed texts were lemmatised and pos-tagged running a pre-trained version of *TreeTagger* (Schmid 1994) on the learner texts. The choice of automatically annotating learner data has been adopted for many other learner corpora, such as ICLE (Granger et al. 2020), MERLIN (Wisniewski et al. 2013), LEONIDE (Glaznieks et al. 2022), and CEDEL2 (Lozano 2021). The pos-tagging was performed using a tagset consisting of 54 tags, which had been developed previously for the annotation of the *Perugia corpus* (Spina 2014). The grammatical category exhibiting the highest degree of internal differentiation is that of verbs, with a total of 23 separate tags. Furthermore, adjectives are differentiated in terms of whether they are qualifying, possessive, indefinite, and demonstrative. Another key feature characterising the annotation of the corpus is that adverbial multiword expressions are tagged as single lemmas. It is the case for *un po'* ('a bit'), *a galla* ('afloat'), *a fondo* ('in depth'). An additional phase of semiautomatic processing of the texts was conducted in order to eliminate the errors of the tagger, related mostly to forms with a high degree of grammatical ambiguity (e.g. *come*, 'how/as/like'; *che*, 'which/that'; *dove*, 'where/which/that') or forms that were not recognised by the tagger, which consequently was unable to assign them to a specific lemma. A final phase of manual processing was needed to remove all remaining forms that were unknown to the tagger. The various post-tagging phases allowed the correction of the tagging errors, which amounted to about 1% of the total forms that were tagged in the corpus, thus increasing the overall annotation accuracy of the corpus.

The potential uses of the CELI corpus are multifold. From a research perspective, the pseudo-longitudinal design of the corpus provides a sound empirical foundation to investigate non-linearity in second language acquisition, which is one of the main tenets of complexity theory (Larsen-Freeman & Cameron 2009). The CELI corpus can also be usefully employed in *Contrastive Interlanguage Analysis* (CIA) studies (Granger 1996; 2015), based on systematic comparisons between varieties of learner language and varieties of L1 language.

From a pedagogical perspective, it can help language teachers and material developers to identify learner difficulties at different levels of proficiency, thus informing the structuring of a curriculum. Furthermore, the corpus can be used directly with the learners, to guide them through the observation of learner patterns at different levels of proficiency, in contrast with L1 speaker patterns, thus fostering metalinguistic awareness (Ackerley 2013).

**References**


Ackerley, K. (2013). A comparison of learner and native speaker writing in online self-presentations: Pedagogical applications. In S. Granger, G. Gilquin, F. Meunier (Eds.). *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Louvain-la-Neuve: Presses Universitaires de Louvain, 1-10.

Bailini, S., Frigerio, A. (2018). CORESPI e CORITE, due nuovi strumenti per l'analisi dell'interlingua di lingue affini. *CHIMERA: Romance Corpora and Linguistic Studies*, 5(2), 313-319.

Boyd, A., Jirka, H., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Stindlová, B., & Chiara Vettori (2014). The MERLIN corpus: Learner Language and the CEFR. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1281-1288.

Bratankova, L. (2015). *Le collocazioni Verbo + Nome in apprendenti di italiano L2*. Tesi di dottorato, Università per Stranieri di Perugia.

Corino, E., Colombo, S., Marello, C. (2017). *Italiano di stranieri: i corpora VALICO e VINCA*. Perugia: Guerra.

Gallina, F. (2015). *Le parole degli stranieri. Il Lessico Italiano Parlato da Stranieri*. Perugia: Guerra.

Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, 8(1), 97-120.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, M. Johansson (Eds.). *Languages in Contrast. Text-based Cross-linguistic Studies*. Lund: Lund University Press, 37-51.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24.

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English. Version 3*. Louvain-la-Neuve: Presses universitaires de Louvain.

Lozano, C. (2021). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, United Kingdom.

Spina, S. (2014). Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In R. Basili, A. Lenci, B. Magnini (Eds.). *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it (Vol. 1*). Pisa: Pisa University Press, 354-359.

Spina, S., Siyanova-Chanturia, A. (2018). The Longitudinal Corpus of Chinese Learners of Italian (LOCCLI). *Poster presented at 13th Teaching and Language Corpora conference*. Cambridge: University of Cambridge, United Kingdom.

Wisniewski, K., Schöne, K., Nicolas, L., Vettori, C., Boyd, A., Meurers, D., Abel A., Hana, J. (2013). MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data. In *ICT for Language Learning 2013, Conference Proceedings*. Firenze: Libreriauniversitaria.it. Edizioni, Firenze, 14-15 novembre 2013.

# Swedish L2 profile – A tool for exploring L2 data

Elena Volodina[1], Therese Lindström Tiedemann[2], Yousuf Ali Mohammed[3]
University of Gothenburg[1,3], University of Helsinki[2]
elena.volodina@svenska.gu.se[1], therese.lindstromtiedemann@helsinki.fi[2],
yousuf.ali.mohammed@svenska.gu.se[3]

Learner corpus researchers, NLP researchers, as well as Digital Humanities and Social Sciences in general, rely on access to various data sets for empirical analysis, statistical insights, and/or for model building. However, interpretation of data is a non-trivial task and there is a need for data visualization tools. One such attempt is the Swedish L2 profile (*SweL2P*) – an ongoing project setting up the first digital tool allowing users to explore written Swedish learner language from a linguistic point of view.

The SweL2P is based on data from two corpora: course books (*receptive* data; Volodina et al. 2014) and learner essays (*productive* data; Volodina et al. 2016). The two corpora have been semi-automatically parsed for verb and noun patterns that currently constitute the core of the *grammatical profile*. Both corpora have also been used to create a word list, *Sen*Lex,* as the main input for the *lexical profile*. Sen*Lex has subsequently been manually enriched with morphological analysis giving rise to the CoDeRooMor resource (Volodina et al. 2021) that has been used as the main input for the *morphological profile*.

The SweL2P features

- *a lexical profile*, organized by words, multi-word expressions, and a few other aspects of vocabulary
- *a grammatical profile*, including noun phrases and verb phrases
- *a morphological profile*, organized into word families and morpheme families

Each item or pattern in the profile can be *filtered* in various ways depending on the category in focus and explored through actual corpus **hits in Korp** (corpus management system; Ahlberg et al. 2013; Borin et al. 2012). Filters appear at the top of the page, providing for each individual sub-profile an individual set of filters. The resource can be explored using several views. (1) The **Table view** lists all items with associated information about each of them. Columns contain descriptive information, such as a clickable **category** (e.g. verb pattern with a clickable link leading to an explanation of the pattern), and clickable **receptive** and **productive** (relative and absolute) frequencies that open a corpus search tool containing hits with those lemgrams. (2) **Graphical view** summarizes the statistics and distribution of various features for the current selection in the two sources – receptive and productive – in graphs with related tables with statistics. (3) In the **Statistical view** we see counts in terms of types, tokens, and type-token ratios per filter category so that we can study the statistical breakdown of each selection contrasting receptive and productive competencies. The entire dataset or filtered data selection can be downloaded.

To conclude, language learning profile resources exist predominantly for English, e.g. *English profile* (O'Keeffe and Mark 2017) and *Pearson's GSE Teacher Toolkit*. Most other languages have nothing similar, the Estonian profile (Üksik et al. 2021) being one of the first non-English profiles. The existing profiles focus mainly on teachers and learners. Even though they have been based on empirical corpus data, this data is not openly provided in connection to the resource, rendering them rather prescriptive. The L2 Swedish profile takes a non-prescriptive view of the language and provides access to the empirical evidence, i.e. all corpus hits and statistics of actual usage. It lets users zoom in on actual data and draw their own conclusions. The provision of both receptive *and* productive frequencies gives a more nuanced picture of language learning. Due to that, and due to the special efforts invested into the visualization of the data, we believe that the SweL2P tool is more readily appropriate for research on second language acquisition than any predecessor known to us. In addition, the open nature of the resource makes it highly usable for future learning apps, for the training of automatic tools, and for teaching.

**References**


Ahlberg, M., Borin, L. Forsberg, M. Hammarstedt, M., Olsson, L-J. Olsson, O., Roxendal, J. & Uppström, J. (2013). Korp and Karp – a bestiary of language resources: the research infrastructure of Språkbanken. In S. Oepen, K. Hagen & J. Bonde Johannessen (Eds.). P*roceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16 (16), 429–433.

Borin, L., Forsberg, M. & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis. (Eds.). *Proceedings of LREC 2012*. Istanbul: ELRA, 474–478.

O'Keeffe, A., & Mark, G. (2017). The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics* 22 (4), 457–489.

Üksik, T., Kallas, J., Koppel, K., Tsepelina, K. & Pool, R. (2021). Estonian as a Second Language Teacher's Tools. In J. Burstein, A. Horbach, E. Kochmar, R. Laarmann-Quante, C. Leacock, N. Madnani, I. Pilán, H.

Yannakoudakis & T. Zesch (Eds.) *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 130–134.

Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G. & Sandell, M. (2016). SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M, Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.). *Proceedings of LREC 2016*, Slovenia. European Language Resources Association (ELRA), 206–212. https://arxiv.org/pdf/1604.06583v1.pdf

Volodina, E., Pilán, I., Rødven-Eide, S. & Heidarsson, H. (2014). You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. *Proceedings of the third workshop on NLP for computer-assisted language learning*. Linköping Electronic Conference Proceedings 107 (10), NEALT Proceedings series 22 (10), 128–144. https://ep.liu.se/ecp/107/010/ecp14107010.pdf

Volodina, E., Mohammed, Y. A. & Lindström Tiedemann, T. (2021). CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish. In S. Dobnik & L. Øvrelid (Eds.). *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa).* Linköping Electronic Conference Proceedings 178 (18), 178–189. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=178&Article_No=18