Book of abstracts
LCR2024 Tartu

7th INTERNATIONAL LEARNER CORPUS RESEARCH CONFERENCE 2024

UNIVERSITY OF TARTU
Institute of Foreign Languages and Cultures

Learner Corpus Association
www.learnercorpusassociation.org

JOHN BENJAMINS PUBLISHING COMPANY

Table of contents

Empirically testing the "Trade-Off Hypothesis": A machine learning experiment 9

Akef, Soroosh; Detmar Meurers, Amália Mendes, Patrick Rebuschat 9

Thematic progression in written argumentative paragraphs of Chinese advanced learners of French 10

Aleksandrova, Tatiana 10

ELLE Text Evaluator – a corpus-based tool for learning and teaching Estonian 11

Allkivi-Metsoja, Kais; Taavi Kamarik, Karina Kert and Silvia Maine 11

To Collocate or not to Collocate: Exploring Verb-Noun Collocations of Turkish EFL Learners 12

Aybek, Sibel 12

The Polish Learner Corpus "FoKo" - main project assumptions 13

Badyda, Ewa and Lucyna Warda-Radys 13

Exploring learner knowledge with Large Language Models fine-tuned with the EFCAMDAT 14

Ballier, Nicolas and Bernardo Stearns 14

Russian Spoken Learner Corpus: Designing and collecting a spoken corpus of Italian university learners of Russian 15

Bejenari, Oxana; Paola Cotta Ramusino, Claudio Macagno, and Tatsiana Maiko 15

Complexity or complexities? A simulation study on lexical complexity in expert and learner texts through the lens of information theory 16

Brasolin, Paolo and Arianna Bienati 16

Levels and Modeling of Variability in L2 Learner Corpora. Insights from a Corpus of Newly Migrated Students in Germany 17

Braunewell, Aylin; Julia Schlauch and Jana Gamper 17

Teaching interpersonal devices with English learner corpora 18

Carrió-Pastor, María Luisa 18

EFL students' use of phrasal verbs of action and motion in spoken and written video-clip descriptions 19

Castello, Erik and Katherine Ackerley 19

Learner corpus genre analysis: What can it reveal about learners' discoursal weaknesses? 20

Charles, Maggie and Karin Whiteside 20

A multidimensional comparison of argumentative writing of English learners from different language families 21

Chen, Mei-Hua, Wei-Fan Chen, Garima Mudgal, and Henning Wachsmuth 21

CREATING A MULTIMODAL CORPUS OF INTERCOMPREHENSION: first reflections on annotation 22

Corino, Elisa and Eugenio Goria 22

Getting down to business with the Apprentice Multiple Business GEnRes (AMBER) corpus 23

De Cock, Sylvie and Jennifer Thewissen 23

Introducing the CLARIN Knowledge Centre for Learner Corpora 24

De Cock, Sylvie; Damien De Meyere, Liesbeth Degand, Thomas François, Gaëtanelle Gilquin, Sylviane Granger, Danqing Huang, Marie-Aude Lefer, Hubert Naets, Magali Paquot, Kristel Van Goethem, Patrick Watrin with the collaboration of Jennifer Carmen-Frey, Alexander König and Egon Stemle 24

What a L1-German Learner Corpus can tell us about the acquisition of Romance Languages: The MuLeCo Project 25

De Crignis, Patricia and Johanna Wolf 25

Tracking phraseological inter-complexity: a corpus based contrastive analysis of learner English and Romanian novice academic writing 26

Dinca, Andreea; Ana-Maria Bucur and Madalina Chitez 26

Syntactic complexity across L1, L2 and L3 27

Dirdal, Hildegunn; Stine H. Johansen and Philip Durrant 27

Interacting to learn: Towards an inclusive corpus of EFL learners' interactions with AI-based chatbots 28

Ferraresi, Adriano and Silvia Bernardini 28

Syntactic and semantic features of concession in high school students' writing: a corpus analysis of Italian 29

Ferrato, Elena 29

DMorphC: A web tool for carrying out the identification, grouping and counting of morphologically complex words. 30

Flores Hernández, Ana Abigahil; Hilda Hidalgo Avilés and Abigail Carretero Hernández 30

The role of learner corpora in needs analysis for ESP (English for Specific Purposes) 31

Flowerdew, Lynne 31

Assessing the validity of new structural complexity metrics as features of L2 proficiency 32

Gaillat, Thomas 32

Exploring Verb-Noun collocations across L2 English proficiency levels in L1 Mandarin and French learners 33

Gaillat, Thomas and Jen-Yu Li 33

Understanding Foundation-level ESL Students' Academic Writing: A Lexico-grammatical Analysis across Genres and Disciplines 34

Gao, Mingyan 34

Structural Characteristics of L3 Norwegian as Spoken by Polish Native Speakers 35

Garbacz, Piotr 35

Productive use of derivational affixes across proficiency levels 36

Gee, Roger W., M. Karen Jogan, Kathleen S. Jogan 36

Cambridge-Belgrade (CamBel) Persian Learner Corpus: Design and Methodology 37

Ghaffari, Mahbod and Saeed Safari 37

DiverSIta: a new corpus for the documentation of L2 Italian 39

Goria, Eugenio and Caterina Mauri 39

Syntactic Complexity Development in Intermediate learner English: A longitudinal pilot study 40

Götz-Lehmann, Sandra; Philine Metzger and Fabian Kettenhofen 40

Introducing ETC, the first learner corpus of non-native teacher English 41

Gráf, Tomáš, Barbora Bulantová and Kryštof Buchal 41

Opening Pandora's box: Is learner corpus research inclusive enough? 42

Gilquin, Gaëtanelle 42

Of-constructions in the interlanguage of Norwegian learners of English. A pseudo-longitudinal study. 43

Hasselgård, Hilde 43

The impact of topic on the use of lexical bundles by EFL and ESL learners 44

Huang, Lingmin 44

Collocations with the verbs have, make/do, give and get in L2 Czech: how corpus research can inform CEFR descriptions of Czech at levels A2-B2 45

Hudousková, Andrea 45

Considerations for planning and developing publicly-shared L2 speech corpora 46

Huensch, Amanda 46

Cross-linguistic influences in different levels of granularity: How are they different and why does it matter for LCR? 47

Ivaska, Ilmari 47

Comparing morphosyntactic features in undergraduate dissertations in Spanish by Estonian and Spanish students: a corpus-driven approach 49

Izquierdo Alegría, Dámaso 49

A Contrastive Analysis of Lithuanian Children's Language 50

Juknevičienė, Rita 50

The effect of self-initiated L2 activities on intermediate-level students' lexical complexity 51

Kaatari, Henrik; Tove Larsson, Ying Wang, Pia Sundqvist, Taehyeong Kim 51

Estonian L2 Learner Corpora: current state and perspectives 52

Kallas, Jelena; Kristjan Suluste, Raili Pool, Helen Kaljumäe 52

Pedagogical applications of learner corpus research: How far have we come? 53

Karlsen, Petter Hagen and Susan Nacey 53

Subject-Verb Agreement in English Learner Texts: A Pseudo-Longitudinal Perspective 54

Karlsen Petter Hagen and Sylvi Rørvik 54

Creation and exploration perspectives of ScientEst, an Estonian Learner corpus of French Academic Discourse 55

Käsper, Marge and Anu Treikelder 55

Design, measurement, and analysis in longitudinal corpus-based SLA research: A systematic review 56

Kim, Minjin and Kevin McManus 56

Lexical complexity indices as markers of proficiency in L2 Russian 57

Kisselev, Olesya V., Mikhail Kopotev and Anton Vakhranev 57

To what extent do P-bursts resemble lexicogrammatically meaningful chunks? An exploratory study on beginner level foreign language writing processes 58

Kruse, Mari 58

KOST, the first learner corpus for Slovene as a second language 59

Kučuk Stritar, Mojca 59

Exploring Noun Lexical Diversity and Noun Phrase Complexity: A Learner Corpus-Based Study of B1 and C1 Spanish EFL Learners' Email Writing 60

Laso Martín, Natalia Judith and María Belén Díez Bedmar 60

Wordless: An integrated corpus tool with multilingual support for the study of language acquisition, pedagogy, and assessment 61

Lei, Ye 61

The acquisition of the article system by Polish advanced learners of English: evidence from legal translations 62

Leńko-Szymańska, Agnieszka; Łucja Biel and Katarzyna Wasilewska 62

A metadata scheme for cross-corpus analyses of L2 acquisition 63

Lenort, Lisa; Annette Portmann, Matthias Schwendemann, Josef Ruppenhofer 63

Bridging academic and technological domains. The new framework for developing the Estonian L1 and L2 preschool children's speech corpus. 64

Lilles, Kelly 64

Construction, transcription and annotation of a longitudinal multimodal interlanguage corpus: An ongoing study with L1 Italian and L2 Chinese 65

Liu, Siyuan 65

A Diasystematic Construction Grammar (DCxG) analysis of Progressive Aspectuality in Multilingual Learners of English as Additional Language 66

Lopopolo, Olga 66

Discussing the categorization of speakers' language background: implicit assumptions and methodological challenges for Learner Corpus Research 67

Lopopolo, Olga; Arianna Bienati, Jennifer-Carmen Frey, Aivars Glaznieks and Stefania Spina 67

Exploring Formulaic Language in Student Academic Writing in L1 and L2 German 68

Lösel, Andrea 68

Quoting and Referencing Mastery: ExpoKo's Insights Into Student Academic Writing Challenges 69

Lösel, Andrea; Matthias Schwendemann and Franziska Wallner 69

Designing and compiling a multi-L1 corpus of L2 Spanish: Lessons learnt from CEDEL2 70

Lozano, Cristóbal 70

Georgian English learner corpus and its lexicographic applicability 71

Makhatadze, Marine 71

Constructing a Spoken Corpus of Cochlear Implant Patients/Users 72

Mazaherylaghab, Hamzeh 72

Navigating the complexity of causality annotation in learner language 73

Miki, Nozomi and Akira Murakami 73

Proficiency-dependent factors influencing L2 dative alternation 74

Murakami, Akira; Masato Terai, Yu Tamura and Junya Fukuta 74

Building the young learner corpus - A longitudinal, bilingual (English-Indonesian) corpus: Challenges of collecting data from different educational settings. 75

Mustun, Senora; Mauselyn Pattikawa, Yosef Tanggu Solo, Vonny Juliana Ruhulessin, Yeremina Karolina Ngabalin, Theresia Astrie Nunumete 75

A call for more data-oriented reporting in Learner Corpus research 76

Nicolas, Lionel, Egon W. Stemle, Magali Paquot, Hubert Naets and Alexander König) 76

SEEFLEX – The Corpus of Secondary School English as a Foreign Language (EFL) Exams 77

Pauls, Tobias 77

The effect of different feedback strategies on lexical sophistication of ESP writing 78

Pojslová, Blanka 78

Capitalisation and consonant doubling in German as a foreign language – an error analysis of learner texts at different CEFR proficiency levels 79

Reitbrecht Sandra 79

Verb phrase versatility as a syntactic complexity indicator in L2-English written texts 80

Reményi, Andrea Ágnes 80

The development of complexity, accuracy and fluency in L2 English writing at secondary school–group and individual learning profiles. 81

Rokoszewska, Katarzyna 81

Syntactic-complexity development in Norwegian learner English 82

Rørvik, Sylvi 82

Automatic generation of target hypotheses for learner language 83

Ruppenhofer, Josef, Torsten Zesch, Katrin Wisniewski and Anette Portmann 83

The Use of Punctuation in a German-to-Basque LTC 84

Sanz Villar, Zuriñe 84

Compiling a multi-corpus database for automatically annotating developmental stages in L2 German 85

Schwendemann, Matthias, Katrin Wisniewski, Torsten Zesch and Lisa Lenort 85

Motion Verbs in the Second/Foreign Language Acquisition of Czech: a corpus-based study on non-native speakers of Czech with Chinese L1 86

Škodová, Svatava and Melissa Shih-hui Lin 86

Triangulation with learner translation corpora 87

Skogmo, Siri Fürst and Susan Nacey 87

Intensification of adjectives in young L2 learners of German and Italian 88

Spina, Stefania, Aivars Glaznieks and Andrea Abel 88

Reporting verbs in Norwegian undergraduate learner English 89

Thormodsæter, Øyvind 89

The use of English articles in essays written as part of the Estonian National Examination of English 90

Compiling oral learner corpora: Is automatic transcription really worth it? 91

Vandeweerd, Nathan 91

Using crowdsourced comparative judgement and rubric-based rating to grade texts in the ICLE corpus: a report on reliability and validity 92

Vandeweerd, Nathan, Peter Thwaites, Magali Paquot and Jiacheng Shen 92

Writing argumentative essays: Discourse strategies in L1- and L2-authored versus ChatGPT-generated text 93

Wan, Shujun, Julián Moreno-Schneider and Georg Rehm 93

The influence of L1 Dutch on cohesion in L2 German academic writing: A contrastive corpus-based analysis 94

Wedig, Helena; Carola Strobl, Jim Ureel, Tanja Mortelmans 94

Frequency vs. accuracy in learner Englishes: A study on tense and aspect 95

Werner, Valentin and Robert Fuchs 95

The Corpus of Young German Learner English 96

Werner, Valentin, Robert Fuchs, Anna Rosen, Lea Bracke and Bethany Stoddard 96

L1 Influence on Chinese English Learners' Use of Individual Senses of "IN" 97

Xu, LingLing 97

Wordless: An integrated corpus tool with multilingual support for the study of language acquisition, pedagogy, and assessment 98

Ye, Lei 98

How does the first language influence the shape of text? A Comparison of Korean and Polish non-native Czech texts 99

Zasina, Adrian Jan 99

Empirically testing the “Trade-Off Hypothesis”: A machine learning experiment

Akef, Soroosh (Center of Linguistics of the University of Lisbon (CLUL)/LEAD Graduate School and Research Network, University of Tübingen), Detmar Meurers (Leibniz Institute für Wissensmedien (IWM), University of Tübingen/LEAD Graduate School and Research Network, University of Tübingen), Amália Mendes (Center of Linguistics of the University of Lisbon) (CLUL), Patrick Rebuschat (Lancaster University/LEAD Graduate School and Research Network, University of Tübingen)

Language proficiency has long been conceptualized by the complexity, accuracy, fluency (CAF) triad (Skehan, 1991). While how each individual aspect characterizes proficiency has been extensively investigated (Hasnain & Halder, 2024), there is a gap in the literature for empirical evidence in support of Skehan’s “Trade-Off Hypothesis” (Skehan, 1998) or its rival hypothesis, the “Cognition Hypothesis” (Robinson, 1995), both of which discuss the interaction between these aspects.

This study addresses this gap by utilizing the Portuguese learner corpus COPLE2, including 1634 texts featuring 49 fine-grained types of grammatical, lexical, orthographical, and discourse error annotations. Machine learning is used as an experimental test bed to study which accuracy features (50 raw counts and 50 length-independent extracted from the corpus) and complexity features (489 automatically calculated using CTAP (Chen & Meurers, 2016)) contribute the most to the performance of a proficiency classifier.

Models were trained for the three-class classification (A, B, C on the CEFR scale) and the five-class classification task (A1-C1), correctly classifying 74.66% and 58.08% of the texts respectively (10-fold cross validation). An analysis of the most predictive features for the fine-grained model revealed the high importance of lexical sophistication, lexical richness, and morphological inflection features. Notably, the only accuracy feature among the top 10 most predictive features was total token accuracy rate (#errors/#tokens) with no other accuracy feature appearing in the top 200, suggesting that the types of errors learners make do not independently characterize proficiency but that a pattern can still be observed in learners’ tendency to make mistakes at different levels of proficiency.

This study demonstrates how machine learning can be utilized as an experimental tool to provide support for second language acquisition theories. To study how the interaction of accuracy and complexity measures and learners’ L1 characterize proficiency, statistical analyses using generalized additive models will be presented.

References

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 (CL4LC) (pp. 113–119). Osaka, Japan: The COLING 2016 Organizing Committee.

Hasnain, S., & Halder, S. (2024). Intricacies of the multifaceted triad-complexity, accuracy, and fluency: A review of studies on measures of oral production. Journal of Education, 204(1), 145-158. https://doi.org/10.1177/00220574221101377

Mendes, A., Antunes, S., Janssen, M., & Gonçalves, A. (2016). The COPLE2 corpus: A learner corpus for Portuguese. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-16) (pp. 3207–3214). Portorož, Slovenia: European Language Resources Association (ELRA).

Robinson, P. (1995). Attention, memory, and the “noticing” hypothesis. Language learning, 45(2), 283-331. https://doi.org/10.1111/j.1467-1770.1995.tb00441.x

Skehan, P. (1991). Individual Differences in Second Language Learning. Studies in Second Language Acquisition, 13(2), 275–298. https://doi.org/10.1017/S0272263100009979

Skehan, P. (1998). A cognitive approach to language learning. Oxford University Press.

Thematic progression in written argumentative paragraphs of Chinese advanced learners of French

Aleksandrova, Tatiana (Université Grenoble Alpes, France)

It is generally recognized that many second language (L2) learners, especially non-European students, have difficulties in writing coherent and cohesive texts and in using appropriate types of thematic progression (Hawes, 2015). According to our research, there are not many studies that examine the thematic progression for Chinese L2 learners of French. We would like to answer the following questions. Firstly, what types of thematic progression do native French speakers and Chinese advanced learners of French use in argumentative paragraphs? Secondly, are there some problems of cohesion and coherence in the productions of these L2 learners?

In our study, we collected a corpus of written productions from 22 Chinese learners at roughly B2 level. The corpus consists of short protest letters (250 words) addressed to the mayor of a city, giving cultural, touristic and economic arguments for not cancelling a concert. Their productions were contrasted to those of French native speakers (n=30) who had similar socio-cultural profiles. All participants were given the same instructions to write by hand under the same examination conditions.

The results show that learners principally use constant progression (46%), while French native speakers prefer linear progression (38% of cases). Indeed, linear progression is much less frequent in the productions of learners (18%). Concerning derived progression, it is used in only 2% of cases by French native speakers and it is entirely absent from the productions of learners. These results show some of the difficulties experienced by the learners in using linear progression in their productions and to privilege constant progression. It confirms the results of previous works on textual cohesion in the L2 productions of non-native students of English (Bloor & Bloor, 1992; Wang, 2007). We can conclude that Chinese L2 learners would benefit from further instruction in notions of thematization and thematic progression.

References

Bloor, M., and Bloor, T. (1992). Given and new information in the thematic organization of text: An application to the teaching of academic writing. Occasional Papers in Systemic Linguistics, 6, 33-43.

Hawes, T. (2015). Thematic progression in the writing of students and professionals. Ampersand, 2, 93-100.

Wang, L. (2007). Theme and rheme in the thematic organisation of text: Implications for teaching academic writing, Asian EFL Journal, 9 (1), 164-176.

ELLE Text Evaluator – a corpus-based tool for learning and teaching Estonian

Allkivi-Metsoja, Kais; Taavi Kamarik, Karina Kert and Silvia Maine (Tallinn University, School of Digital Technologies)

Estonian Language Learning and Analysis Environment (ELLE) combines an Estonian learner corpus with a toolkit allowing users to analyze the corpus resource as well as texts of their own choice. Its aim is to bring language learners, educators, and researchers together on one platform that supports both direct and indirect pedagogical application of corpus analysis. While benefiting from various linguistic tools and resources, learners can submit their writings to the corpus and facilitate further work on corpus-based research and study materials.

ELLE can be considered a renewed and enhanced user interface of the Estonian Interlanguage Corpus compiled since the 2000s. The ongoing development follows a prototype created by using participatory design methods, involving target users in the process (Norak & Põldoja, 2021). Our demonstration focuses on one of ELLE's central tools called Text Evaluator. The learner-oriented writing assistant and evaluator incorporates error correction and improvement suggestions, proficiency level assessment, and text complexity/readability analysis, which are rarely all found in one tool. Learner language data has been essential in developing and testing the application.

Firstly, Text Evaluator integrates the spelling and grammatical error correction models developed by the University of Tartu and Tallinn University (Allkivi-Metsoja & Kippar, 2023; Luhtaru et al., 2024). Errors are grouped by category. Secondly, text proficiency level is predicted on a scale of A2–C1. The assessment is based on supervised machine learning, considering lexical and grammatical complexity together with error frequency. Thirdly, the text is assessed on a scale of easy – intermediate – difficult based on complexity indices. Measures of lexical complexity are calculated separately. Long words and sentences, word repetitions, relatively rare words, abstract nouns, and content words are highlighted. Whereas useful for measuring the development of writing skills, the tool also helps correct and simplify professional texts or choose learning materials.

References

Allkivi-Metsoja, K., & Kippar, J. (2023). Spelling Correction for Estonian Learner Language. In T. Alumäe, & M. Fishel (Eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics*. *NEALT Proceedings Series, 52* (pp. 782−788). ACL Anthology

Luhtaru, A., Korotkova, E., & Fishel, M. (2024). No Error Left Behind: Multilingual Grammatical Error Correction with Pre-trained Translation Models. In Y. Graham, & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, *1* (pp. 1209–1222). ACL Anthology.

Norak, K., & Põldoja, H. (2021). Designing a Virtual Learning Environment Based on a Learner Language Corpus. In W. Zhou, & Y. Mu (Eds.), *Advances in Web-Based Learning – ICWL 2021. Lecture Notes in Computer Science*, *13103* (pp. 40−51). Springer. https://doi.org/10.1007/978-3-030-90785-3_4

To Collocate or not to Collocate: Exploring Verb-Noun Collocations of Turkish EFL Learners

Aybek, Sibel (Cukurova University)

Many studies in learner corpora argue that L2 learners have a strong need for the collocation knowledge to become proficient and fluent (Boers et al., 2006). Such mastery distinguishes advanced learners from intermediate ones (Thornbury, 2002). The importance of these multiword units considered as the "building blocks of discourse in spoken and written registers" (Biber & Barbieri, 2007, p. 263) has been repeatedly emphasized in previous studies (Sinclair, 1991; Granger 1998; Wray, 2002; Schmitt, 2004). Verb-noun collocations were found the most problematic for L2 learners (Nesselhauf, 2005; Peng, 2016) due to 'arbitrariness and unpredictability' of these constructions (Peng, 2016, p. 19). The present study investigates the use of English verb-noun collocations in the written productions of Turkish EFL learners at three proficiency levels A1-A2; B1-B2; C1-C2, as defined by the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). The corpus used in this particular study is the Cambridge Learner Corpus (CLC), the largest annotated test performance corpora which enables the investigation of the linguistic and rhetorical features of the learner performances in CEFR proficiency bands. After retrieving the most frequent nouns in the CLC, concordances have been created and the most frequent verb-noun collocations have been extracted based on frequency data and mutual information (MI) scores. Extracted collocations have been checked with the British National Corpus (BNC) corpus to determine the frequency bands and the appropriateness of the collocations used. Also, erroneous uses of these collocations have been analyzed. Both quantitative and qualitative analysis have been carried out in order to reveal the most salient collocations and their relative frequencies across three proficiency levels A1-A2; B1-B2; C1-C2 of Turkish EFL learners, examine and categorize the errors and to observe the improvement of collocation accuracy with the increasing proficiency level. Preliminary findings suggest that the frequency of collocations increases as the proficiency level increases and Turkish EFL learners make collocation errors such as appropriate verb choice, verb agreement and incorrect tense of verb. The results also suggest insights on learners' collocation usage and difficulties, and that the use of collocations is affected by the lack of L2 knowledge.

References

Biber, D., & Barbieri, F. (2007). Lexical Bundles in University spoken and written registers. *English for Specific Purposes, 26*, 263–286.

Boers, F., Eyckmans, J. Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. Language Teaching Research, 10(3), 245–261.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge, U.K: Press Syndicate of the University of Cambridge.

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: Clarendon Press.

Granger, S., Dagneaux, E. & Meunier, F. (2009). *The international corpus of learner English version 2. handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus.* Amsterdam/ Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1017/S0272263107270068

Peng, X. L. (2016). *Use of Verb-Noun Collocations by Advanced Learners of Chinese. Ph.D. thesis*, University of Pennsylvania.

Schmitt, N. (Ed.). (2004). Formulaic sequences: Acquisition, processing, and use. Amsterdam: John Benjamins Publishing Company.

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

Thornbury, S. (2002). How to teach vocabulary. London: Longman.

Wray A (2002). Formulaic language and the lexicon. Cambridge: Cambridge University Press.

The Polish Learner Corpus "FoKo" - main project assumptions

Badyda, Ewa and Lucyna Warda-Radys (University of Gdańsk)

The Polish Learner Corpus "FoKo", developed at the University of Gdańsk (Poland), gathers written works from individuals learning Polish as a foreign/second language, specifically students aged 10-19, who attend schools in the Pomeranian region of Poland. The corpus consists of their handwritten essays, created under the supervision of a teacher, on various assigned topics. Additionally, it includes speech bubble texts from comic drawings depicting everyday communicative situations. Each student's folder in the corpus also contains a questionnaire filled out by their teacher, providing sociolinguistic data and details about the language situation of the child before coming to Poland and during the linguistic study. So far, approximately 300 students' works have been collected, with the majority originating from Ukraine. The students' periods of stay in Poland and their levels of language proficiency in Polish vary significantly. Separate sub-corpora within "FoKo" include works collected under the same principles from Polish children studying in Polish schools located abroad and those who started their school education in the educational systems of other countries.

The corpus functions in the software environment developed at the University of Gdańsk and on its platform, but employing the TEITOK environment is being considered. There are two versions of the corpus available: one with marking corrections made by the authors (with limited access) and one without annotations (with open access). Morpho-syntactic annotation of texts is planned in the future.

Currently, the challenge is to transcribe handwritten texts, which means coping with difficulties such as interpreting graphs of individual shapes and graphs of similar shapes but different phonetic values in the student's first language (L1). Anticipated challenges at the morpho-syntactic annotation stage include automatic lemmatization and grammatical interpretation of incorrect textual forms and automatic syntactic interpretation of statements, as the texts contain many linguistically distorted forms.

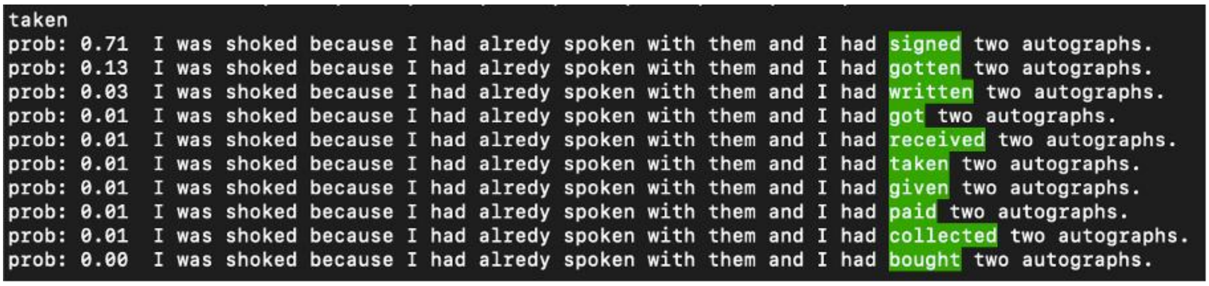
Exploring learner knowledge with Large Language Models fine-tuned with the EFCAMDAT

Ballier, Nicolas (Université Paris Cité / CLILLAC-ARP & LLF) and Bernardo Stearns (University of Galway)

Large Language Models (LLMs) like BERT (Bidirectional Encoder Representations from Transformers) have been trained to predict a masked token in a sentence (a bit like the Clauzure test in English). These LLMs have proved to be successfully retrained (fine-tuned) for specific tasks, such as question answering (cf. ChatGPT). Our paper explores the opportunities for learner corpus research of retraining an LLM like BERT (Devlin et al, 2016) with learner data to explore learner lexical knowledge.

We fine-tuned BERT models with samples from the EFCAMDAT (Geertzen et al, 2013) to investigate how these models may simulate learner behaviours of a given L1 or CEFR level. As an exploratory validation procedure, we used a subset of the FCE dataset (Yannakoudakis et al. 2011) that has been annotated for errors to evaluate the ability of our retrained model to predict the token produced by the learners. Only 57.5 % of the erroneous tokens according to the FCE annotators were predicted by our model with the first prediction, but 82,2% of the tokens were correctly predicted among the top 10 predictions.

We explain why analyzing the probabilities (prob in Figure 1) contribute to computational learner modeling (Abyaa et al, 2019). Examining the first ten predictions that we extracted from our fine-tuned model can be used as a window on lexical and grammatical knowledge, a virtual exploration of what a learner of the same level could have said. We highlight some of the challenges raised by this new way to revisit paradigmatic knowledge.



```
taken
prob: 0.71  I was shoked because I had alredy spoken with them and I had signed two autographs.
prob: 0.13  I was shoked because I had alredy spoken with them and I had gotten two autographs.
prob: 0.03  I was shoked because I had alredy spoken with them and I had written two autographs.
prob: 0.01  I was shoked because I had alredy spoken with them and I had got two autographs.
prob: 0.01  I was shoked because I had alredy spoken with them and I had received two autographs.
prob: 0.01  I was shoked because I had alredy spoken with them and I had taken two autographs.
prob: 0.01  I was shoked because I had alredy spoken with them and I had given two autographs.
prob: 0.01  I was shoked because I had alredy spoken with them and I had paid two autographs.
prob: 0.01  I was shoked because I had alredy spoken with them and I had collected two autographs.
prob: 0.00  I was shoked because I had alredy spoken with them and I had bought two autographs.
```

Figure 1: The first 10 predictions of the native (uncased) BERT model for the masked token (learner choice : taken) in the sentence I was shoked because I had alredy spoken with them and I had [MASK] two autographs.

References

Abyaa, A., Khalidi Idrissi, M., & Bennani, S. (2019). Learner modelling: systematic review of the literature from the last 5 years. Educational Technology Research and Development, 67, 1105-1143.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013, October). 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 (pp. 240-254).

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. Applied linguistics, 30(4), 461-473.

Khetan, V., Ramnani, R., Anand, M., Sengupta, S., & Fano, A. E. (2022). Causal BERT: Language models for causality detection between events expressed in text. In Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 1 (pp. 965-980). Springer International Publishing.

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. 180-189). Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2002472.2002496

Russian Spoken Learner Corpus: Designing and collecting a spoken corpus of Italian university learners of Russian

Bejenari, Oxana; Paola Cotta Ramusino, Claudio Macagno, and Tatsiana Maiko (Università degli Studi di Milano)

Interest in creating Russian learner corpora has emerged relatively recently, within the last decade (Kisselev 2023). The largest project to date, the Russian Learner Corpus (RLC, http://web-corpora.net/RLC), contains samples of primarily written (2,165,488 words) and partially spoken (93.381 words) samples produced by learners of Russian as a Foreign language and speakers of Heritage Russian with 48 dominant languages (Rakhilina et al. 2016). Currently, we are not aware of any other projects dedicated to creating a spoken learner corpus of Russian, aside from collections of spoken production samples gathered by researchers to address specific research questions. To the best of our knowledge, no attempts have been made previously to create a longitudinal corpus of data (both spoken and written) from learners whose native language is Italian.

This presentation introduces a work-in-progress project that aims to build a spoken learner corpus of Russian. It comprises longitudinal and pseudo-longitudinal oral data produced by Italian learners of Russian. In the longitudinal part of the project, data collection is conducted twice a year within the same group of students throughout their three/five-year study program. The pseudo-longitudinal subcorpus includes data produced by students across different proficiency levels, from A0→1 to C1. In addition to learner data, the corpus also includes two reference subcorpora. One subcorpus contains interviews with native speakers of Russian, while the other one consists of interviews with bilingual (Italian-Russian) speakers. The interviews are transcribed following explicit conventions. The database contains audio files, their transcripts, and detailed metadata about the interviewee, the interviewer, and the tasks. We will provide a brief overview of the project’s objectives, its structure, metadata, and discuss the progress made so far.

References

Kisselev, Olesya. 2023. “Russian Learner Corpora Research: State of the Art and Call for Action”. Bakhtiniana Revista de Estudos do Discurso 18 (1): 8-29.

Rakhilina, Ekaterina V., Anastasia S. Vyrenkova, Elmira Mustakimova, Alina Ladygina, Ivan Smirnov. 2016. “Building a Learner Corpus for Russian.” In Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, edited by Elena Volodina, Gintarė Grigonytė, Ildikó Pilán, Kristina Nilsson Björkenstam, Lars Borin, 66–76. Linköping: LiU Electronic Press.

Complexity or complexities? A simulation study on lexical complexity in expert and learner texts through the lens of information theory

Brasolin, Paolo (independent researcher) and Arianna Bienati (University of Modena and Reggio Emilia / Eurac Research)

The debate about linguistic complexity in general and lexical complexity in particular has been extremely lively, exploring both the validity of indices in capturing the construct (e.g., McCarthy & Jarvis, 2010; Kyle et al., 2021; Zenker and Kyle, 2021) and the theoretical foundations of the construct itself (e.g., Bulté and Housen, 2012; Jarvis, 2013; Pallotti, 2015). Intuitively, complexity transcends "the number and variety of an item's constituent elements" to include "the elaboratedness of their interrelational structure" (Rescher, 2020:1). This echoes the concept of Gell-Mann effective complexity in information theory, which emphasizes the amount of non-random information in a system, which peaks in the intermediate stage between order and disorder. Gell-Mann complexity is often opposed to Kolmogorov complexity, i.e., the total amount of information in a system, which monotonically increases from maximum order to maximum disorder.

This study explores which information-theoretical notion of complexity (Kolmogorov vs. Gell-Mann) is measured by widely used complexity indices, via a simulation study on four Italian corpora, representing the spectrum from expert to learner texts. New texts are synthesized from the originals by altering them in two directions: increased order is obtained as the repetition of increasingly smaller subsections of the original text, whereas increased disorder is obtained as the shuffling of increasingly smaller fragments of it. Additionally, we generate texts with uniform word distribution, simultaneously altering both the structure and the original word distributions. For each corpus, the synthetic data allow us to explore the spectrum from total order to total disorder. All texts are analyzed using type-token-ratio-based and surprisal-based metrics, including fluctuation complexity (Bates and Shepard 1993). Examining the distribution of the computed values shows that TTR-based metrics, except MATTR, are sensitive to increased order but not disorder. Surprisal-based measures, on the other hand, do show interesting Kolmogorov (entropy) or Gell-Mann behavior (normalized entropy and fluctuation complexity), enhancing their mutual interpretability when combined. Our results indicate that fluctuation complexity in particular could complement linguistic complexity tools, since it captures the intuitive notion of complexity in a text.

References

Bates, J.E., & Shepard, H.K. (1993). Measuring complexity using information fluctuation. Physics Letters A, 172(6), 416-425.

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 (pp. 21–46). John Benjamins. https://doi.org/10.1075/lllt.32.02bul

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. Language Learning, 63(s1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. Language Assessment Quarterly, 18(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Pallotti, G. (2015). A simple view of linguistic complexity. Second Language Research, 31(1), 117-134.

Rescher, N. (2020). Complexity: A Philosophical Overview. Routledge. https://doi.org/10.4324/9780429336591

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. Assessing Writing, 47, 100505. https://doi.org/10.1016/j.asw.2020.100505

Levels and Modeling of Variability in L2 Learner Corpora. Insights from a Corpus of Newly Migrated Students in Germany

Braunewell, Aylin; Julia Schlauch and Jana Gamper (Justus Liebig University Giessen)

This poster presents a corpus of newly migrated students in German schools, which is currently being developed (Schlauch 2022, Schlauch *et al*. in prep.). Newly migrated L2-learners of German face the challenge of acquiring the specific communicative requirements of school in a particularly short time and in a very heterogeneous context. In Germany, this challenge is met with the help of so-called intensive classes (cf. Massumi/von Dewitz 2015) that prepare newly migrated students for attending regular classes.

The corpus contains longitudinal oral and written data that was systematically collected from 15 L2-learners in intensive classes over 18 months. The aim of the corpus is to systematically capture and subsequently model inter- and intraindividual variability of language development processes (cf. Wisniewski et al. 2022). The corpus also aims to capture situational and registerspecific variation (cf. Lüdeling et al. 2022). To achieve this, a methodology inspired by the language situation approach by Wiese (2020) was developed which elicits various communicative tasks. Participants were instructed to describe and report six different illustrated events orally and to produce corresponding written texts.

We will present the corpus and discuss the specific challenges associated with the compilation and processing of a corpus of early learner varieties in a heterogeneous context, one of which is the high variability of the data. While variability is considered a central characteristic of learner language (cf. Schwendemann 2022) and investigating it systematically promises profitable insights into language acquisition, it also poses particular challenges: For example, preparing the annotation linguistically requires significant effort (cf. Wisniewski et al. 2022) and suitable methods must be identified to map the variability (Shadrova 2020, Schwendemann 2022). We will address these issues and present our specific approach to them.

References

Lüdeling, A., Alexiadou, A., Adli, A., et al. (2022). Register: Language Users' Knowledge of Situational-Functional Variation. Register Aspects of Language in Situation, 1, 1–58. https://doi.org/10.18452/24901

Massumi, M., & von Dewitz, N. (2015). Neu zugewanderte Kinder und Jugendliche im deutschen Schulsystem. Bestandsaufnahme und Empfehlungen. Herausgegeben vom Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache und vom Zentrum für LehrerInnenbildung der Universität zu Köln.

Schlauch, J. (2022). Erwerb der Verbstellung bei neu zugewanderten Seiteneinsteiger: innen in der Sekundarstufe. Eine Fallstudie aus dem DaZ-Lerner: innenkorpus SeiKo. Korpora Deutsch Als Fremdsprache, *2*(2), 43–62. https://doi.org/DOI: 10.48694/kordaf.3550.

Schlauch, J., Braunewell, A., & Gamper, J. (in prep.). Das Seiteneinsteiger: innenkorpus SeiKo. Ein Lerner: innenkorpus neu zugewanderter Schüler: innen in Intensivklassen.

Schwendemann, M. (2022). Variabilität als Faktor in der zweitsprachlichen Entwicklung syntaktischer Strukturen. Teilergebnisse einer longitudinalen Einzelfallstudie. Korpora Deutsch Als Fremdsprache, 2(2), 63–92. https://doi.org/10.48694/kordaf.3546

Shadrova, A. (2020). Measuring coselectional constraint in learner corpora: A graph-based approach [Dissertation]. Humboldt-Universität zu Berlin.

Wiese, H. (2020). Language Situations: A Method for Capturing Variation within Speakers' Repertoires. In Y. Asahi (Ed.), Proceedings of Methods XVI: Papers from the sixteenth international: Vol. v.59 (pp. 105–117). Peter Lang.

Wisniewski, K., Lüdeling, A., & Czinglar, C. (2022). Zum Umgang mit Variation in der Lernersprachenanalyse. Perspektiven aus und für DaF / DaZ. Deutsch als Fremdsprache, 4, 195–206. https://doi.org/10.37307/j.2198-2430.2022.04.03

Teaching interpersonal devices with English learner corpora

Carrió-Pastor, María Luisa (Universitat Politècnica de València, Spain)

It can be observed in the descriptors of the Common European Framework of References for Languages (CEFR 2001, 2018) that the teaching of the pragmatic strategies used by learners has been neglected in the skill of writing. In this sense, this study aims to fill this gap and to provide a detailed list of the devices that should be taught for the B2 and C1 proficiency levels of English learners. Thus, the objectives of this paper are, first, to identify the rhetorical strategies used by English learners in levels B2 and C1 and then to propose a grid to incorporate progressively the use of interpersonal devices in the different proficiency levels.

The methodology followed was based on a corpus-driven approach. First, opinion essays were compiled from the FineDesc corpus (https://web.ujaen.es/investiga/finedesc/index.php). Second, the different interpersonal devices were classified following the classification proposed by Thompson (2001) and using a tool designed for this purpose, METOOL. Finally, the interpersonal devices used in the B2 and C1 levels of language proficiency were integrated into a grid to be used by language instructors. The results not only show the different interpersonal strategies to be taught at B2 and C1 levels of language proficiency but also those to be acquired by English learners. This study focuses on a concern that is key to writing, the acquisition of rhetoric. Then, the findings could be helpful to specifically identify some of the pragmatic devices to be implemented in the CEFR and foreign language teaching.

References

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

Council of Europe. (2018). Common European Framework of Reference for Languages: learning, teaching, assessment. Companion volume with new descriptors. Council of Europe.

Thompson, Geoff. (2001). Interaction in academic writing: learning to argue with the reader. Applied Linguistics, 22–1, 58–78. https://doi.org/10.1093/applin/22.1.58

EFL students’ use of phrasal verbs of action and motion in spoken and written video-clip descriptions

Castello, Erik and Katherine Ackerley (University of Padua)

The use of phrasal verbs (PVs) differs according to register. Biber et al. (2021) show that PVs occur most frequently in both fiction and conversation, and less so in news/academic prose. Studies on learner corpora, on the other hand, reveal that learners tend to use PVs more frequently in written than in spoken production, a possible reason being their “lack of automaticity in the production of phrasal verbs under unplanned conditions” (Gilquin 2015: 81). There is, however, a lack of research comparing EFL learners’ use of PVs in texts responding to the same prompt, but produced in spoken and written mode.

This study investigates Italian learners’ use of PVs in written and spoken narrations, identifying their frequency and variety across proficiency levels. Texts produced by 201 students for version 2 of the COREFL corpus (Lozano et al. 2021) were analysed. The prompt was a video clip from Chaplin’s The Kid. The learner corpus was split into levels according to the results of a proficiency test. We are currently collecting spoken texts using the same prompt, and intend to use the COREFL NS written and spoken corpora for reference.

An analysis of the written data has revealed a consistent increase in both the frequency and variety of PVs at each level. Given the extensive variety of PVs identified, we decided to focus mainly on those used to express actions and motion in space (Ibarretxe-Antuñano 2017).

Our research questions are:

1. Which lexical verbs and particles combine to convey action and motion in the two registers across proficiency levels?
2. What differences are there in terms of frequency of phrasal verbs in spoken and written EFL narratives?

We will conclude by discussing how the results can inform approaches to teaching PVs for spoken and written production.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (2021) Grammar of Spoken and Written English. John Benjamins, Amsterdam/Philadelphia.

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.

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

Lozano, C., Díaz-Negrillo, A., & Callies, M. (2020) “Designing and compiling a learner corpus of written and spoken narratives: COREFL”. In C. Bongartz and J. Torregrossa (eds.), What’s in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy, Peter Lang, Bern, pp. 21-46.

Learner corpus genre analysis: What can it reveal about learners’ discoursal weaknesses?

Charles, Maggie (Oxford University) and Karin Whiteside (Reading University)

Genre analysis is extensively used in English for Academic Purposes (EAP), but typically involves the analysis of expert or L1 texts for instruction purposes (Swales, 2004). We would argue, however, that the genre analysis of learner corpus data provides a unique window into the discoursal problems faced by L2 writers as they attempt to produce successful instances of a genre. Specifically, disagreements by analysts on the function of stretches of learner text are likely to indicate key sites of learner weakness e.g. signalling issues or missing steps in logical development. Our learners are exiled Syrian academics on the Council for At-Risk Academics (Cara) Syria Programme. They have Arabic L1 (CEFR levels B1-C1) and are enrolled on the Cara EAP programme. They submitted proposals for research funding to Cara and our corpus consists of 102 of the proposal summaries (40,936 tokens). The corpus was examined using AntConc (Anthony, 2020) and a 3-move 10-step genre analysis framework based on Feng and Shi (2004) was developed by the Syrian-UK team. A pilot analysis (32 summaries) established move/step definitions and coding protocols. The remaining 70 summaries (28,651 tokens) were analysed by 2 team members independently; these are the data examined here.

Research questions:

1. What discrepancies are found between analysts in allocating text to moves/steps?
2. What do these discrepancies reveal about learners’ discoursal weaknesses when writing proposal summaries?

In total, 156 discrepancies were found (average: 2.2/summary). The most frequent discrepancy (43 occurrences) was between the analysis of text as Establishing the Territory or Indicating a Niche. Summaries lacked the necessary metadiscoursal signals to enable analysts to distinguish clearly between statements about the research background and statements indicating the research gap to be filled. This paper provides further details on learner weaknesses revealed by discrepancies in the genre analysis and discusses their pedagogical implications.

References

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

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

Swales, J. M. (2004). Research genres: Exploration and applications. Cambridge University Press.

A multidimensional comparison of argumentative writing of English learners from different language families

Chen, Mei-Hua (Tunghai University), Wei-Fan Chen (Rhenish Friedrich Wilhelm University of Bonn), Garima Mudgal (Paderborn University), and Henning Wachsmuth (Leibniz University Hannover)

The examination of how argumentative structures manifest among learners varied cultural backgrounds has attracted significant critical attention. While many studies have compared argumentative writing between native and non-native speakers (Plantin, 2020; Suzuki, 2010), there have been few attempts to investigate the argumentative communication styles of culturally diverse learners. To fill this research gap, we undertake a cross-cultural investigation to gain insights into how different learners construct their argumentative essays. In the analysis, we categorized over 6,000 argumentative essays from 16 diverse language backgrounds into eight language families, allowing for comparing and contrasting three dimensions: argumentation scores, common patterns, and the directness of communication style.

To fulfill these objectives, we applied argument mining techniques to identify components like major claims, claims, premises, and non-argumentative units (Stab & Gurevych, 2014) as well as to assess both the organization scores and argument strength scores of the essays (Persing, Davis, & Ng, 2010; Persing & Ng, 2015). Three major findings were identified: (1) Essays produced by learners, irrespective of their language background, may exhibit an appropriate argument structure, but the quality of the arguments might not reach optimal levels. (2) Learners from language families such as Bantu, Romance, Slavic, Japanese, and Oghuz tend to employ more premises than claims. In contrast, the Sinitic and Germanic language families prefer using claims over premises. The Uralic language does not display a distinct pattern. (3) The evaluation of communication style's directness involves two criteria: the placement of claims at the text's outset and the support of claims with reasons or premises in a paragraph. In general, over 80% of essays across all eight language families position their claims at the beginning of the essay. However, except for Japanese and Oghuz languages, fewer than 50% of essays demonstrate adequately supported claims.

References

Persing, I., Davis, A., & Ng, V. (2010). Modeling organization in student essays. Paper presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

Persing, I., & Ng, V. (2015). Modeling argument strength in student essays. Paper presented at the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Plantin, C. (2020). Argumentation through languages and cultures. In Argumentation Through Languages and Cultures (pp. 1-7): Springer.

Stab, C., & Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. Paper presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Suzuki, S. (2010). Forms of written arguments: A comparison between Japan and the United States. International Journal of Intercultural Relations, 34(6), 651-660.

CREATING A MULTIMODAL CORPUS OF INTERCOMPREHENSION: first reflections on annotation

Corino, Elisa and Eugenio Goria (University of Turin)

While the use of corpora as a basis for linguistic research is a well-established practice and corpus-based - if not corpus-driven - practice is in fact a shared and acknowledged methodology, there are areas in which there is a certain lack of structured and queryable data collections on which to carry out analyses and reflections. Intercomprehension (IC), a relatively young subject in the composite field of linguistics and educational linguistics, is one of these. IC is "la capacité de comprendre une langue étrangère sur la base d'une autre langue sans l'avoir apprise" (Meissner 2004). In short, it is a matter of communicating in a multilingual context using one's mother tongue and implementing decoding strategies based on the comparison of the elements of the two idioms.

Based on these considerations, the research group of the PLUS-SI (PLUrilinguism for Social Sustainability and Youth Inclusion) project, which shares with the UNITA European Alliance the principles and aims of applied research related to plural approaches, has initiated the construction of a multimodal corpus of IC interactions (UnICo - Unita Ic Corpus) annotated with ELAN at different levels, which will be freely accessible by the scientific community.

This contribution will present the reflections that led to the definition of the corpus architecture, the different levels of labelling (Jefferson, PoS, Lexicon, Pragmatics and Linguistic Acts, Plurilingualism, Non-verbal Language among the most relevant), will show some initial examples of annotated files as the result of the design process, and will perform some queries to show the possible research applications of UnICo's tagset and content.

References

Aijmer K, Rühlemann C. (Eds). Corpus Pragmatics. A Handbook. Cambridge: Cambridge University Press.

Bonvino E., Garbarino S. 2022. Intercomprensione. Bologna: Caissa Italia.

Bosco C., Ballarè S. et al. 2020. KIPoS @ EVALITA2020: Overview of the Task on KIParla Part of Speech Tagging. In Basile V., Croce D., Maro M.,&amp; Passaro L. C. (Eds). EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop. Torino: Accademia University Press. DOI: 10.4000/books.aaccademia.7743

Brunner M. L., Diemer S. 2021. Multimodal meaning making: The annotation of nonverbal elements in multimodal corpus transcription. In Research in Corpus Linguistics. 9:1. 63–88. DOI: 10.32714/ricl.09.01.05

Garbarino S., Leone P.. 2021. "Je suis pas sûre d'avoir compris la dernière phrase...". Capirsi e collaborare in contesti di intercomprensione.

Jefferson G. 2004. Glossary of transcript symbols with an introduction. In Lerner (Ed). Conversation Analysis. Studies from the first generation. John Benjamins Publishing Company.

Meißner F. J. 2021. The Core Vocabulary of Romance Plurilingualism. French and Italian word lists. Giessen University Library Publications.

Meißner, F.-J. 2004. EuroComRom-Les sept tamis: lire les langues romanes des le depart, Shaker Verla, Aachen, Germany.

Getting down to business with the Apprentice Multiple Business GEnRes (AMBER) corpus

De Cock, Sylvie (UCLovain) and Jennifer Thewissen (Universiteit Antwerpen; UCLouvain)

Business texts represent a wide variety of genres, which tend to be identified on the basis of communicative purpose(s) and situation(s) (Bhatia 1993, Koester 2010). A broad distinction (Nelson 2000) can be made between genres used to do business and communicate to get work done within the framework of companies'/organisations' activities (e.g. business meetings, social media posts, minutes of meetings, reports) and genres not issued by companies/organisations that are used to talk or write about business (e.g. news articles about the world of business, business studies lectures). An examination of the Learner Corpora around the World webpage (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html) reveals that learner corpora made up of a variety of business genres are rather few and far between and tend not to be readily accessible to the research community.

This poster sets out to introduce the recently launched Apprentice Multiple Business GEnRes (AMBER) corpus to the learner corpus community and to further develop the AMBER network of partners. The poster aims to explain the rationale behind this new multi-L1 learner corpus collection project which first concentrates on productions in English (e.g. providing a strong empirical basis to investigate genre awareness among learners / novice users of business genres) and to outline the corpus design criteria with a specific focus on the business genres included (e.g. press releases, cover letters), and on the various task and learner variables that lie at the heart of the AMBER project. Information about the core metadata recorded is also provided.

References

Bhatia, Vijay Kumar. 1993. Analysing genre: Language use in professional settings. London: Longman.

Koester, Almut. 2010. Workplace discourse. London and New York: Continuum.

Nelson, Mike. 2000. A corpus-based study of the lexis of business English and business English teaching materials. Unpublished PhD thesis. University of Manchester.

Introducing the CLARIN Knowledge Centre for Learner Corpora

De Cock, Sylvie; Damien De Meyere, Liesbeth Degand, Thomas François, Gaëtanelle Gilquin, Sylviane Granger, Danqing Huang, Marie-Aude Lefer, Hubert Naets, Magali Paquot, Kristel Van Goethem, Patrick Watrin with the collaboration of Jennifer Carmen-Frey, Alexander König and Egon Stemle

The main objective of this poster is to introduce the CLARIN Knowledge Centre for Learner Corpora (CKL2CORPORA), present its core missions and report on current projects.

CKL2CORPORA builds on more than 30 years of expertise in learner corpus research at the Centre for English Corpus Linguistics (CECL). CKL2CORPORA also relies on the expertise of staff from two additional research centres from the Institute for Language and Communication, UCLouvain, i.e. CENTAL and VALIBEL. Together, CKL2CORPORA members have expertise in learner corpus design (e.g., metadata, transcription, file formatting, ethics), annotation (e.g., POS tagging, parsing, error annotation), and analysis (e.g., error analysis, contrastive interlanguage analysis, transfer studies) in a range of languages (English, French, Dutch, etc.).

CKL2CORPORA's sharing of expertise can take various forms, from answering (theoretical, methodological, technical) questions sent via the helpdesk to providing training services. For example, CKL2CORPORA members organized the fifth edition of the Learner Corpus Research Summer School last summer. Additionally, they actively maintain a webpage aiming to list and classify existing learner corpora in various languages (Learner Corpora around the World) and have contributed to the development of corpus query tools such as Corpor@uclouvain, an online interface that provides access to corpora developed by UCLouvain members, and the UCLouvain Error Editor (UCLEE), a tool specifically designed to facilitate the insertion of error tags and corrections into learner texts, as well as their subsequent processing (Granger et al., 2023).

Among the other projects undertaken by CKL2CORPORA members is the creation of FABRA (Wilkens et al., 2022), initially designed for readability research but versatile enough to compute an extensive array of linguistic complexity measures for L2 French. Furthermore, the CKL2CORPORA members have actively contributed to the development of a Core Metadata Schema for Learner Corpora that can be used to document a learner corpus (Paquot et al., 2023).

What a L1-German Learner Corpus can tell us about the acquisition of Romance Languages: The MuLeCo Project

De Crignis, Patricia and Johanna Wolf (LMU Munich)

This contribution presents data from the MuLeCo learner corpus (L1 German, Romance learner languages), focusing on differential object marking in Spanish and gender assignment in Spanish and French. The data was collected as part of field research with the picture story Frog, where are you? at schools. In the case of the Spanish data, 66 participants were included in the study, 37 of whom were learners of Spanish (A2 and B1) and 29 of whom were L1 speakers of Castillian Spanish (control group). A total of 925 constructions with a direct object were analyzed, with 338 constructions (36.54 %) in the learner group and 587 constructions (63.46 %) in the L1 control group.

In our study on gender assignment and agreement in French 42 texts of A2 learners and 37 texts of B1 learners were analyzed. For Spanish the analysis is still ongoing, so far 10 texts of learners with A2-level were analyzed. Regarding the French data, a total of 1268 DPs and a total of 275 APs were analyzed at the A2-level and a total of 1249 DPs and a total of 318 APs were analyzed at the B1-level.

Using the results for differential object marking, we can show in particular how important it is to include L1 control groups in studies on SLA and to use their productions to define what constitutes an error: the results of the standard error annotation and the productions of the L1 control group differ considerably in our corpus for animate no-human direct objects. The data on gender assignment show that there are differences between learners of French and Spanish which are probably due to the language systems and therefore need different approaches in FLT, e.g. to facilitate the processing of the phenomenon already in the input.

References:

Ayoun, D. (2020). A Longitudinal Study in the L2 Acquisition of the French TAM System. Languages, 5(4), 42, 1-23

Brezina, Vaclav, and Gabriele Pallotti (2019). Morphological complexity in written L2 texts. Second language research 35(1), 99-119.

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

De Clercq, Bastien, and Alex Housen (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. Second Language Research 35(1), 71-97.

Delbecque, Nicole (1994). Hacia una aclaración cognitiva del acusativo preposicional. Boletín de la Soc. Española para el Procesamiento del Leng. Natural 14, 33-45.

Gilquin, Gaëtanelle, Sylviane Granger, and Magali Paquot (2007). Learner corpora: The missing link in EAP pedagogy. Journal of English for Academic Purposes 6(4), 319-335.

Gilquin, Gaëtanelle (2022). Cognitive corpus linguistics and pedagogy: From rationale to applications. Pedagogical Linguistics 3(2), 109-142.

Gilquin, Gaëtanelle (2022). One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. Language Teaching 55(1), 87-99.

Granger, Sylviane (2014). The computer learner corpus: a versatile new source of data for SLA research. Learner English on computer. Routledge, 3-18.

Herschensohn, Julia. (2001). Missing inflection in second language French: Accidental infinitives and other verbal deficits. Second Language Research 17, 273-305.

Heusinger, Klaus von, Kaiser, Georg A. (2011). Affectedness and Differential Object Marking in Spanish. Morphology 21, 593–617.

Hirschmann, Hagen, et al. (2022). FALKO. Eine Familie vielseitig annotierter Lernerkorpora des Deutschen als Fremdsprache. Korpora Deutsch als Fremdsprache 2.2.

Manyasa, Jonace (2019). Analysis of French grammatical gender errors committed by learners in Tanzanian universities. Journal for Foreign Languages 11(1), 65-86.

Mayer, Mercer (1969). Frog, where are you? New York: The Dial Press.

Römer, U. (2006). Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. Zeitschrift für Anglistik und Amerikanistik, 54(2), 121-134.

Tracking phraseological inter-complexity: a corpus based contrastive analysis of learner English and Romanian novice academic writing

Dinca, Andreea; Ana-Maria Bucur and Madalina Chitez

Learner corpora have been used to analyze learner language from multiple perspectives, which include, for example the phraseological repertoire in L2 (e.g. Hasselgård, 2019), types of errors (e.g. Gilquin et al., 2007) or genre features (Staples & Reppen, 2016). Several studies have used L2 writers' native language datasets to extract interlanguage patterns (e.g. Paquot, 2013). In this paper, we contrast two datasets representing student academic writing at Romanian universities: academic writing in English L2 and Romanian L1. The datasets are part of the ROGER corpus (Chitez et al., 2022), which is freely available and searchable via the self-developed ROGER platform (Striletchi et al., 2022). We use the EXPRES corpus (Chitez et al, 2023), to compare phraseology in L1 and L2 at the expert level as well. The texts included in the corpus are distributed in eight disciplinary fields (e.g. Humanities, Economics).

In the present paper, we aim at detecting phraseological inter-complexity, manifesting as the correlation between the complexity of phrases used in L1 and L2. Our target group are Romanian learners of English. Phraseological complexity is operationalized through the measurement of phraseological diversity and sophistication in adjectival modifiers (amod), adverbial modifiers (advmod), and direct objects (dobj) (see e.g. Vandeweerd et al., 2023). We use the same approach for both English and Romanian language data. For example, Romanian L1 novice writers use less sophisticated direct objects, with a pointwise mutual information (PMI) score of 1.97 compared to native English expert academic writing, which typically displays a PMI score exceeding 3. Our findings also indicate that L2 writers sometimes "borrow" phraseology from the native language into English (e.g. Ro. dobj. pune accent > En. dobj. put emphasis). The results of this study can be used for pedagogical and applied research.

References:

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

Chitez M, Bercuci L, Dincă A, Rogobete R, Csürös K. Corpus of Romanian Academic Genres (ROGER) (2022). West University of Timisoara. Available at https://roger-corpus.org/.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner Corpora: The Missing Link in EAP Pedagogy. Journal of English for Academic Purposes, 6(4), 319–335.

Hasselgård, H. (2019). Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In: V. Wiegand and M. Mahlberg (eds.) Corpus Linguistics, Context and Culture (pp. 339-362). De Gruyter.

Paquot, M. (2013). Lexical bundles and L1 transfer effects. International Journal of Corpus Linguistics, 18(3), 391–417. https://doi.org/10.1075/ijcl.18.3.06paq

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. Journal of Second Language Writing, 32, 17-35.

Striletchi, C., Chitez, M. and Csurös, K. (2022). Building Roger: Technical Challenges While Developing a Bilingual Corpus Management and Query Platform. In H.-G. Fill, M. van Sinderen and L. Maciaszek (Eds.), Proceedings of the 17th International Conference on Software Technologies (ICSOFT), 11 - 13 July 2022, Lisbon, Portugal (pp. 390-398). Setúbal: SciTePress.

Vandeweerd, N., Housen, A., & Paquot, M. (2023). Comparing the longitudinal development of phraseological complexity across oral and written tasks. Studies in Second Language Acquisition, 45(4), 787-811. doi:10.1017/S0272263122000389

Syntactic complexity across L1, L2 and L3

Dirdal, Hildegunn (University of Oslo), Stine H. Johansen (University of Oslo) and Philip Durrant (University of Exeter)

Although cross-linguistic influence has been found to be important in all areas of language acquisition, only a few complexity studies have focused on such effects (e.g. Lu & Ai 2015; Ehret and Szmrecsanyi 2019; Dirdal 2022) and there is a lack of studies taking a properly multicompetence perspective investigating the various languages of the same individuals (Cenoz 2013). This study aims to investigate correlations in the complexity of L1, L2 and L3 writing. There has recently been a call for more detailed measures of syntactic complexity (Biber et al. 2016; Biber et al. 2020), and as a response we focus on the various forms of clause types that learners need to master, asking the following question:

Does L1 complexity predict complexity in L2 and L3 (1) as evidenced by overall subordination rates and (2) as evidenced by particular clause types?

We analyse a set of approx. 650 texts written by 57 first-year upper-secondary students in the three languages they learn at school: Norwegian, English and French/German/Spanish. Texts are hand-annotated for features of clausal and phrasal structure using a custom-developed grammatical framework that can be applied across languages. A range of both broad and fine-grained complexity measures are then quantified for each text and summary measures calculated for each writer within each language. Correlations are used to examine the associations between complexity measures in writers' first and second languages. Regression models with L1 complexity, L2 complexity, and language (French, Spanish, German) as dependent variables are used to examine the extent to which L3 complexity is predicted by each of these variables.

The results will increase our knowledge about interactions between the three languages of an individual and will be valuable for educational development, particularly with respect to cooperation across language subjects.

References

Biber, D., Gray, B. & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. Applied Linguistics 37(5): 639–668. https://doi.org/10.1093/applin/amu059

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. Journal of English for Academic Purposes 46, Article 100869. https://doi.org/10.1016/j.jeap.2020.100869

Cenoz, J. (2013). The influence of bilingualism on third language acquisition: Focus on multilingualism. Language Teaching 46, 71–86. https://doi.org/10.1017/S0261444811000218

Dirdal, H. (2022). Development of L2 writing complexity: Clause types, L1 influence and individual differences. In A. Leńko-Szymańska & S. Götz (Eds.), Complexity, Accuracy and Fluency in Learner Corpus Research (pp. 81–113). John Benjamins. https:/doi.org/10.1075/scl.104.04dir.

Ehret, K. & Szmrecsanyi, B. (2019). Compressing learner language: An information-theoretic measure of complexity in SLA production data. Second Language Research 35(1): 23–45. https://doi.org/10.1177/0267658316669559

Lu, X. & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. Journal of Second Language Writing 29, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

Interacting to learn: Towards an inclusive corpus of EFL learners' interactions with AI-based chatbots

Ferraresi, Adriano and Silvia Bernardini (University of Bologna)

This work-in-progress report describes the design criteria and data collection procedure of a novel learning corpus resource that is currently being produced to investigate written interactions by learners of English as a Foreign Language (EFL) with chatbots. Chatbots, i.e. automated agents designed to engage in meaningful interaction with humans, have attracted the interest of teachers and researchers for a number of reasons, including the variety of learning contexts and tasks for which they can be used and their virtually unlimited availability for students (Huang et al. 2022). Several studies to date have focused on the effectiveness of chatbots in enhancing learners' language competence and/or motivation (Bibauw et al. 2022), but to the best of our knowledge none have collected/analysed learners' output within such interactions, nor have they compared these interactions with similar ones among humans (e.g. from TeleKorp; Belz 2006).

The corpus under development features written interactions of Italian university students with chatbots based on Large Language Models, which are used in different EFL tasks/scenarios (task-oriented roleplay vs. small talk). Learners are recruited among Bachelor students in non-language-related degree programmes, targeting specifically learners at lower proficiency levels (B2 or lower). With the ultimate aim of favouring equal access to learning opportunities (CAST 2018), the resource includes the production of learners with disabilities and specific learning disorders. The report will highlight the educational affordances of the corpus, which gives prominence to a text production mode likely to gain increasing importance in educational contexts (Konke et al. 2023), and prioritizes the inclusion of diverse learners as a relevant criterion in learner corpus design (cf. Gilquin 2015). At the same time, it will reflect on the nature of data obtained through learner interaction with Artificial Intelligence applications with respect to central notions of corpus linguistics such as authenticity and representativeness.

References

Belz, J. (2006) At the intersection of telecollaboration, learner corpus analysis, and L2 pragmatics: Considerations for language program direction. In J. Belz and S. Thorne (eds.) Internet-mediated Intercultural Foreign Language Education. Heinle.

Bibauw, S., W. Van Den Noortgate, T. François and P. Desmet (2022). "Dialogue systems for language learning: a meta-analysis". Language Learning & Technology, 26(1).

CAST. 2018. Universal Design for Learning guidelines. http://udlguidelines.cast.org

Gilquin (2015) "From design to collection of learner corpora". In S. Granger, G. Gilquin and F. Meunier (eds.) The Cambridge Handbook of Learner Corpus Research. Cambridge University Press.

Huang, W., K. F. Hew and L. Fryer (2022). "Chatbots for language learning — Are they really useful?". Journal of Computer Assisted Learning, 38(1).

Kohnke, L., B. L. Moorhouse and D. Zou (2023). "ChatGPT for language teaching and learning". RELC Journal, 54(2).

Syntactic and semantic features of concession in high school students' writing: a corpus analysis of Italian

Ferrato, Elena (University of Verona & Free University of Bolzano and Eurac Research)

This study analyzes concessive structures in 227 texts extracted from the ITACA corpus [1], encompassing 635 texts produced by high school students (median age: 18) whose language of instruction is Italian1. It seeks to answer two research questions: (1) What are the syntactic and semantic features of concession in these texts? (2) What markers introduce concession?

In contrast to prior investigations that relied on contrived examples [2] or extracts from narrative [3] and technical texts [4], this study advocates for a corpus-based approach using naturalistic data.

The ITACA corpus highlights challenges in coherence (mean values: 4.1/10 for text structuring, 6.1/10 for text segmentation, and 6.5/10 for comprehensibility) but provides a valuable perspective on the evolving language of the new generation entering higher education or the workforce. Syntax-wise, the present study identifies the emergence of new structures, such as "composite" structures combining previously recognized categories and a broader form of juxtaposition marked by concessive relations but separated by intervening textual segments.

Regarding concession markers, the study corroborates the presence of frequently occurring markers found in prior research [2–4], e.g ma (but) or anche se (even if). However, it also unearths 35 multi-word expressions not neatly fitting into established categories (e.g., ma è anche vero che, but it is also true that) a noteworthy observation rarely explored in Italian language studies (for English and Spanish see [5]).

In the semantic domain, the study aligns with Izutsu's [6] parameters for distinguishing concession from semantic opposition [7] but falls short of definitively determining the number and characteristics of semantic sub-types of concession. Previous debates propose varying classifications (from three to ten semantic subtypes, cf. [2] and [8]), emphasizing the need for further investigation.

In conclusion, this research offers an exploration of concessive structures in high school students' Italian texts, revealing the emergence of expressions of concession that have not been previously examined. Further investigation is needed to unravel intricate semantic sub-types within the concession category.

References

[1] The ITACA Corpus. PORTA Eurac Research Learner Corpus Portal. URL: https://www.porta.eurac.edu/lci/itaca/ (visited on 01/27/2024).
[2] Marco Mazzoleni. "Le frasi concessive". Grande grammatica italiana di consultazione, nuova ed. 3 vols. Il Mulino, 2001, pp. 784–817.
[3] Giovanni Battista Moretti. Riflessioni sulla concessione e sulla ammissione nell'italiano contemporaneo. Le Edizioni Università per stranieri, 1983.
[4] Claudio Di Meola. Der Ausdruck der Konzessivität in der deutschen Gegenwartssprache. Theorie und Beschreibung anhand eines Vergleichs mit dem Italienischen. Niemeyer, 1997. 1The 72.2% of 545 students declared Italian as their L1, whereas the 21.8% two or more languages, the 3.3% other language, the 1.5% German, the 0.9% Ladin.
[5] Maite Taboada and María de los Ángeles Gómez-González. "Discourse markers and coherence relations: Comparison across markers, languages and modalities". Linguistics and the Human Sciences (2010), pp. 17–41. (Visited on 10/12/2022).
[6] Mitsuko Narita Izutsu. "Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations". Journal of Pragmatics (2008), pp. 646–675. ISSN: 03782166. DOI: 10.1016/j.pragma.2007.07.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378216607001178 (visited on 10/26/2022).
[7] Robin Lakoff. "If's, And's and But's About Conjunction". Studies in Linguistic Semantics. Irvington, 1971, pp. 3–114.
[8] Ilde Consales. La concessività nella lingua italiana (secoli XIV - XVIII). Studi linguistici di storia della lingua italiana. Aracne, 2005.

DMorphC: A web tool for carrying out the identification, grouping and counting of morphologically complex words.

Flores Hernández, Ana Abigahil; Hilda Hidalgo Avilés and Abigail Carretero Hernández (Universidad Autónoma del Estado de Hidalgo)

Finding derivational suffixes in a corpus is a laborious and time-consuming task typically undertaken manually. This process involves initially employing a word-ending list followed by a meticulous verification of the morphological composition of complex words in various reference sources such as lists, grammars, or dictionaries (Naismith & Kanwit, 2021; Namtapi, 2020; Liu & Shen, 2012; Morita, 2020; García, 2009; Lee & Miller, 2020; Arslan, Mahmood & Rasool, 2020).

This proposal introduces the development of the Derivational Morphemes Counter (DMorphC), a corpus-based tool designed to identify 147 derivational suffixes in any given corpus. DMorphC operates through the integration of two key components: Firstly, it employs a meticulously curated suffix list adapted from Laws and Ryder (2014), the Oxford English Dictionary (2023), and Stein (2007). Secondly, it incorporates a comprehensive reference list of derived words sourced from Morphoquantics (Laws & Ryder, 2014) and Morpholex (Sánchez-Gutiérrez, 2018) corpora, supplementing it with inflections as needed. This innovative tool significantly reduces the time required for searching word forms and derivational suffixes, with statistical validation confirming its equivalence to manual counts, as demonstrated through its application to a learner corpus of written essays by Mexican learners of English (Flores, 2019).

As DMorphC operates by identifying complete lexical units, it presents a limitation in recognizing neologisms and unconventional uses of morphological elements, especially prevalent in learner language. However, this limitation is effectively addressed by generating a list of unaccounted words. In instances where words are not captured by the tool, searches can be manually conducted with relative ease, shortening the time required compared to scanning the entire corpus.

The manual processing of large volumes of text for derivational morphemes is hindered by human limitations, such as imperfect training and fatigue, leading to errors in identifying complex word forms or misinterpreting their morphological structures. This highlights the advantage of an automated approach, which offers superior accuracy in identifying the units included in the reference lists integrated in DMorphC.

References

Arslan, M., Mahmood, M. & Rasool, A. (2020). A corpus-based comparative study of derivational morphemes across ENL, ESL, EFL learners through ICNALE. Linguistic Forum 2 (4), 1-12.

Flores, A. (2019). Adquisición de sufijos derivativos en L2 (Unpublished doctoral dissertation). Universidad Autónoma de Querétaro.

García, B. (2009). The diminutive suffix “-et/-ette”: the role of the internet in its study. Revista Canaria de Estudios Ingleses, 58, 133-145. Gilquin & F. Meunier (Eds), the Cambridge handbook of learner corpus research (pp. 537-566). Cambridge University Press.

Laws, J. & Ryder, C. (2014). Getting the measure of derivational morphology in adult speech a corpus analysis using MorphoQuantics. Language studies working papers, 6, 3-17.

Lee, C. & Miller, J. (2020). A corpus-based list of commonly used English medical morphemes for students learning English for specific purposes. English for Specific Purposes 58, 102–121.

Liu, W. & Shen, H. (2012). A Corpus-based Analysis of English Suffix –esque. Theory and Practice in Language Studies, 2(4), 767-772.

Morita, J. (2020). A corpus-based study of derivational morphology and its theoretical implication. Proceedings of 4th International Conference of Computational Linguistics in Bulgaria 2020, 8-16.

Naismith, B. & Kanwit, M. (2021). A corpus study of the English suffixes -ness and -acy: productivity, genre, and implications for l2 learning. Canadian Journal of Applied Linguistics, 24, (1), 115-137.

Namtapi, I. (2020). A corpus-based study of English adjective formation using the suffix –ish. Parichart Journal Thaksin University, 34(3), 150-165.

Oxford UP (2023). Oxford English dictionary online. http://www.oed.com

Sánchez-Gutiérrez, C., Mailhot, H., Deacon, H. and Wilson, M. (2018). MorphoLex: A derivational morphological database for 70,000 English words. Behaviour Research Methods, 50, 1568–1580.

Stein, G. (2007). A dictionary of English affixes their function and meaning. LINCOM. Suhandoko and Ningrum, D. (2020). A corpus-based list of academic English derivational suffixes. Indonesian Journal of Applied Linguistics, 10, 481-490.

The role of learner corpora in needs analysis for ESP (English for Specific Purposes)

Flowerdew, Lynne (Birkbeck, University of London, UK)

Needs analysis, carried out to establish the ‘what’ and the ‘how’ of an ESP (English for Specific Purposes) course, is the essential first step in course development. Brown’s (2016) informative volume on needs analysis for English for specific purposes has a short section on the use of corpora to which other references could be added. For example, Nesi (2015) underscores the need to compile ESP corpora which focus on the types of texts learners will engage in. Granger and Paquot (2015) describe a needs-driven online writing aid accessing subject-specific corpora. Meanwhile, Jeaco (2020) illustrates how learners can engage in needs analyses of their own by exploring specific corpora of texts they have built themselves.

Learner corpora clearly also have an important role to play in needs analyses for ESP, thus bridging the gap between learner corpus research and pedagogic applications (Gilquin et al. 2007). However, this role is not always explicitly stated (Author, 2021). The aim of this presentation is to discuss how the findings from key studies involving various types of ESP learner/apprentice written and spoken academic and professional workplace corpora have been used to inform needs analyses and hence course design, which can also be corpus-based. Most of these studies are of a contrastive nature comparing learner/apprentice language with expert discourse (Author, in press 2024). These ESP needs analyses involving specialised learner corpora span a wide range of academic and professional disciplines including healthcare, aviation, law and engineering, addressing ‘real world’ concerns. Looking to the future, Xu and Sun (2023) have built an AI generated English essay corpus, which can usefully serve as a reference corpus for learner corpus research. Such generative technology would lend itself to construction of other specific types of reference corpora for comparison with learner corpora in ESP needs analyses.

References

Author (2021).
Author (in press, 2024).
Brown, J.D. (2016). Needs Analysis and English for Specific Purposes. London: Routledge.
Gilquin, G., Granger, S. & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. Journal of English for Academic Purposes, 6 (4): 319-335.
Granger, S. & Paquot, M. (2015). Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. Lexicographica: International annual for lexicography 31 (3): 118-141.
Jeaco, S. (2020). DIY needs analysis and specific text types: Using the Prime Machine to explore vocabulary in readymade and homemade English corpora. In M. Dodigovic & M. Agustin-Llach (Eds.) Vocabulary in Curriculum Planning: Needs, strategies and tools (pp. 199-223). New York: Springer.
Nesi, H. (2015). ESP corpus construction: A plea for a needs-driven approach. Asp La Revue du Geras, 68: 7-24.
Xu, J. & Sun, M. (2023). aiTECCL: An AIGC English essay corpus. Beijing: National Research Centre for Foreign Language Education, Beijing Foreign Studies University. Available online: http://corpus.bfsu.edu.cn/info/1082/1913.htm

Assessing the validity of new structural complexity metrics as features of L2 proficiency

Gaillat, Thomas (Université Rennes 2)

The study of the development of grammatical complexity in writings has been approached with the use of holistic complexity metrics (Bulté & Housen, 2012; Norris & Ortega, 2009) and the use of syntagmatic forms mapped to specific functions (Biber et al., 2011, 2020, 2023). In both cases, lexico-grammatical patterns are counted along the syntagmatic axis to compute proportions or ratios, and analyzed in terms of correlations with proficiency/developmental stages. For all their benefits, these studies have left paradigmatic production out of the equation. Few studies have focused on analysing the internal structure of a form-function mapping despite evidence of paradigmatic instability within mappings (Bresnan et al., 2007; Gaillat et al., 2021).

Based on the construct of microsystems (Gentilhomme, 1980; Py, 1980), we operationalise the measurement of the probability of occurrence of a form vs its paradigmatic competitors within mappings. For instance, we compute the probability of occurrence of proform THIS vs THAT or IT. To do so, after pre-processing the EFCAMDAT corpus (Shatz, 2020) with UDPipe (Straka et al., 2016), we extract the forms of a microsystem with GrewMatch (Guillaume, 2021) and train a model (multinomial logistic regression with Nnet library in R) on a randomized training subset (80%). We apply the same extraction strategy and the model prediction on the test set. We use ordinal logistic regression to evaluate associations between probabilities and CEFR-based proficiency. Premiminary results for the proform microsystem show that the odds of the median of the forms' probabilities per text are associated with proficiency (Kruskal-Wallis rank sum test across CEFR: $p < .001$). The ordinal regression indicates that THAT and THIS probabilities are strongly associated with higher levels, whilst IT probability is associated with low levels. Other candidates for microsystems include groups of determiners, relative pronouns and multiword structures. These new metrics provide measures of paradigmatic dimension in the assessment of writing.

References

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. https://doi.org/10.5054/tq.2011.244483

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. Journal of English for Academic Purposes, 46, 100869.

Biber, D., Szmrecsanyi, B., Reppen, R., & Larsson, T. (2023). Expanding the scope of grammatical variation: Towards a comprehensive account of genitive variation across registers. English Language & Linguistics, 1–39. https://doi.org/10.1017/S1360674323000497

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In B. Gerlof, I. Kramer, & J. Swarts (Eds.), Cognitive foundations of interpretation (pp. 69–94). Royal Netherlands Academy of Arts and Sciences.

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

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, 34(2). https://doi.org/10.1017/S095834402100029X

Gentilhomme, Y. (1980). Microsystèmes et acquisition des langues. Encrages, Numéro spécial, 79–84.

Guillaume, B. (2021, April 19). Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics. https://hal.inria.fr/hal-03177701

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. https://doi.org/10.1093/applin/amp044

Py, B. (1980). Quelques réflexions sur la notion d'interlangue. Revue Tranel (Travaux Neuchâtelois de Linguistique), 1, 31–54.

Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. International Journal of Learner Corpus Research, 6(2), 220–236. https://doi.org/10.1075/ijlcr.20009.sha

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), 4290–4297. https://aclanthology.org/L16-1680

Exploring Verb-Noun collocations across L2 English proficiency levels in L1 Mandarin and French learners

Gaillat, Thomas and Jen-Yu Li (Université Rennes 2)

Collocations, regarded as an essential component of L2 lexical competence (Granger & Larsson, 2021), fall under the umbrella of phrasemes. Wang & Shaw (2008) discovered that advanced students, irrespective of their native language, employ a similar quantity of collocations, whereas L2 English learners use fewer collocations than their native peers. Yoon (2016) noted that high-intermediate students favored high-frequency collocations, while native speakers used three times as many infrequent ones. Paquot's (2018) study found no systemic increase in collocation diversity with proficiency levels. Based on Mutual Information (MI) score bins, Eguchi and Kyle (2023) identified clear patterns in collocation use across proficiency levels. However, the influence of L1 and linguistic features on these patterns remains unclear. For example, light verb constructions, exemplified by "do," "get," and "make," may exhibit distinct patterns compared to other verb types.

This study investigates Verb-Noun (VN) collocations in the EFCamDat (Geertzen et al., 2013; Shatz, 2020) across CEFR proficiency levels. The research question addresses whether there are specific structural features of collocations that associate with different proficiency levels, in terms of MI-score profiles. Collocations were extracted using a predefined algorithm (Author, 2020) for two L1 groups, French (Fr) and Mandarin (Mn). The NLP libraries UDpipe (Straka et al., 2016) and NLTK (Bird et al., 2009) were used for text parsing and collocation extraction, respectively. The analysis considered L1s, MI scores as well as some linguistic features (verb type, distance between VN, sentence type, etc.). Several statistical comparisons were performed. For instance, Mann-Whitney U rank test showed that the difference of MI scores between Mn-C1 and Fr-C1 groups is statistically significant ($p<0.01$), but the difference between Mn-C1 and Mn-B2 groups is not. Results indicate non-uniform and non-identical relationships between L1s, linguistic features and proficiency levels. Findings enhance our comprehension of learners' collocation use in diverse linguistic contexts.

References:

Bird, S., & Loper, E., & Ewan Klein, E. (2009), Natural Language Processing with Python. O'Reilly Media Inc.

Eguchi, M., & Kyle, K. (2023). L2 collocation profiles and their relationship with vocabulary proficiency: A learner corpus approach. Journal of Second Language Writing, 60, 100975.

Granger, S., & Larsson, T. (2021). Is core vocabulary a friend or foe of academic writing? Singleword vs multi-word uses of thing. Journal of English for Academic Purposes, 52, 100999.

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.

Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. Language Assessment Quarterly, 15(1), 29–43.

Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. International Journal of Learner Corpus Research, 6(2), 220–236.

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 4290–4297.

Wang, Y., & Shaw, P. (2008). Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. ICAME Journal, 32, 201–232.

Yoon, H.-J. (2016). Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. Journal of Second Language Writing, 34, 42–57.

Understanding Foundation-level ESL Students’ Academic Writing: A Lexico-grammatical Analysis across Genres and Disciplines

Gao, Mingyan (University of Exeter)

In UK higher education, foundation-level education is targeted at Foreign/Second language (L2) learners who do not meet the entry criteria for undergraduate programmes regarding their English proficiency or formal qualifications. Foundation Programmes include both English for Academic Purposes (EAP) skills and discipline-specific content modules. These play a crucial role in English Medium Instruction contexts as they provide pathways to undergraduate study by helping students develop academic literacy skills. Successful learning on such programmes could be facilitated by an understanding of the specific genres that students are required to write. However, despite extensive research into the genre effect in university-level academic writing, language use across genres in foundation-level academic writing remains under-researched. This study addresses the gap by investigating linguistic features across genres in foundation-level students’ assessed writing. To fulfil this, my research questions are (1) what genres are found in foundation-level writing? (2) how do these genres differ in their typical lexical and grammatical features? (3) what is the relationship between writing in EAP skills and discipline-specific content modules?

This poster presentation will discuss the categorisation of genres and measures of vocabulary in this project. Summative assignments written by foundation-level students in a British institution will be collected to build a learner corpus. Adopting Nesi and Gardner’s (2012) taxonomy, learner texts will be classified into genres based on social purposes and stages. For measures of vocabulary, the diversity of academic vocabulary and word frequency across genres will be analysed as the main lexical features. The study will provide knowledge of genre classification and differences in lexical sophistication across genres in foundation-level academic writing. This will have pedagogical implications for foundation programmes catering to diverse student populations and contribute to our understanding of the effects of genres on linguistic features in written production.

Reference:

Nesi, H., & Gardner, S. (2012). Genres across the disciplines: Student writing in higher education. Cambridge University Press.

Structural Characteristics of L3 Norwegian as Spoken by Polish Native Speakers

Garbacz, Piotr (University of Oslo)

Over the past two decades, Poles have emerged as the predominant immigrant group in Norway, with 107,442 Polish nationals recorded as immigrants by 2023 (according to Statistics Norway). They typically work in the construction and service industries, but also as qualified experts such as doctors and scientific researchers. As EU citizens, Polish immigrants do not qualify for free Norwegian language training, leading many to self-study or enroll in paid language courses.

While the use of Norwegian among Polish construction workers has been explored by studies from Kraft (2017) and Urbanik (2021), a comprehensive structural analysis of L3 Norwegian as spoken by Polish natives is absent, aside from a limited study by Biskup (1988) and Skommer's (2014) examination of phonological challenges faced by Polish learners of Norwegian. The presentation will condense a description based on the NORINT corpus (https://www.hf.uio.no/iln/om/organisaison/tekstlab/prosjekter/norint/, limited to the speech part of the corpus) and data from the NorPol: L2 Communication among Polish Migrants in Norway project (http://www.hf.uio.no/multiling/prosjekter/flaggskipprosjekter/norpol/), funded by the Norwegian Research Council between 2020 and 2024.

This presentation aims to outline phonological, morphological, and syntactic elements of L3 Norwegian as used by Polish L1 speakers. Findings show considerable phonological transfer, while morphological and syntactic variations are more aligned with patterns seen in Norwegian spoken by LX speakers, regardless of their first language, which raises questions about the specific nature of morphological and syntactic transfer, cf. Odlin (2022:45–64).

Refrences:

Biskup, Zofia. 1988. «Noen grunnleggende forskjeller mellom polsk. En kontrastiv minigrammatikk.» NOA norsk som andrespråk 1988 (6): 1–22.

Kraft, Kamilla. 2017. Constructing migrant workers: Multilingualism and communication in the transnational construction site. PhD-thesis. University of Oslo.

Odlin, Terence. 2022. Explorations of Language Transfer. Second Language Acquisition vol. 144. Bristol – Jackson: Multilingual Matters.

Skommer, Grzegorz. 2014. «Norsk fonologi – en utfordring for polske innlærere?» NOA norsk som andrespråk 30 (2): 67–85.

Urbanik, Pawel Kazimierz. 2021. ”Directives in the construction site: Grammatical design and work phases in second language interactions with crane operators.” Journal of Pragmatics, vol. 178, pp. 43–67.

Productive use of derivational affixes across proficiency levels

Gee, Roger W. (Holy Family University, USA), M. Karen Jogan (Albright College, USA), Kathleen S. Jogan (University of Arkansas, USA)

Previous research has found that English L2 learners have limited knowledge of derivational affixes (Iwaizumi & Webb, 2021), that affix difficulty varies by L1 (Stewart et al., 2023), that there are few studies of productive use (Stewart et al.), and no single study has focused on the effect of L2 proficiency through all levels, A1 through C2. This research investigates the development of derivational affixes in writing by Spanish L1 EFL students ranging from A1 through C2.

The texts, from COREFL (Lozano et al., 2020), were produced in response to a consistent prompt without cues or instructions for affixes. The corpus of 436 texts contained 85,856 words. Affixes were identified with Morpholex (https://www.lextutor.ca/), dividing affixes into levels using Bauer and Nation's (1993) frequently cited scheme, but the data was aggregated without regard to affix levels.

A multinomial test showed a significant difference between the observed and expected use of affixes at different CEFR levels: $\chi 2$ (5) = 58.15, $p<.001$. Visual inspection of charts, after words and affixes at each proficiency level were normalized per 100 words, suggested that the number of derived tokens steadily increased after the A2 level, with greatest growth from C1 to C2; that the number of types produced showed greatest growth from C1 to C2; and little increase in different types used after the B1 level. Overall, productive affix use was developmental, with greatest increases in tokens and types beginning with C1 and few new types added after B1.

There are at least two limitations to this study. First, the results must be considered with caution as Spanish was the only L1. Further research with non-Romance languages is needed. Also, there was no determination of what frequent words might have been used as wholes, rather than decomposable words. This too is an area for future research.

References

Bauer, L., & Nation, P. (1993). Word families. International journal of Lexicography, 6(4), 253-279.

Iwaizumi, E., & Webb, S. (2021). To what extent do learner-and word-related variables affect production of derivatives? Language Learning, 73(1), 301-336.

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 (pp. 21-46). Peter Lang. https://doi.org/10.3726/978-3-653-05182-7

Stewart, J., Brown, D., Bennett, P., Robles-García, P., Sánchez-Gutiérrez, C. H., Miguel, N. M., & McLean, S. (2023). The contribution of affixes to productive English vocabulary knowledge for Chinese, German and Spanish learners: A comparison. System, 115, 103035.

Cambridge-Belgrade (CamBel) Persian Learner Corpus: Design and Methodology

Ghaffari, Mahbod (University of Cambridge) and Saeed Safari (University of Belgrade)

Although Persian is categorised as a Less Commonly Taught Language (LCTL), in recent years, more and more attention has been paid to the academic teaching of Persian. However the creation and development of Persian learner corpora, as a main research tool for language instructors and developers of teaching materials has not received sufficient attention. This paper addresses the growing need for comprehensive resources in Persian language instruction by introducing the design and methodology of Cambridge-Belgrade (CamBel) Persian learner corpus. CamBel is an error-marked (coded) learner corpus jointly compiled, recorded, and administered as part of an academic research partnership between the Persian Studies teams at the University of Cambridge and the University of Belgrade. Being a corpus of Persian written texts produced by learners of Persian, CamBel adopts an "error-tagged" approach and applies a specific methodology for error annotation that combines linguistic categories, surface structure taxonomies, and error types. In this study, we try to address the following questions:

What challenges and problems arise in designing and creating the Persian learner corpus?

How is data representativeness determined? and

What methods are employed to assess the proficiency levels of Persian learners with diverse language backgrounds?

Additionally, considering the diversity of errors and the grammatical features of Persian, we also investigate the basis for developing an error tagset that comprehensively captures common and frequent errors. On the other hand, While the development of CamBel and the analysis of errors holds considerable potential for language teaching research, the limitations and challenges associated with this approach need to be recognised. These include determining the type of the production content, assessing language proficiency level, appropriate error tagging, and ensuring the representativeness of collected data. This study investigates these issues in depth and also proposes effective solutions to establish solid standards for them. CamBel provides a range of reports that allow for a thorough examination of the frequency and types of errors learners make. Therefore, these analyses contribute to understanding Persian interlanguage at different CEFR levels, advancing theoretical insights into language learning development, and facilitating linguistic error analysis through comparative studies. Most importantly, they can identify the challenges learners face during their learning process, as detailed in the paper.

References

Bennett, G. R. (2010). Using corpora in the language learning classroom: Corpus linguistics for teachers. University of Michigan Press.

Callies, M. (2015). Learner corpus methodology. In F. Meunier, G. Gilquin, & S. Granger (Eds.), The Cambridge Handbook of Learner Corpus Research (pp. 35–56). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.003

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

Dulay, H. C., Burt, M. K., & Kreshen, S. (1982). Language Two (p 150-160). New York: Oxford University Press.

Ellis, R. (2008). The study of second language acquisition. Oxford University Press.

Ghaffari, M. (2020). Persian as an interlanguage. In P. Shabani Jadidi, The Routlege Handbook of Second Language Acquisition and Pedagogy of Persian (1st ed., pp. 546-566). London: Routledge.

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 (1st ed., pp. 9–34). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.002

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), Language Learning & Language Teaching (Vol. 6, pp. 3–33). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.6.04gra

Granger, S. (2008a). Learner corpora in foreign language education. In N. H. Hornberger (Ed.), Encyclopedia of Language and Education (pp. 1427–1441). Springer US. https://doi.org/10.1007/978-0-387-30424-3_109

Granger, S. (2008b). Learner Corpora. In Lüdeling, A. & Kytö, M. (eds.), Corpus Linguistics. An International Handbook, 259–275. Berlin & New York, NY: Walter de Gruyter.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), Studies in Corpus Linguistics (Vol. 33, pp. 13–332). John Benjamins Publishing Company. https://doi.org/10.1075/scl.33.04gra

Granger, S. (2013). Error-tagged learner corpora and CALL: A promising synergy. CALICO Journal, 20(3), 465–480. https://doi.org/10.1558/cj.v20i3.465-480

Granger, S. (2014). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger, Learner English on Computer (1st ed., pp. 3–18). Routledge. https://doi.org/10.4324/9781315841342-1

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: Learner corpus research – past, present and future. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The Cambridge Handbook of Learner Corpus Research (1st ed., pp. 1–6). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.001

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. https://doi.org/10.1075/ijlcr.20001.lar

Le Bruyn, B., & Paquot, M. (Eds.). (2020). Learner corpus research meets second language acquisition (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108674577

Lee, N. (1990). Notions of "Error" and Appropriate Corrective Treatment, Hong Kong Papers in Linguistics and Language Teaching, 14, pp. 55-70.

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and sla: Employing technology to analyze language use. Annual Review of Applied Linguistics, 39, 74–92. https://doi.org/10.1017/S0267190519000096

McEnery, T., & Hardie, A. (2011). Corpus linguistics: Method, theory and practice (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-Based Language Studies: An Advanced Resource Book. London: Routledge.

Nesselhauf, N. (2004). Learner corpora: learner corpora and their potential for language teaching. In J. McH. Sinclair (Ed.), Studies in Corpus Linguistics (Vol. 12, pp. 125–152). John Benjamins Publishing Company. https://doi.org/10.1075/scl.12.11nes

Richards, J. C. (1971). A non-contrastive approach to error analysis1. ELT Journal, XXV(3), 204–219. https://doi.org/10.1093/elt/XXV.3.204

Safari, S. (2018). The salam Farsi learner corpus-Introducing the error tagging system. Anali Filološkog Fakulteta, 30(2), 249–263. https://doi.org/10.18485/analiff.2018.30.2.13

Saville-Troike, M. (2006). Introducing second language acquisition. Cambridge University Press.

Tono, Y. (2013). Using learner corpora for L2 lexicography: Information on collocational errors for EFL learners. Lexikos, 6(1). https://doi.org/10.5788/6-1-1028

DiverSIta: a new corpus for the documentation of L2 Italian

Goria, Eugenio (University of Turin) and Caterina Mauri (University of Bologna)

This paper is dedicated to the presentation of the corpus design and methodology underlying the implementation of DiverSIta, a new module within the Kiparla corpus of spoken Italian (Mauri et al. 2019), which will be specifically dedicated to mobile speakers with a plurilingual repertoire.

While the first release of the corpus only included typical L1 speakers of Italian, the new data collection will involve speakers of Italian with a migratory background; these may include: (i) individuals who were born in other countries, who have foreign languages as their L1s, typically arrived in Italy as adults, and speakers of an L2 variety of Italian; and (ii) individuals who were born in Italy from migrant families, and who have Italian as (one of their) their L1(s), but also have some degree of bilingualism involving other languages present in the family. In this respect, the DiverSIta corpus qualifies as a learner corpus, in that at least part of the interactions under collection involve adult learners of Italian who typically acquire Italian during adulthood and after the critical periods.

Data collection will be carried out throughout 2024, in the cities of Torino and Bologna, and will rely on a sampling based on the languages (and therefore communities) that are most represented in each city. These are Arabic, Spanish, Chinese and Romanian for Torino, and Arabic, Bangla, Chinese and Ucrainian for Bologna.

By collecting naturalistic data (interviews and casual conversation), we will be able not only to carry out corpus-based studies on how these learner varieties of Italian are structured, but also to achieve a better insight on the type of linguistic practices in which these varieties are used.

Reference

Mauri, C., Ballarè, S., Goria, E., Cerruti, M., & Suriano, F. (2019). KIParla Corpus: A New Resource for Spoken Italian. Proceedings of the Sixth Italian Conference on Computational Linguistics. Bari, Italy, November 13-15, 2019.

Syntactic Complexity Development in Intermediate learner English: A longitudinal pilot study

Götz-Lehmann, Sandra; Philine Metzger and Fabian Kettenhofen (Philipps University Marburg)

Syntactic complexity has featured prominently in Second Language Acquisition research over the last few decades (cf. Larsen-Freeman 2009). Recent developments of tools that can automatically extract a large number of complexity measures (e.g. the Tool for Automatic Analysis of Lexical Sophistication; Kyle & Crossley 2015) have led to very detailed descriptions of L2 English complexity development (e.g. Lu 2010; Kyle, Crossley & Verspoor 2021). Broadly, we can assume a steadily increasing level of complexity with an increase in learners' proficiency levels, although studies typically report on large degree of variation, so that generalizations are often hard to make. Additionally, truly longitudinal corpus-based studies tracing the complexity development of intermediate learners of English remain very rare (cf., however, Kyle, Crossley & Verspoor 2021). Studies that are complemented by teacher assessments, have – to the best of our knowledge – not been conducted yet.

Against this background, in the proposed paper, we would like to present the findings of a study that investigates how syntactic complexity develops in intermediate German learners of English over four school years while comparing our findings to teachers' assessments of the learner texts to check if they correlate or deviate. More specifically, the proposed paper addresses the following research questions:

1. (How) does syntactic complexity develop in written L2 English from grade 9 to grade 12?
2. Do learning context variables have an effect on the development of syntactic complexity of intermediate written L2 English?
3. Are quantitative assessments of syntactic complexity in line with teachers' assessments of learner writing?

In order to answer these research questions, we will analyze a subset of the longitudinal Marburg Corpus of Intermediate Learner English (MILE; Kreyer 2015), consisting of written learner data by 90 intermediate learners of English between grade 9 and grade 12, totaling 1,080 texts and more than 500,000 words. In our proposed paper, we zoom in closely on 5 learners' developments over 4 years, who submitted 4 texts each year (i.e. 80 essays). These texts were first subjected to an automatic analysis of Lu's (2010) 14 syntactic complexity parameters using the TAASSC tool (e.g. mean length of T-unit, dependent clauses per T-unit, etc.), manually checked for accuracy Châu & Bulté 2022) and then subjected to a statistical data analysis using mixed effects regression modelling. One first look into the data suggests that some global complexity variables appear to be robust predictors to discriminate the grade levels (e.g. the mean length of T-units; cf. also Larsen-Freeman 1978), whereas other variables did not have stable effects (e.g. T-units per sentence). Evaluations of the teacher ratings are largely in line with these findings, however, the assessments also revealed some striking differences, which will be discussed in terms of their language-pedagogical implications.

References:

Châu, Q. H. & B. Bulté (2022). A comparison of automated and manual analyses of syntactic complexity in L2 English writing. International Journal of Corpus Linguistics 28, 232–262.

Kreyer, R. (2015). The Marburg Corpus of Intermediate Learner English (MILE). In Learner Corpora in Language Testing and Assessment, M. Callies & S. Götz, eds. John Benjamins. 13-34.

Kyle K. & S. A. Crossley & M. Verspoor. 2021. Measuring longitudinal writing development using indices of syntactic complexity and sophistication. Studies in Second Language Acquisition 43, 781–812.

Larsen-Freeman, D. (1978). An ESL index of development. TESOL Quarterly, 12, 439–448.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in Second Language Acquisition. Applied Linguistics 30(4): 579–589.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15, 474–496.

Introducing ETC, the first learner corpus of non-native teacher English

Gráf, Tomáš (Department of English Language and ELT Methodology, Faculty of Arts, Charles University, Prague), Barbora Bulantová (Department of English Language and ELT Methodology, Faculty of Arts, Charles University, Prague) and Kryštof Buchal (Institute of Phonetics, Faculty of Arts, Charles University, Prague)

The presentation aims to introduce a unique and a thoroughly innovative project of a compilation of a corpus of L2 English-teacher language, as a specific variety of learner language. The project is unique in that no corpus of spontaneous L2 English-teacher language is yet available, although corpora of classroom interaction do exist. Secondly, the compilation had a strong pedagogical bias in that it involved a collaborative effort of seasoned learner-corpus experts and a group of BA teacher-trainees who helped design, record, transcribe and align the corpus, and consequently gained a sound insight into the principles of learner corpus linguistics, reliable data collection and the use of AI for the processing of data. The other pedagogical aspect which makes this the project unique is that the speech tasks in the corpus focused on a variety of aspects of the ELT profession thus increasing the trainees' awareness of the profession from the perspective of experienced teachers. This also makes the collected data open for content analysis using corpus-linguistic techniques.

The resulting corpus (25 non-native- and 15 native-speaker teachers, c. 100,000 tokens, 12.5 hours of recorded text) paves way for systematic analysis of teacher language and teacher language proficiency, bearing in mind that teacher language is one of the most important sources of input for language learners. Besides, it shows how the individual speech tasks can be designed to match a specific context while still encouraging spontaneous language production, and how AI can be used to streamline corpus-compilation subprocesses.

Besides describing the sophisticated process of the story of this first-of-its-kind corpus we will also introduce the first results of teacher-language analysis based on our corpus data allowing insights into teacher-language fluency and accuracy, showing that this particular variety of learner language ought to receive more attention.

References

Chambless, K. S. (2012). Teachers' Oral Proficiency in the Target Language: Research on Its Role in Language Teaching and Learning. Foreign Language Annals, 45(s1). https://doi.org/10.1111/j.1944-9720.2012.01183.x

Choi, E., & Lee, J. (2016). Investigating the relationship of target language proficiency and self-efficacy among nonnative EFL teachers. System, 58, 49–63. https://doi.org/10.1016/j.system.2016.02.010

Faez, F., & Karas, M. (2017). Connecting Language Proficiency to (Self-Reported) Teaching Ability: A Review and Analysis of Research. RELC Journal, 48(1), 135–151. https://doi.org/10.1177/0033688217694755

Faez, F., Karas, M., & Uchihara, T. (2021). Connecting language proficiency to teaching ability: A meta-analysis. Language Teaching Research, 25(5), 754–777. https://doi.org/10.1177/1362168819868667

Freeman, D. (2017). The Case for Teachers' Classroom English Proficiency. RELC Journal, 48(1), 31–52. https://doi.org/10.1177/0033688217691073

Freeman, D., Katz, A., Garcia Gomez, P., & Burns, A. (2015). English-for-Teaching: Rethinking teacher proficiency in the classroom. ELT Journal, 69(2), 129–139. https://doi.org/10.1093/elt/ccu074

Richards, J. C. (2010). Competence and Performance in Language Teaching. RELC Journal, 41(2), 101–122. https://doi.org/10.1177/0033688210372953

Richards, J. C. (2017). Teaching English through English: Proficiency, Pedagogy and Performance. RELC Journal, 48(1), 7–30. https://doi.org/10.1177/0033688217690059

Thi Hong Nhung, P. (2018). General English Proficiency or English for Teaching? The Preferences of In-service Teachers. RELC Journal, 49(3), 339–352. https://doi.org/10.1177/0033688217691446

Tsang, A. (2017). EFL/ESL Teachers' General Language Proficiency and Learners' Engagement. RELC Journal, 48(1), 99–113. https://doi.org/10.1177/0033688217690060

Opening Pandora’s box: Is learner corpus research inclusive enough?

Gilquin, Gaëtanelle UCLouvain

Following the LCR 2024 organizers’ call to “champion corpus research involving learning of smaller languages and by learners with a small L1, as well as language learning in plurilingual settings”,[1] this presentation aims to move the debate beyond the L1s and L2s studied in learner corpus research to ask whether the field as a whole is inclusive enough. In doing so, it also seeks to contribute to the broader trend of discussing and promoting inclusion in linguistics, a trend that has been growing lately (see, e.g., Charity Hudley et al. 2024).

I will start by showing how learner corpus research has been inclusive in some respects and how recent developments in the field have led to enhanced inclusion, for example through the consideration of a wider range of languages or the adoption of a “Diversity and Inclusion Statement” by the Learner Corpus Association.[2] I will then review different aspects of learner corpus research for which some of our practices may have been less than optimal in fostering inclusion. These aspects will include the learner populations investigated, the choice of norms in learner corpus studies (see also Gilquin 2022), the terminology used in the field and the pedagogical applications based on learner corpora. For each of these aspects, the importance of more inclusive approaches will be underlined, with examples taken from the literature, and suggestions will be made on how to further enhance inclusion. To conclude, it will be emphasized that developing inclusive learner corpus research is a slow process, but one to which we should devote the attention it deserves, both as individuals and as a community.

References

Charity Hudley, Anne H., Mallinson, Christine & Bucholtz, Mary. 2024. Inclusion in Linguistics. Oxford: Oxford University Press.

Gilquin, Gaëtanelle. 2022. ‘One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching’. Language Teaching 55(1): 87-99.

---

[1] https://lcr2024.ut.ee/ (last accessed on April 14, 2024).

[2] https://www.learnercorpusassociation.org/wp-content/uploads/2022/05/LCA-Diversity-and-Inclusion-and-Code-of-Conduct-statements.pdf (last accessed on April 14, 2024).

Of-constructions in the interlanguage of Norwegian learners of English. A pseudo-longitudinal study.

Hasselgård, Hilde (University of Oslo)

Prepositions are notoriously difficult to learn in an L2. The preposition of may be especially hard because its uses are mainly abstract. Unlike other prepositions, of rarely forms phrases functioning as adjuncts, but more typically "combines with preceding nouns" to form complex NPs (Sinclair 1991: 83; cf. Hunston 2008). Norwegian lacks a functionally similar preposition (Author A), and Norwegian advanced learners have been found to use the framework "the N of the N" in a qualitatively different way from native speakers of English (Author B). In particular, partitive constructions were overrepresented (e.g. the end of the story) while nominalizations were underrepresented (e.g. the unraveling of the truth).

This study focuses on intermediate learners of English and to of-constructions in general. The material comes from the English part of the longitudinal TRAWL corpus, which contains texts by pupils in Norwegian schools. 200 instances of of were extracted randomly from years 8, 10 and 11 (ages c. 13, 15 and 16) to answer the following research questions:

1. In what lexicogrammatical constructions do learners use of?
2. Do the patterns of of change qualitatively across the school years?

The youngest learners are expected to use of in relatively simple constructions and formulaic phrases, e.g., a lot of, kind of, take care of, while open-choice complex NPs may become more common as the learners advance. TRAWL is compared to the Norwegian component of the ICLE corpus, which represents a higher study level than TRAWL.

A first analysis of ICLE-NO and years 8 and 10 in TRAWL seems to support the hypotheses. Set quantifying and partitive phrases (a lot of, some of the) and complex prepositions (in front of, instead of) constitute larger proportions in TRAWL than in ICLE, while ICLE has a larger share of complex NPs. More of the younger learners' of-constructions denote measurements and part-whole relationships, while possession and support (e.g. a matter of opinion; Sinclair 1991) are more prevalent in ICLE-NO.

References

Author A
Author B
Hunston, S. 2008. "Starting with the small words. Patterns, lexis and semantic sequences". International Journal of Corpus Linguistics 13:3, 271–295.
Sinclair, J. 1991. Corpus, Concordance, Collocation. Oxford University Press.
The International Corpus of Learner English, Norwegian component (ICLE-NO), see https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html
The TRAWL Corpus - Tracking Written Learner Language, see https://tekstlab.uio.no/trawl/

The impact of topic on the use of lexical bundles by EFL and ESL learners

Huang, Lingmin (University of Louvain)

Recent research has demonstrated that the use of lexical bundles shows differences across registers (e.g., Biber et al., 1999, 2004; Gablasova et al., 2017), disciplines (e.g., Cortes, 2004; Hyland, 2008; Durant, 2017), genres (Author, 2023) and task types (Kyle & Crossley, 2016). However, little attention has been given to the impact of topic, with some rare exceptions such as Li & Yao's (2023) study which reported that topic significantly influences bigram and trigram use. Additionally, the impact of topic on n-gram use moderated by education settings and proficiency levels also remains largely unexplored.

Against this background, this study aims to explore the topic effect on trigram frequency and trigram mutual information (MI) in English writing by taking education settings (EFL vs ESL) and proficiency levels into account. We analyzed a dataset of 5,200 written essays totaling c. 1.2 million words from the International Corpus Network of Asian Learners of English (ICNALE: Ishikawa, 2023) on two distinct topics (non-smoking and part-time job). The dataset includes 3,800 texts by 1,900 EFL learners, and 1,400 texts by 700 ESL learners with proficiency levels of A2, B1_1, B1_2 and B2+ (revised CEFR scale).

Trigram frequency and trigram MI indices were extracted from the Corpus of Contemporary American English via the Tool for Automated Analysis of Lexical Sophistication (TAALES: Kyle et al., 2018). Linear mixed-effects models were employed by using lme4 package in R (Bates et al., 2015). The results showed that essays on the non-smoking topic had significant higher trigram frequency than those on the part-time job topic and the topic differences were moderated by proficiency levels and education settings. Furthermore, the results revealed that essays addressing the non-smoking topic had significantly lower trigram MI in comparison to those on the part-time job topic and their differences were moderated only by learners' proficiency levels.

References

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. Applied Linguistics, 25(3), 371-405.

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

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. English for Specific Purposes, 23(4), 397-423.

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. Applied Linguistics, 38(2), 165-193

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. Language Learning, 67(S1), 155-179.

Author (2023). Revisiting the relationships of n-gram measures to L2 writing proficiency: Comparisons between genres and connections to vocabulary levels. System, 118, 103136.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), 4-21.

Ishikawa, S. (2023). The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English. Routledge.

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

Kyle, K., Crossley, S. & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. Behavior Research Methods Res 50, 1030-1046.

Li, H., & Yao, Y. (2023). Formulaic competence in college-level Asian English learner's argumentative writing: Examining the effects of language background and topic. The Asia-Pacific Education Researcher, 32(6), 793-803.

Collocations with the verbs have, make/do, give and get in L2 Czech: how corpus research can inform CEFR descriptions of Czech at levels A2-B2

Hudousková, Andrea (Faculty of Arts, Charles University, Prague)

The verbs have, make/do, give and get are among the twenty most common verbs in Czech. They are also frequently used by L2 Czech learners. The problem with the use of these verbs in the learners' language is that they occur repeatedly in quite a small number of collocations, which remains rather limited even at higher proficiency levels and lags far behind the use of native speakers (Nesselhauf, 2005). There are two reasons for this state of affairs: first, L2 learners tend to rely on 'safe', well-known collocations (Hasselgård, 2019); second, although these verbs are encountered early in L2 instruction, they are neglected later on (Granger & Altenberg, 2001).

The aim of the talk is to present the development of collocations with the above-mentioned verbs throughout the levels A2-B2, based on data from two available Czech learner corpora, i.e. CzeSL and MERLIN. The results are compared 1) with glossaries in referential descriptions of L2 Czech (Čadská et al., 2005; Šára et al., 2001; Holub et al., 2005), which contain lexical items and collocations that learners at the proficiency levels A2-B2 are supposed to be familiar with, and 2) with the lists of collocations from the most recent representative corpus of native Czech, SYN2020.

It is argued that corpus research can contribute to the improvement of referential descriptions of L2 Czech that had been published before the L2 Czech corpora appeared (Boyd et al., 2014). At the A2-B1 level, data from learner corpora can be used to extend and adapt glossaries to better match the level requirements and learners' communicative needs. At the B2 level, on the other hand, learner corpora do not provide sufficient guidance, so the synchronic corpus of native Czech should be also used as a source for informed revision and elaboration of the referential descriptions.

References:

Altenberg, B., and Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. Applied Linguistics, 22(2), 173–195.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B. and Vettori, C. (2014). The MERLIN corpus: Learner language and CEFR. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 1281–1288. Reykjavik: European Language Resources Association (ELRA).

Čadská, M., Bidlas, V., Confortiová, H., and Turzíková, M. (2005). Čeština jako cizí jazyk. Úroveň A2. Praha: Tauris.

Hasselgård, H. (2019). Phraseological teddy bears. Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In Wiegand, V., and Mahlberg, M. (Eds.) Corpus Linguistics, Context and Culture, pp. 339–362. Berlin / Boston: Walter de Gruyter.

Holub, J., Adamovičová, A., Bischofová, J., Cvejnová, J., Gladkova, H., Hasil, J., Hrdlička, M., Mareš, P., Nekvapil, J., Palková, Z., Šára, M. (2005). Čeština jako cizí jazyk. Úroveň B2. Praha: Tauris.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., Škrabal, M. (2020). SYN2020: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK, Praha. http://www.korpus.cz

MERLIN – Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context. http://merlin-platform.eu

Nesselhauf, N. (2005). Collocations in a Learner Corpus. Amsterdam / Philadelphia, John Benjamins Publishing.

Šára, M., Bischofová, J., Confortiová, H., Cvejnová, J., Čadská, M., Holub, J., Lánská, L., Palková, Z., and Turzíková, M. (2001). Prahová úroveň – čeština jako cizí jazyk. Strasbourg: Council of Europe.

Šebesta, K., Bedřichová, Z., Šormová, K., Štindlová, B., Hrdlička, M., Hrdličková, T., Hana, J., Petkevič, V., Jelínek, T., Škodová, S., Poláčková, M., Janeš, P., Lundáková, K., Skoumalová, H., Sládek, Š., Pierscieniak, P., Toufarová, D., Richter, M., Straka, M., Rosen, A. (2014).

Considerations for planning and developing publicly-shared L2 speech corpora

Huensch, Amanda (University of Pittsburgh)

Spoken corpora are less common than written corpora, and within existing spoken corpora, L2 phonological corpora (i.e., those that that include audio/video data with time-aligned annotation of a phonological feature) are quite rare (Gut, 2014). Yet, such corpora are critical to answer questions related to the acquisition of L2 phonological skills. Therefore, when developing speech corpora, researchers should plan to publicly share data and protocols to help broaden impact (Myles, 2015). During this talk, lessons learned from multiple projects involving the collection of publicly-shared L2 speech corpora are presented. These projects all incorporated a commitment to open science initiatives and represent a breadth of research questions and experimental designs: (a) an 11-year longitudinal investigation of L2 attrition, maintenance, and development and (b) a multi-site project examining intelligibility, comprehensibility, and accentedness in L2 Spanish, and (c) a cross-sectional investigation of the relationships among L2 speech perception/production and general cognitive skills, specifically inhibitory control.

Two key takeaways are discussed in this presentation: (a) the benefits of preregistration (i.e., placing a time-stamped research plan on a public repository) and (b) the need for detailed research protocols and data management plans. The piloting of protocols for all levels of a project (participant recruitment, data collection, data processing, data coding, etc.) supports project efficiency and success. Data management plans document key questions about types of data shared (e.g., sound files, transcriptions, annotations), how and where they are stored (e.g., public or institutional repositories), and the types of documentation (e.g., metadata) provided. Planning ahead with decisions as specific as how to name data files or what the structure of data directories should be to larger-scale considerations such as developing training plans for research members across sites increases feasibility and project success. Examples of protocols, data management plans, and preregistration elements from prior projects are provided.

References

Gut, U. (2014). Corpus phonology and second language acquisition. In J. Durand, U. Gut, & G Kristofferson (Eds.), The Oxford handbook of corpus phonology. Oxford University Press.

Mitchell, R. Tracy-Ventura, N., & McManus, K. (2017). Anglophone students abroad: Identity, social relationships and language learning. Routledge.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The Cambridge handbook of learner corpus research (pp. 309–332). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.014

Cross-linguistic influences in different levels of granularity: How are they different and why does it matter for LCR?

Ivaska, Ilmari (University of Turku)

Cross-linguistic influences (CLIs) between an individual's second language (L2) and their first language (L1) have been one of the most central topics within learner corpus research. While the methodological approaches to addressing CLIs have been discussed extensively, yielding widely used comparable methodological frameworks (Jarvis 2000; Jarvis 2010; see also Jarvis & Crossley 2012; Jarvis & Paquot 2015), the conceptual nature of these influences has received less attention. Building in part on the work of Håkan Ringbom (2007), this presentation discusses construction level CLIs and system level CLIs as profoundly different types of phenomena. Consider examples (1)–(4) that all express more or less the same meaning ‘you run in the forest’.

(1) *(Sinä)* *juokse-t* *metsä-ssä* (Finnish)
you run-SG2 forest-in

(2) *(Sa)* *jookse-d* *metsa-s* (Estonian)
you run-SG2 forest-in

(3) *(Tu)* *corr-i* *ne-l* *bosco* (Italian)
you run-SG2 in-DEF forest

(4) *Tu* *cour-s* *dans* *la* *forêt* (French)
you run-SG2 in DEF forest

The four languages belong genealogically to two different genera: Finnish and Estonian in the Finnic genus of the Uralic language family, and Italian and French in the Romance genus of the Indo-European language family, respectively. The linguistic subsystems that underlie the examples follow this categorization in many ways both lexically (e.g. *juokset* and *jooksed* as opposed to *corri* and *cours*) and grammatically (e.g. locative expression using case endings as opposed to prepositions). In some ways, however, these languages group differently: the overtly expressed grammatical subject (*sinä*, *sa*, *tu*) is obligatory only in French. While both system level and construction level CLIs have been widely accepted and extensively studied within learner corpus research, there has so far been only little discussion on their relationship, and even attempts to systematically tease the two apart are few and far between (however, see Murakami 2016). Crucially to theorizing the generalizability of CLIs, the two types of CLIs differ in that the system level is typically an open-ended categorization with a large or unknown number of possible values, as opposed to the construction level, which is typically a closed-class categorization with only few possible values.

In this paper, I argue that the system level similarities and differences (in the above example, Finnish and Estonian in contrast to Italian and French), and those specific to individual constructions (in the above example, the obligatoriness of the overt subject marking) can both induce cross-linguistic influences. What is more, I argue that they can combine in multiple ways: so that system level CLIs take place while construction level CLIs do not, so that construction level CLIs take place and system level do not, and so that they both play a role. I propose that learner corpus studies addressing CLIs should account for the difference in the type of CLI, and that this should optimally be visible both in the data and in the method. I will address this argument from the point-of-view of Finnish as an L2, and looking at several different linguistic phenomena (e.g. the above-described subject marking, case marking of grammatical objects, constructional strategies involved in tense expressions, as well as those involved in locative expressions). The data come from various corpora of L2 Finnish (ICLFI, see Jantunen 2011; LAS2, see Ivaska 2014; TOPLING, see University of Jyväskylä 2016) to allow for comparing L2 learners of

Finnish with a range of both typologically and genealogically diverging L1s. I will also sketch some methodological options for capturing this fundamental difference of different types of CLIs, but would like to invite the LCR community to take up on the task of finding new appropriate yet feasible solutions.

References

Ivaska, Ilmari. 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.

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

Jarvis, Scott. 2000. Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon. Language Learning 50(2). 245–309. https://doi.org/10.1111/0023-8333.00118.

Jarvis, Scott. 2010. Comparison-based and detection-based approaches to transfer research. EUROSLA Yearbook 10. 169–192.

Jarvis, Scott & Scott Crossley (eds.). 2012. Approaching Language Transfer through Text Classification: Explorations in the detection-based approach. Bristol, Buffalo, Toronto: Multilingual Matters.

Jarvis, Scott & Magali Paquot. 2015. Learner corpora and native language identification. In Sylviane Granger, Gaëtanelle Gilquin & FannyEditors Meunier (eds.), The Cambridge Handbook of Learner Corpus Research (Cambridge Handbooks in Language and Linguistics), 605–628. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.027.

Murakami, Akira. 2016. Modeling Systematicity and Individuality in Nonlinear Second Language Development: The Case of English Grammatical Morphemes. Language Learning 66(4). 834–871. https://doi.org/10.1111/lang.12166.

Ringbom, Håkan. 2007. Cross-linguistic Similarity in Foreign Language Learning. Clevedon: Multilingual Matters. https://doi.org/10.21832/9781853599361.

University of Jyväskylä. 2016. The Finnish Subcorpus of Topling - Paths in Second Language Acquisition. Data set. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2016111802.

Comparing morphosyntactic features in undergraduate dissertations in Spanish by Estonian and Spanish students: a corpus-driven approach

Izquierdo Alegría, Dámaso (ICS, University of Navarra, Spain)

Writing undergraduate or postgraduate dissertations in foreign languages is viewed as a challenge for many students (Dong 1998; Strauss 2012; Kaufhold 2015; Eriksson & Nordrum 2018; Ahern & Hernando 2020; Wu & Paltridge 2021). This seems more demanding when dissertations are written in L3/L4/L5+ languages, as in many curricula in modern languages different from English. This is the case of students of Spanish Language and Literature in Estonia, where most of them have no prior knowledge of Spanish when they gain admission to this degree (Kruse 2018).

Previous research on the features of the Spanish interlanguage of Estonian speakers is limited and focuses on particular linguistic areas, such as articles, temporal adverbials, and false friends (Kruse 2018,2020; Eller 2022; Rapún & Tramallino 2023), and the development of learner corpora in Spanish by Estonians is still in its infancy (ELEACTAR: Kruse 2018,2020; CLEAE: Rapún 2023).

The aim of this presentation is to uncover the most salient morphosyntactic features in dissertations in Spanish by Estonians from a corpus-driven perspective. Two corpora have been compiled: a corpus of 82 undergraduate dissertations in Spanish by Estonian students of Spanish Language and Literature at the University of Tartu, and a comparable corpus of 82 dissertations by native speakers in 5 Spanish universities. More specifically, the tags and n-grams with higher *keyness* scores are analysed in each corpus through the *keywords* function in *Sketch Engine*.

Some tags have been found to be underused (e.g. future tense verbs) or overused (e.g. present tense verbs) by Estonians probably because of the influence of their mother tongue, while others reveal that Estonians' linguistic strategies in Spanish for typical discursive functions in academic writing (introducing quotations, referring to other sections) are more limited. Results also suggest that the academic style of Estonians is less homogenous than that of native speakers.

References

Ahern, A. K., & Hernando, A. (2020). Los trabajos de fin de grado en español y en inglés. Retos, y un intento de mejora, de la alfabetización académica en formación inicial de profesorado. Tendencias pedagógicas, 36, 9-24.

Dong, Y. R. (1998). Non-native graduate students' thesis/dissertation writing in science: Self-reports by students and their advisors from two US institutions. English for Specific Purposes, 17(4), 369-390.

Eller, H. (2022). Adquisición/aprendizaje del artículo en español por hablantes. PhD dissertation. Universidad Complutense de Madrid (Spain).

Eriksson, A., & Nordrum, L. (2018). Unpacking challenges of data commentary writing in master's thesis projects: an insider perspective from Chemical Engineering. Research in Science & Technological Education, 36(4), 499-520.

Kaufhold, K. (2015). Conventions in postgraduate academic writing: European students' negotiations of prior writing experience at an English speaking university. Journal of English for Academic Purposes, 20, 125-134.

Kruse, M. (2018). La transferencia en personas plurilingües: los falsos amigos como un obstáculo y una oportunidad en la enseñanza y aprendizaje de lenguas extranjeras. PhD dissertation. University of Tartu (Estonia).

Kruse, M. (2020). Palabras cognadas en el vocabulario académico del inglés, español y estonio. E-Aesla, 6, 253-265.

Rapún, V. (2023). CLEAE: Corpus Longitudinal de Español de Aprendientes Estonios. Diseño y estudio piloto. In E. Álvarez García & M. Villayandre Llamazares (Eds.), Creatividad, innovación y diversidad en la enseñanza del español como LE/L2 (pp. 351-362). León: Universidad de León.

Rapún, V. & Tramallino, C. P. (2023). Acquisition of Spanish temporal adverbials by multilingual Estonian learners. Sustainable Multilingualism, 23(1), 63-90.

Strauss, P. (2012). 'The English is not the same': challenges in thesis writing for second language speakers of English. Teaching in Higher Education, 17(3), 283-293.

Wu, B., & Paltridge, B. (2021). Stance expressions in academic writing: A corpus-based comparison of Chinese students' MA dissertations and PhD theses. Lingua, 253, 103071

A Contrastive Analysis of Lithuanian Children’s Language

Juknevičienė, Rita (Vilnius University)

The last decades witnessed a wave of migration in Lithuania with families first moving to Western Europe and then, after years of life abroad, returning to their homeland. To children of such migrant families, it brought many language-related challenges. Having started school in the language of the new country, they continue schooling in Lithuania after their families return. Yet inadequate communicative competence in the Lithuanian language has been reportedly pointed out as one of the major obstacles to successful integration (Bagdonaitė 2020). This research was undertaken to shed more light on differences between L1 and L2 Lithuanian produced by young learners of 8-12 years and to highlight which aspects of linguistic competence require more attention in the teaching process.

The study was set up as a contrastive interlanguage analysis (Granger 2015), involving the corpus-driven approach to learner language, namely, the analysis of keywords (Kilgariff 2009) and 3-word lexical bundles (Biber et al. 2004). The data comes from the EMVAKA corpus (Bikelienė et al. 2022) which includes short written texts and spoken adult-child interactions based on a scripted scenario. The total corpus size is 72,287 words.

The analysis revealed several differences between L1 and L2 varieties of Lithuanian, with some indicating developmental interlanguage features and others pointing to different cultural experiences of the children. The keywords analysis showed that the most conspicuous differences lie in the frequencies of functional words, namely, pronouns and discourse markers. The analysis of lexical bundles revealed a statistically significant correlation between functional categories of lexical bundles and a variety of children’s languages. While the language of children who spent all their lives in Lithuania contains more referential expressions, children from migrant families, rather unexpectedly, demonstrate a considerably broader variety of stance expressions.

References

Bagdonaitė, J. 2020. Remigration in Lithuania in the 21st century: the readiness of the education system to accept students from returning families. Scientific Research in Education, 3: 1–15.

Biber, D., S. Conrad and V. Cortes. 2004. If You Look at … : Lexical Bundles in University Teaching and Textbooks. Applied Linguistics, 25(3): 371–405.

Bikelienė, L., Author, N. Poderienė, A. Tamulionienė. 2022. Grįžusių emigrantų vaikų kalba: kelios įžvalgos [Engl. The Language of Returning Emigrant Children: Preliminary Insights] Verbum 13, p. 4. doi: 10.15388/Verb.30.

Granger, S. 2015. Contrastive interlanguage analysis: A reappraisal. International Journal of Learner Corpus Research, 1(1): 7–24.

Kilgariff, A. 2009. Simple Maths for Keywords. In M. Mahlberg, V. González-Díazand and C. Smith (Eds) Proceedings of the Corpus Linguistics Conference, 20–23 July 2009. Liverpool: University of Liverpool.

The effect of self-initiated L2 activities on intermediate-level students' lexical complexity

Kaatari, Henrik (University of Gävle), Tove Larsson (Northern Arizona University), Ying Wang (Karlstad University), Pia Sundqvist (University of Oslo), Taehyeong Kim (Northern Arizona University)

Frequent engagement in extramural English (EE) activities (English-language activities that students engage in outside of the classroom) has been shown to positively influence students' L2 receptive and productive skills (e.g., Sundqvist, 2009, 2019). There are also indications in previous studies that the *type* of EE input students receive affects their production (Kaatari et al., 2023). Extending this line of research, we test the role of the type of input students receive through EE activities focusing specifically on their effect on lexical complexity. To do so, we look at junior and senior high school student writing in L2 English from the Swedish Learner English Corpus (SLEC; Kaatari et al., forthcoming). SLEC contains information about how many hours per week students engage in five EE activities: reading, watching, conversation, social media, and gaming. We use three types of psycholinguistic lexical sophistication measures (contextual distinctiveness, concreteness, and age of exposure), along with one measure of lexical diversity (moving average type-token ratio). Specifically, we build on previous research and ask the following research questions that also serve as our hypotheses:

1. Does frequent engagement with spoken input (conversation and watching) result in a higher degree of linguistic diversity than other types of EE exposure?
2. Does frequent engagement with longer written input (reading) result in a higher degree of linguistic sophistication, than other types of EE?

To test these specific hypotheses, we use Structural Equation Modeling (SEM; Larsson et al., 2021). Competing measured variable path analysis models were fitted, systematically testing our hypotheses. The best-fitting model (χ^2: 0.14, df: 20, CFI: 0.99, RMSEA: 0.033[0.00–0.064], SRMR: 0.067) confirmed both of our hypotheses. It thus seems crucial to avoid grouping EE activities together into a single category, and instead consider what type of input students are exposed to.

References

Kaatari, H., Wang, Y., & Larsson, T. Forthcoming. Introducing the Swedish Learner English Corpus: A corpus that enables investigations of the impact of extramural activities on L2 writing. Corpora, 19(1).

Kaatari, H., Larsson, T., Wang, Y., Acikara-Eickhoff, S., & Sundqvist, P. 2023. Exploring the effects of target-language extramural activities on students' written production. Journal of Second Language Writing, 62, 101062.

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.

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

Estonian L2 Learner Corpora: current state and perspectives

Kallas, Jelena (Institute of the Estonian Language), Kristjan Suluste (Institute of the Estonian Language), Raili Pool (Institute of the Estonian Language/University of Tartu), Helen Kaljumäe (Institute of the Estonian Language)

The paper aims to provide an overview of the Institute of the Estonian Language's (EKI) activities in gathering, maintaining, and publishing Estonian learner corpora. Until now, in Estonia, two corpus query systems for learner corpora—EMMA and ELLE—have been developed.

Starting from 2023, EKI is taking a leading role in the ESF project 'The Development of Estonian language teaching and learning', with one of the key activities involving the systematic gathering of learner corpora. The objective is to gather datasets, including text corpora and speech corpora, from target groups such as young and adult learners of Estonian as a Second Language. This involves adding both existing learner corpora from EMMA and collecting new data, particularly examination and test materials obtained from The Education and Youth Board, along with data from the L2 teaching program held at Tallinn University (Argus et al. 2023), into a unified system. The data will go through anonymization and pseudonymization processes before being stored in the language data repository called LADU. After processing, which includes lemmatization, morphology, syntax, and error annotation, the data will be available through the Corpus Query System KORP. The progress so far is that the repository is available for storing materials; however, there is a need for improvement in metadata and resource registration policy.

The existing learner corpora have played an essential role in developing EKI's learner dictionary portal, Sõnaveeb for Learners (Tavast et al. 2018), and Estonian L2 Teacher's Tools (Kallas et al. 2021; Üksik et al. 2021). In the second part of the presentation, we will showcase the application of learner corpora in Estonian L2 research, particularly focusing on the description of learners' vocabulary and grammar competence within Teacher's Tools. Learner data has been primarily used in creating wordlists for different proficiency levels and analyzing grammar acquisition.

References:

Argus, Reili; Baird, Piret; Meristo, Merilyn; Rüütmaa, Tiina 2023. Results of the assessment of children's linguistic development in the pilot project "Professional Estonian-language teacher in educational institutions". Limited sample assessment. Spring 2023 Tallinn.

EMMA https://korpused.keeleressursid.ee/emma/

ELLE https://evkk.tlu.ee/

ESF project https://portaal.eki.ee/keeleope.html

Kallas, Jelena; Pool, Raili; Üksik, Tiiu; Koppel, Kristina; Argus, Reili; Kerge, Krista; Bauer, Annika; Alp, Pilvi; Epner, Anu; Tsepelina, Katrin 2022. Eesti keele kui teise keele õpetaja tööriistad 2022. DOI: 10.15155/3-00-0000-0000-0000-08C04L

KORP http://korp.eki.ee

LADU https://console.object.hpc.ut.ee/login

Sõnaveeb for Learners https://sonaveeb/lite

Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018. Ed. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek. Ljubljana University Press, Faculty of Arts, 749−761.

Üksik, Tiiu; Kallas, Jelena; Koppel, Kristina; Tsepelina, Katrin; Pool, Raili (2021). Estonian as a Second Language Teacher's Tools. Proceedings of the Sixteenth Workshop on Innovative Use of NLP for Building Educational Applications, 130−134.

Pedagogical applications of learner corpus research: How far have we come?

Karlsen, Petter Hagen and Susan Nacey (Inland Norway University of Applied Sciences)

This paper presents a systematic review of the pedagogical applications of learner corpora as reported in the learner corpus research (LCR) literature between 2014-2023 – roughly the last decade. Our research question is the following: “What are the characteristics and trends of LCR over the last decade with respect to pedagogical applications?”

Since the inception of LCR in the late 1980s, the field has been held to have great potential for positive impact in the language classroom not only through affording a richer understanding of foreign language learning and development, but also through contributing towards both the enhancement of pedagogical materials and the creation of data-driven learning opportunities (see Meunier, 2016). Yet Granger (2015) found that “pedagogical ‘implications’ are much more numerous than ‘applications’” in the research literature (p.487), while Gilquin writes as late as 2023 that learner corpora “have not played a central role in learner teaching so far” (p. 283).

Our paper reviews all relevant LCR publications identified though a targeted search of research databases, together with corpus linguistics journals. We identify 487 peer-reviewed publications reporting empirical LCR findings, using their abstracts to code them for factors including corpus type (written, spoken, multimodal, cross-sectional, longitudinal, translation) and application type. Our four application types lie along a continuum of learners’ and/or teachers’ interaction with learner data: publications detailing 1) theoretical contributions alone (knowledge production), 2) pedagogical implications of knowledge production, 3) indirect applications of learner corpora (e.g. dictionary development), and 4) direct applications of learner corpora (e.g. classroom activities).

Our overall findings show a greater emphasis in LCR publications on knowledge production than on transfer of that knowledge into educational practice, similar to earlier observations. That said, we also find some exemplary pieces of research that represent intriguing avenues for classroom relevant LCR and future directions for the field.

References:

Gilquin, G. (2023). Written learner corpora to inform teaching. In The Routledge Handbook of Corpora and English Language Teaching and Learning (pp. 281-295). Taylor and Francis. https://doi.org/10.4324/9781003002901-25

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In The Cambridge Handbook of Learner Corpus Research (pp. 485-510). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.022

Meunier, F. (2016). Learner corpora and pedagogical applications. In The Routledge Handbook of Language Learning and Technology (pp. 376-387). Taylor and Francis. https://doi.org/10.4324/9781315657899-41

Subject-Verb Agreement in English Learner Texts: A Pseudo-Longitudinal Perspective

Karlsen Petter Hagen and Sylvi Rørvik (Inland Norway University of Applied Sciences)

This study investigates the use of subject-verb agreement in argumentative texts written by three different groups of Norwegian learners of English: pupils in Years 10 and 11 (material from the TRAWL corpus, cf. Dirdal et al 2022), and first-year university students of English (material from the International Corpus of Learner English, cf. Granger et al 2009).

The study is pseudo-longitudinal and aims to identify differences in the frequency and form of concord mistakes made by the writers in each subcorpus, as well as potential indications of developmental trajectories in the acquisition of concord competence.

The study has been guided by the following research questions:

1. What is the frequency and form of concord mistakes in argumentative texts written by Norwegian learners of English in pre-tertiary (Years 10 and 11) and tertiary education?
2. Can any potential developmental trajectories be identified in the pseudo-longitudinal data?

All finite verb phrases in the texts (excluding those containing modals) were coded for presence or absence of concord mistakes, as well as the form of the verb and the subject in cases of such mistakes, employing a coding scheme inspired by Garshol (2019) and Killie (2019). Preliminary results indicate a high degree of individual variation, but the following trends can be observed: The frequency of concord mistakes seems to decrease with increased age/proficiency. In Years 10 and 11 the subjects are relatively evenly distributed between those realized by subjects and those realized by noun phrases, while in the ICLE material the majority of mistakes occur with noun phrases functioning as subjects. Finally, an increasing proportion of concord mistakes involve relative pronouns functioning as subjects, i.e. the percentage of such cases is higher in the ICLE material than in Year 11, and higher in Year 11 than in Year 10.

References

Garshol, L. (2019). I JUST DOESN’T KNOW: Agreement errors in English texts by Norwegian L2 learners: Causes and remedies [Doctoral dissertation]. University of Agder.

Killie, K. (2019). The Acquisition of Subject-Verb Agreement Among Norwegian (Teenage) Learners of English: Focus on the Subject. Journal of Linguistics and Language Teaching, 10: 2: 183-206.

Creation and exploration perspectives of ScientEst, an Estonian Learner corpus of French Academic Discourse

Käsper, Marge and Anu Treikelder (University of Tartu)

In the context of our scientific collaboration with the LIDILEM laboratory at the University of Grenoble Alpes, we are currently engaged, as an extension of their Scientext project (2007-2010), in the development of ScientEst corpus, a corpus comprising about one hundred Bachelor's dissertations (102) written in French by Estonian learners affiliated to the Department of Romance Languages at the University of /.../. The forthcoming corpus will be made accessible on the Lexicoscope platform (Kraif 2016), facilitating its automated lexico-syntactic exploration and its comparison with other corpora, such as the corpus of articles in French and English in SHS ParaSHS, as well as the sources available on the ScienQuest (Falaise, Tutin, Kraif 2011) and LST (Hatier 2016) platforms developed by the same laboratory.

In terms of the source context, it could be interesting to compare these texts written in French by Estonian students with the database being created by the Bwrite project (Hint et al. 2022) assembling scientific texts written in Estonian. In terms of the perspective of teaching French as a Foreign Language (FLE), this corpus will be a valuable tool for scrutinizing the elements of practice and the representations on these practices concerning the discursive routines of academic French seen by our students (cf Yan 2017 for further didactic context). The aim of our communication is to examine the specificities of this type of learner corpus. Through concrete examples, we aim to delineate its potential applications in the contrastive analysis of scientific discursive routines and in the pedagogical endeavor of imparting academic discourse skills to Estonian students.

References

Falaise, A., Tutin, A., & Kraif, O. (2011). Exploitation d’un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques. Traitement Automatique des Langues Naturelles, 88.

Fløttum, K., T. Dahl & T. Kinn (2006). Academic Voices – across languages and disciplines. Amsterdam/Philadelphia: John Benjamins.

Hatier, S. (2016). Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS. Thèse de doctorat, Université Grenobles Alpes. https://www.theses.fr/2016GREAL027

Hint, Helen; Leijen, Djuddah A.J.; Jürine, Anni (2022). Eestikeelse akadeemilise teksti tunnustest [About the features of Estonian academic writing]. Keel ja Kirjandus, 4, 327−353. DOI: 10.54013/kk772a3.

Hyland, K. (2011). Academic discourse. In Hyland, K. & Paltridge, B. (eds.) Continuum Companion to Discourse Analysis. London: Continuum, 171–184.

Kraif, Olivier (2016). Le lexicoscope : un outil d’extraction des séquences phraséologiques basé sur des corpus arborés. Cahiers de Lexicologie, Phraséologie et linguistique appliquée, 1 (108), 91–106. ⟨10.15122/isbn.978-2-406-06281-3.p.0091⟩. ⟨hal-01884944⟩

Suomela-Salmi, E. & F.Dervin (2009). Cross-Linguistic and Cross-Cultural Perspectives on Academic Discourse, John Benjamins Publishing Company.

Tutin, A. (2014). La phraséologie transdisciplinaire des écrits scientifiques: des collocations aux routines sémantico-rhétoriques. In A. Tutin & F. Grossmann (Éd.), L’écrit scientifique: du lexique au discours. Autour de Scientext. Rennes: Presse universitaire de Rennes, 271–44.

Tutin, A. et F. Grossmann (2014). L’écrit scientifique : du lexique au discours. Autour de Scientext, Presses de l’Université de Rennes.

Tutin, A., Ji, Y., Kraif, O. (2022). Repérage et analyse des routines sémantico-rhétoriques dans le discours scientifique : application aux routines de guidage du lecteur. Langages, Armand Colin (Larousse jusqu'en 2003), 2022, 1/2022 (225). ⟨hal-03654193⟩

Yan, R. (2017). Étude des constructions verbales scientifiques dans une perspective didactique : utilisation des corpus dans le diagnostic des besoins langagiers en FLE à l'aide des techniques de TAL. Linguistique. Université Grenoble Alpes. ⟨NNT : 2017GREAL007⟩. ⟨tel-01691923v2⟩

Lexicoscope 2.0 : http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/

ScienQuest : https://corpora.aiakide.net/scientext20/?do=SQ.setV5iew&view=corpora

LST : http://lst.demarre-shs.fr/

Design, measurement, and analysis in longitudinal corpus-based SLA research: A systematic review

Kim, Minjin and Kevin McManus (The Pennsylvania State University)

Longitudinal data are both necessary and critical to documenting and understanding the routes and rates of second language (L2) development (Tracy-Ventura, et al., 2020; Vyatkina, 2012). In response to this awareness in the field, longitudinal corpus-based studies of L2 development are becoming more frequent. With this interest, however, careful attention must be paid to multiple characteristics unique to longitudinal research so that corpus-based findings of development can meaningfully contribute to knowledge, understanding, and theory development in the field, including questions about design, measurement, and analysis. To better understand questions of rigor and quality in the design and conduct of corpus-based, longitudinal SLA studies, this current study systematically examined the designs, measurements, and data analysis methods of 127 longitudinal corpus-based L2 studies.

Each study was coded for 21 items under 4 categories: study identification (e.g., author, publication year, type, journal), sampling and design (e.g., sample size, L1, observation period, frequencies, and intervals), measurement (e.g., modality, task type), and analysis technique (e.g., methods, tools, statistical analysis). In addition to multiple strengths, our review identified several concerns compromising the validity of such findings, such as unspecified intervals between observations, unclear reporting of tasks, and little reliability testing in measurement and analysis. At the same time, the rigor and quality of longitudinal corpus-based studies of L2 development appear to be improving, in line with broader trends in L2 research (e.g., larger sample sizes, inferential with descriptive statistics). In this presentation, we review these trends in the field while also discussing areas for growth and improvement, especially in terms of selecting an appropriate data collection interval for longitudinal studies, areas of L2 proficiency that are understudied, appropriate use of automated tools, and the potential of advanced statistical modelling.

References

Tracy-Ventura, N., Huensch, A., & Mitchell, R. (2020). Understanding the long-term evolution of L2 lexical diversity: The contribution of a longitudinal learner corpus. In B. Le Bruyn & M. Paquot (Eds.), Learner corpus research meets second language acquisition (1st ed., pp. 148–171). Cambridge University Press. https://doi.org/10.1017/9781108674577.008

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. https://doi.org/10.1111/j.1540-4781.2012.01401.x

Lexical complexity indices as markers of proficiency in L2 Russian

Kisselev, Olesya V. (University of South Carolina), Mikhail Kopotev (Stockholm University) and Anton Vakhranev (independent researcher)

Linguistic complexity has served as an important measure of second language (L2) writing development and an important construct in language assessment (Lu, 2011; Verspoor et al., 2012). Complexity indices, however, rarely feature in the studies of the Less Commonly Taught Languages (LCTLs) due to the lack of corpus-based tools. Additionally, studies assessing complexity in L2 data have been criticized for the lack of consistency in defining proficiency (Gablasova et al., 2017; Ortega, 2012). In this paper, we address these gaps by a) testing a newly designed set of codes that measure lexical indices in one such LCTL, Russian, and b) by exploring writing development in learner Russian, while paying specific attention to operationalization of the notion of proficiency. The study is based on a corpus of 601 essays (103,150 words) written by learners at different proficiency levels, operationalized as ratings on a standardized writing proficiency test. Following previous research (e.g., Bulté & Housen, 2014; Díez-Ortega & Kyle, 2023) we explored the following indices as potential indices of lexical complexity: average token length, average vowel count, average morpheme count, unique tokens/text, unique lemmas/text, function words, HD-D and MTLD for tokens, HD-D and MTLD for lemmas.

In the analysis, we employed descriptive statistics, paired samples t-tests, and the Games-Howell Posthoc test. Our results demonstrate that most lexical indices that we investigated (specifically, average token length, average vowel count, unique tokens, unique lemmas, and different versions of TTR both by lemma and token) showed significant correlation with proficiency levels. However, average morpheme count or the proportion of function words did not help distinguish proficiency levels.

The findings largely align with patterns identified in previous L2 complexity research: consistent with the results of many previous studies, overall lexical complexity in the texts of our learners increased with increased proficiency.

References

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

Díez-Ortega, M., & Kyle, K. (2023). Measuring the development of lexical richness of L2 Spanish: A longitudinal learner corpus study. Studies in Second Language Acquisition, 1-31.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring Learner Language Through Corpora: Comparing and Interpreting Corpus Frequency Information: Exploring Learner Language Through Corpus. Language Learning, 67(S1), 130–154. https://doi.org/10.1111/lang.12226

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. https://doi.org/10.5054/tq.2011.240859

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 (pp. 127–155). De Gruyter.

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.

To what extent do P-bursts resemble lexicogrammatically meaningful chunks? An exploratory study on beginner level foreign language writing processes

Kruse, Mari (University of Tartu)

Evidently, language does not consist of isolated words but rather chunks of lexicogrammatical elements that have semantic and syntactic roles, and these can be combined in increasingly complex phrases and paragraphs (Ungerer & Hartmann 2023). We also know that language production takes place in bursts, so both in oral form and in writing processes there are shorter or longer pauses between sequences produced. In writing, however, such pauses can also take place mid-word and separate single words while the writer is engaged in formulating their text. Until now, only a few authors (Cislaru and Olive 2017, Gilquin 2020, 2022ab) have suggested exploring such formulation sequences or meaningful chunks of words produced together in keylogger data, yet this qualitative view could greatly enhance our understanding of how language is stored and produced, as well as how foreign language learning takes place.

In this study, a meaningful burst or production chunk was defined as a sequence of two or more words produced in succession without a pause over the established threshold (500/1200 ms). The data comes from a longitudinal corpus of beginner Spanish L3 (first 9 months of study). Five participants who had provided all the 9 sample texts along with their production logs recorded with Inputlog (Leijten and VanWaes 2013) were selected for exploratory analysis. Preliminary results indicated added interest from using a double threshold: bursts centered around nuclei of notion and indicated some groups of words that are clearly processed together. Also, having absolute beginners alongside participants with some previous experience revealed interesting developmental dynamics, showing how a headstart allows for more stabile and substantial growth in fluency, both from a quantitative and a qualitative point of view.

References:

Cislaru, G. & Olive, T. (2017) Segments répétés, jets textuels et autres routines. Quel niveau de pré-construction? Corpus 17.

Gilquin, G. (2020) In search of constructions in writing process data. Belgian Journal of Linguistics 34: 99-109.

Gilquin, G. (2022a) Bursts of writing in process data: A new way of approaching constructions. Oral communication at the 11th International Conference on Construction Grammar (ICCG11), University of Antwerp (online).

Gilquin, G. (2022b) The Process Corpus of English in Education: Going beyond the written text. Research in Corpus Linguistics 10(1): 31-44. DOI: https://doi.org/10.32714/ricl.10.01.02

Ungerer, T. & Hartmann, S. (2023) Constructionist Approaches: Past, Present, Future. Cambridge University Press. DOI: https://doi.org/10.1017/9781009308717

KOST, the first learner corpus for Slovene as a second language

Stritar Kučuk, Mojca (University of Ljubljana)

KOST 2.0 (Korpus slovenščine kot tujega jezika, https://www.cjvt.si/korpus-kost/) is the first learner corpus of Slovene language as a second or foreign language. As a written corpus of approx. 1.5 million tokens, it consists of texts written by non-native speakers of Slovene who are learning Slovene in various programmes at the University of Ljubljana, so mostly in an academic setting. Although the authors of the texts speak more than 30 first languages, 48% of all authors are speakers of languages closely related to Slovene, i.e. the Central South Slavic languages (Bosnian, Croatian, Montenegrin and Serbian) and Macedonian. The included texts are mostly essays on a wide array of topics, written mainly digitally. In the last years, however, we have been focusing on texts written in exam conditions to avoid the use of machine translation and AI which has become widespread among the authors of our texts.

KOST 2.0 is lemmatised and POS-tagged. More than 24% of the texts are annotated with error tags, following a specifically created error taxonomy with 23 error tags. The errors were annotated manually in the application Svala (https://orodja.cjvt.si/svala/), adapted to Slovene specifics from its Swedish original. To browse the error tags more easily and to access the rich KOST metadata, we also developed a new concordancer which, among other, allows us to view the original and normalised versions of texts simultaneously, to analyse the concordances in a transparent manner and to browse the corpus by different error tags or by other meta data.

KOST can be openly accessed in the new concordancer (https://kost.cjvt.si/). It is also available for download as a single XML file, compliant with the TEI schema and the design of other corpora for Slovene language.

References:

Arhar Holdt, Š., Kosem, I., Stritar Kučuk, M. (2022): Metode in orodja za lažjo pripravo korpusov usvajanja jezika. Simpozij Obdobja 41: Na stičišču svetov: Slovenščina kot drugi in tuji jezik: 23–30. Ljubljana: Založba Univerze v Ljubljani.

Granger, S. (2003): Error-tagged learner corpora and CALL: A promising synergy. CALICO Journal, (20) 3: 465–480.

Granger, S. (2008): Learner corpora. Ludeling A., Kyto, M. (ur.), Corpus Linguistics. An International Handbook (259–275). Mouton de Gruyter.

James, C. (1998). Errors in Language Learning and Use: Exploring Error Analysis. Longman. https://doi.org/10.4324/9781315842912

Kosem, I., Stritar, M., Može, S., Zwitter Vitez, A., Arhar Holdt, Š., Rozman, T. (2012): Analiza jezikovnih težav učencev: Korpusni pristop. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Lüdeling, A., Walter, M., Kroymann, E., Adolphs, P. (2005). Multi-level error annotation in learner corpora. Proceedings of corpus linguistics 2005. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2005-journal/languagelearninganderror/multilevelerror.doc

Mikelić Preradović, N. (2020): Označavanje pogrešaka u CroLTeC-u (računalnom učeničkom korpusu hrvatskog kao stranog jezika). Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, (46) 2: 899–920.

Pirih Svetina, N. (2005): Slovenščina kot tuji jezik. Ljubljana: Izolit.

Rakhilina, E., Vyrenkova, A., Mustakinova, E., Ladygina, A., Smirnov, I. (2016): Building a learner corpus for Russian. Joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition, SLTC: 66–75.

Reznicek, M., Lüdeling, A., Krummes, C. (2012): Das FALKO-Handbuch: Korpus Aufbau und Annotationen, Version 2.01. Berlin: Humboldt-Universität zu Berlin.

Rosen, A., Hana, J., Hladká, B., Jelínek, T., Škodová, S., Štindlová, B. (2020): Compiling and annotating a learner corpus for a morphologically rich language: CzeSL, a corpus of non-native Czech. Praga: Karlova univerza.

Stritar, M. (2012): Korpusi usvajanja tujega jezika. Ljubljana: Zveza društev Slavistično društvo Slovenije.

Stritar Kučuk, M. (2022): KOST med korpusi usvajanja tujega jezika. Simpozij Obdobja 41: Na stičišču svetov: Slovenščina kot drugi in tuji jezik: 323–334. Ljubljana: Založba Univerze v Ljubljani.

Stritar Kučuk, M. (2023): Priročnik za označevanje napak. https://www.cjvt.si/korpus-kost/wp-content/uploads/sites/24/2023/10/Prirocnik-za-oznacevanje-napak-v-KOST-u-2023-10-17.pdf

Stritar Kučuk, M. (2024): Prvi korpus slovenščine kot tujega jezika KOST 1.0. Razvoj slovenščine v digitalnem okolju. Ljubljana: Založba Univerze v Ljubljani. 93 – 117.

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C., Sundberg, G., Wirén, M. (2019): The SweLL Language Learner Corpus: From Design to Annotation. Northern European Journal of Language Technology 6: 67–104.

Wirén, M., Matsson, A., Rosén, D., Volodina, E. (2018): SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. Selected papers from the CLARIN Annual Conference 2018. Linköping Electronic Conference Proceedings 159: 227–239.

Exploring Noun Lexical Diversity and Noun Phrase Complexity: A Learner Corpus-Based Study of B1 and C1 Spanish EFL Learners' Email Writing

Laso Martín, Natalia Judith (Universitat de Barcelona) and María Belén Díez Bedmar (Universidad de Jaén)

NP literature on EFL writing mainly focuses on linguistic complexity and accuracy (Ai & Lu, 2013; Bulté & Housen, 2014; Crossley & McNamara, 2014; Parkinson & Musgrave, 2014; Xu, 2019; Author, 2020; Kim, 2021), as well as lexical richness and phraseological competence (Nation, 2001; Hyland, 2008; Šišková, 2012; Peters, 2016; Vedder & Benigno, 2016; Paquot, 2019; Du et al., 2022). However, little is known about the relation between the lexical diversity of nouns and the syntactic complexity of the NPs of which those nouns are heads. To bridge this gap in the literature the main objective of this study is to investigate the relationship between lexical diversity and NP complexity in learner writing.

This paper examines 680 noun lexemes in the 2039 NPs produced in 44 (B1 level) and 46 (C1 level) emails from the *FineDesc* learner corpus. All texts were POS tagged using Freeling (Padró et al., 2010; Padró & Stanilovsky, 2012) and all noun lexemes were first disambiguated by means of the UKB option (Agirre et al., 2018) in Freeling and were later annotated using WordNet (Fellbaum, 1998). All NPs were then manually annotated using a taxonomy which encapsulates the semantic and NP complexity variables found in the NPs.

The results reveal a) a lower range of nouns in the B1 sample than in the C1 one; b) a low percentage of B1 and C1 nouns at both levels, considering the English Vocabulary Profile; and c) differences in the NP complexity types employed with prototypical hypernyms typical of B1 and C1 levels, as the concreteness or abstractness of the head-nouns and the students' CEFR level have an effect on NP complexity.

This paper underscores the importance of complementing the studies of lexical diversity and NP complexity to gain a better understanding of NP use. These results may inform teaching, learning and assessment of L1 Spanish EFL learners.

References

Agirre, E., López de Lacalle, O., & Soroa, A. (2018). The risk of sub-optimal use of Open Source NLP Software: UKB is inadvertently state-of-the-art in knowledge-based WSD. NLP-OSS workshop at ACL (arXiv:1805.04277).

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In Díaz-Negrillo, A., Ballier, N., & Thompson, P. (Eds.) *Automatic treatment and analysis of learner corpus data* (pp. 249–264). Amsterdam: John Benjamins Publishing Company.

Author (2020)

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

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66–79.

Du, X., Afzaal, M., & Al Fadda, H. (2022). Collocation Use in EFL Learners' Writing Across Multiple Language Proficiencies: A Corpus-Driven Study. *Frontiers in Psychology*, 13:752134.

Fellbaum, Ch. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, 4–21.

Kim, J. (2021). Measuring NP complexity in Korean EFL writing across CEFR levels A2, B1 and B2. *Korean Journal of English Language and Linguistics*, 21, 341-358.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Padró, Ll., Reese, S., Agirre, E., & Soroa, A. (2010). Semantic Services in FreeLing 2.1: WordNet and UKB. In Bhattacharyya, P., Fellbaum, C., & Vossen, P. (Eds.) *Principles, Construction, and Application of Multilingual Wordnets* (99-105). Mumbai, India: Narosa Publishing House.

Padró, Ll. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*. Istanbul, Turkey.

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

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.

Peters, E. (2016). The learning burden of collocations: the role of interlexical and intralexical factors. *Language Teaching Research*, 20, 113–138.

Šišková, Z. (2012). Lexical Richness in EFL Students' Narratives. *Language Studies Working Papers*, 4, 26-36.

Vedder, I., & Benigno, V. (2016). Lexical richness and collocational competence in second-language writing. *International Review of Applied Linguistics in Language Teaching*, 54(1), 23-42.

Xu, L. (2019). Noun phrase complexity in integrated writing produced by advanced Chinese EFL learners. *Papers in Language Testing and Assessment*, 8(1), 31-51.

Wordless: An integrated corpus tool with multilingual support for the study of language acquisition, pedagogy, and assessment

Lei, Ye (Shanghai International Studies University)

This paper presents Wordless (Ye, 2023), an integrated corpus tool with multilingual support for the study of language, literature, and translation. Wordless is free, cross-platform, and open-source. It aims to eliminate the barriers to the utilization of bleeding-edge technologies by language researchers and provide a better alternative to other corpus tools currently available such as WordSmith, AntConc, Sketch Engine, and ParaConc.

Wordless features a user-friendly graphical interface, catering specifically to the needs of non-technical users. It adopts a modular design and each of the 9 main modules provides functionalities for corpus profiling, concordancing, parallel concordancing, dependency parsing, wordlist generation, n-gram generation, collocation extraction, colligation extraction, and keyword extraction respectively.

Wordless is “batteries-included”: it has built-in multilingual support for multiple natural language processing tasks including sentence/word/syllable tokenization, part-of-speech tagging, lemmatization, stop word filtering, dependency parsing, sentiment analysis, and more. It can process and analyze corpora in more than 120 languages across the globe, including many non-Indo-European languages as well as those spoken in some Southeast Asia countries which require special tokenization handling such as Burma, Chinese, Khmer, Lao, Japanese, Thai, Tibetan, and Vietnamese.

Wordless is capable of calculating a wide range of statistical measures, including indicators of readability and lexical diversity of the whole text, measures of dispersion and adjusted frequency indicating the distributive evenness of tokens/n-grams in the corpus, and methods of Bayes factor and effect size used to complement tests of statistical significance in collocation/colligation/keyword extraction. Wordless also comes with a variety of data visualization options including dispersion plots, line charts, word clouds, network graphs, and dependency graphs. Other convenience features include zapping options for quick creation of cloze tests to be used in classroom settings and functionalities for parallel concordancing.

Resource

Ye, L. (2023). Wordless (Version 3.4.0) [Computer software]. Github. https://github.com/BLKSerene/Wordless

The acquisition of the article system by Polish advanced learners of English: evidence from legal translations

Leńko-Szymańska, Agnieszka; Łucja Biel and Katarzyna Wasilewska (Institute of Applied Linguistics, University of Warsaw)

This study examines the use of articles by advanced learners of English whose L1 is Polish, a [-ART] language. The investigation zooms in on legal texts. Legal genres provide interesting data to study article use, which so far have not been explored from the acquisitional perspective. Syntactically, legal language is characterized by a high degree of nominalization and complex NP structures (Bathia, 1994). At the semantic and pragmatic level, the definiteness and specificity of reference are particularly relevant categories in legal contexts. Moreover, the learner legal texts analyzed in this study are translations produced by students in an MA translation programme. This choice of data elicitation technique combines the advantages of a closed task and of free production, as it reflects a controlled but authentic (even if not typical and common) language use situation.

The present study addresses the following questions:

1. Is there a difference in the frequency of articles between legal translations produced by Polish advanced learners of English, translations performed by professional translators and comparable non-translated legal texts in English? Is so, does it point to an underuse of articles by learners, observed in earlier studies examining more common genres (Ekiert, 2004)?
2. Is there a variation in the use of articles in L2 learners' texts? If so, does it depend on specific semantic, pragmatic and grammatical contexts? How does it compare with the use of articles in expert translations and in comparable non-translated texts?

The data has been drawn from five genre-specific corpora: (1) a learner parallel corpus containing 78 Polish-English translations of the same legal text ; (2) an expert parallel corpus of 11 translations of the same text ; (3) a comparable parallel corpus of expert translations; (4) two comparable corpora of non-translated legal texts as a benchmark.

The results demonstrate that the learner translations display exceptionally high frequencies of the definite article and low frequencies of the indefinite article in comparison with the non-translated texts. However, they are not different than the respective frequencies in the expert translations of the same text. At the same time, the most frequent noun phrases in the learner corpus show different degrees of variation in article use which can be linked to the semantic, pragmatic and grammatical categories of the head noun referents.

References

Bhatia, V. K. (1994). Analysing Genre: Language Use in Professional Settings. New York: Longman.
Ekiert, M. (2004). Acquisition of the English article system by speakers of Polish in ESL and EFL settings. Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 4(1): 1-23.

A metadata scheme for cross-corpus analyses of L2 acquisition

Lenort, Lisa (Leipzig University), Annette Portmann (Leipzig University), Matthias Schwendemann (Leipzig University), Josef Ruppenhofer (Fernuniversität Hagen)

Reliable and informative metadata make learner corpora more useful and more easily re-usable for a variety of purposes. To make metadata more comparable across corpora, the learner corpus research community has advanced initiatives to foster metadata standardization (e.g., core learner metadata (Granger & Paquot, 2017) or the LC core metadata draft (König et al., 2022, Paquot et al., 2023).

When it comes to researching second language acquisition (SLA), variability plays a crucial role, a fact that has recently received much attention (again) (Bayley et al., 2022; Shadrova et al., 2021; Wulff & Gries, 2021) . Metadata in learner corpora used for SLA research purposes therefore ought to capture potential factors for variation that often go beyond core metadata. However, for many reasons (e.g., legal, practicality, ethical), only few corpora contain metadata regarding, e.g., the age of onset, the amount/quality of L2 input, or a comprehensive language biography of learners.

In our presentation, we present a fine-grained, SLA-sensitive metadata scheme that was developed in a research project which links various German learner corpora (N>40). The scheme is a project-specific adaptation of the core metadata mentioned above which adds aspects of SLA-theory and of metadata contained in those corpora pulled together.

We first present the methodology – and challenges – of the documentation of the metadata of the datasets to be harmonised and the development of our scheme. Introducing the scheme itself, we will then focus on selected variable groups that are of particular theoretical and methodological relevance to the field. These include the challenge to define metadata variables that adequately capture comprehensive language learning biographies, a task that showed very heterogeneous (or implicit, or even a lack of) definitions of L1, L2, L3 in the corpora at hand. We further discuss metadata for proficiency estimates/classifications and also the difficult classification of task types.

References

Bayley, R., Li, X., & Preston, D. R. (2022). Variation in Second and Heritage Languages. Variation in Second and Heritage Languages, 1–377.

Granger, S. & Paquot, M. (2017). Core Metadata [Schema] for Learner Corpora Draft 1.0. https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/61

König, A., Frey J.-C., Stemle, E., Glaznieks, A. & Paquot, M. (2022). Towards standardizing LCR metadata. Paper presented at Learner Corpus Research 6, 22-24 September 2022, University of Padua, Italy.

Paquot, M., König, A., Stemle, E. & Frey, J.-C. (2023). Core Metadata Schema for Learner Corpora. UC Louvain, V1. https://dataverse.uclouvain.be/dataset.xhtml?persistentId=doi:10.14428/DVN/4CDX3P

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, 1–29. https://doi.org/10.3389/fpsyg.2021.716485

Wulff, S., & Gries, S. Th. (2021). Exploring Individual Variation in Learner Corpus Research: Methodological Suggestions. In B. Le Bruyn & M. Paquot (Hrsg.), Learner Corpus Research Meets Second Language Acquisition (S. 191–213). Cambridge University Press. https://doi.org/10.1017/9781108674577.010

Bridging academic and technological domains. The new framework for developing the Estonian L1 and L2 preschool children's speech corpus.

Lilles, Kelly (Tallinn University, Institute of the Estonian Language, ALPA Kids)

This presentation addresses the critical gap in corpus-based research on the acquisition of grammatical constructions in Estonian as L2 during early childhood. It highlights the absence of a corpus featuring both L1 and L2 language productions by preschool-aged children, a concern raised in numerous studies (e.g., Lu 2023; Eslon et al. 2010).

Currently, there are several learner text corpora available for Estonian (i.e. ELLE and EMMA) that contain texts written by adult and young learners. Additionally, within the CHILDES network, there exists an Estonian children's speech corpus containing transcriptions from monolingual children. Furthermore, the RusLAPSED corpus includes materials collected from Estonian Russian-speaking children.

Nevertheless, a notable absence persists — a preschool children's spontaneous speech corpus collected based on identical principles from both L1 and L2 speakers, along with comprehensive metadata. To bridge this gap, we propose a framework for the creation of an extensive speech corpus utilizing the ALPA Kids platform. Recognized as the most preferred e-learning games platform among Estonian preschool teachers (CASS 2023), it provides a promising avenue for fulfilling this crucial need.

ALPA Kids, with 100,000+ monthly users across 14 languages, currently aggregates anonymized data from 20,000+ Estonian children aged 3-8 with information on age, gender, native language, and in-app activities. A forthcoming enhancement includes a story-dice game designed to stimulate spontaneous speech production, allowing parents to contribute children's stories for research purposes. The speech will be transcribed, metadata added together with lemmatization, morphological, syntactic, and error annotation.

The author will detail the technical and legal issues for data collection and explore possibilities for enhancing children's speech recognition technologies and implementing state-of-the-art corpus analysis methodologies for L2 acquisition. This project not only promises to enrich Estonian language resources but also demonstrates a public-private collaborative development of linguistic tools that bridge academic and technological domains.

References:

ALPA Kids https://alpakids.com/
CASS (University of Tartu Centre for Applied Social Sciences) 2023. The Usage of Estonian Edtech Solutions. https://www.edtechestonia.org/resources
CHILDES https://childes.talkbank.org/access/Other/Estonian/MAIN.html
ELLE https://evkk.tlu.ee/
EMMA Corpus https://korpused.keeleressursid.ee/emma/
Eslon, Pille; Õim, Katre; Kaivapalu, Annekatrin; Argus, Reili 2010. Kuidas uurida esimese ja teise keele omandamist? Lähivõrdlusi. Lähivertailuja, 20 (2010), 11–48.
Lu, Xiaofei 2023. Corpus Linguistics and Second Language Acquisition: Perspectives, Issues, and Findings, Cognitive Science and Second Language Acquisition Series. New York London: Routledge, Taylor & Francis Group.
RusLAPSED https://pedagogicum.ut.ee/et/teadusrakk/laste-kone-elektroonne-korpus-ruslapsed

Construction, transcription and annotation of a longitudinal multimodal interlanguage corpus: An ongoing study with L1 Italian and L2 Chinese

Liu, Siyuan (University of Bologna)

This study focuses on a novel corpus, the MICICL (a Multimodal Interlanguage Corpus for Italian Chinese Learners), which is an ongoing corpus project. The research goal of this project is to find the most appropriate path to construct a Chinese multimodal interlanguage corpus, and test finally constructed corpus's applications in an actual Chinese classroom teaching/learning context. For the corpus content, the MICICL will include data in three modalities: pictures, audio and text. For the data processing, the MICICL will transcribe picture data and audio data respectively. The picture data will be transcribed in the form of Chinese characters, and the audio data will be transcribed in the Chinese phonetic system - Pinyin; and the MICICL will include error annotation and basic linguistic information annotation. For the form of the corpus, the MICICL will be a longitudinal corpus, with the goal of tracking each participant for at least one semester.

This paper outlines the data collection processes, and data processing and analysis approaches of MICICL:

1. Data collection process: Data collection primarily occurs at the University of Bologna, employing both online and offline methods. This section discusses methodological aspects of data collection, including how to control variables to the maximum extent and guide students to express their true proficiency.

2. Data processing and analysis approaches: MICICL employs distinct processing methods for different modalities. The challenges in designing transcription/annotation methods for different modalities, especially dealing with a large number of incorrectly written Chinese characters produced by learners, are addressed. The processing of incorrectly written characters, which cannot be directly typed or stored/displayed in universal computer characters, poses a significant challenge. Determining error annotation standards for various modalities, defining granularity and versatility of error annotation, and striking a balance between redundant and insufficient annotation information are additional challenges faced by this research.

References

Cui X.L. & Zhang B.L. (2011).Quanqiu Hanyu Xuexizhe Yuliaoku Jianshe Fangan. Yuyan Wenzi Yingyong (02)

Lee, L. H. et al, (2018). Building a TOCFL learner corpus for Chinese grammatical error diagnosis. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)

Li, L. (2021). A spoken Chinese corpus: development, description, and application in L2 studies: a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Applied Linguistics at Massey University, Manawatū, New Zealand (Doctoral dissertation, Massey University)

Iurato, A. (2022). Learner Corpus Research Meets Chinese as a Second Language Acquisition: Achievements and Challenges. Annali di Ca’ Foscari. Serie orientale

A Diasystematic Construction Grammar (DCxG) analysis of Progressive Aspectuality in Multilingual Learners of English as Additional Language

Lopopolo, Olga (Eurac Research)

All natural languages, whether or not they have a designated grammatical category conventionally referred to as progressive, can convey the idea that an event is progressing dynamically over a time frame opened up by an utterance (Mair 2012: 803). However, it is important to distinguish between a semantic-cognitive notion of progressive aspectuality, which is universal and transportable across languages, and the corresponding formal expression for this notion, i.e. the progressive aspect, found in various languages, which can be obligatory or optionally marked on lexical verbs or verb phrases.

The present study explores how multilingual learners of English express the universal semantic-cognitive concept of progressive aspectuality. The framework that will be applied is the Diasystematic Construction Grammar (DCxG) (Höder 2012, 2014, 2021) which postulates the existence of shared constructions ("form-meaning-function constellations", Goldberg 2003) which speakers view as similar across multiple languages, termed 'diaconstructions,' and language-specific constructions, termed 'idioconstructions'. Learners' multilingual constructicon will be presented in an integrated network of dia- and idioconstructions organized along inheritance links connecting more specific and more schematic constructions to the supraordinate abstract concept of progressive aspectuality. The constructional link among constructions is based on either the formal or the functional side of the constructions themselves.

The analysis is conducted on learner data (Leonide – Glaznieks et al. 2022) composed by lower-secondary school students living in a multilingual environment, i.e. South Tyrol, in which they are exposed in different ways to the official languages of the environment, i.e. Italian and German, alongside English as a language subject. Preliminary results show preferences of dominant German learners of English in using language-specific idioconstructions, in contrast to dominant Italian learners of English in their use of diaconstructions shared across the two languages, because they fulfill the same function and exhibit the same form. Since DCxG aims to model multilinguals' linguistic knowledge in a socio-cognitively and dynamic realistic fashion (Höder et al. 2021: 314), the study considers learners' repertoires as prototypical dominant language constellations (DLC) (Aronin 2006, 2016) embedded in the South Tyrolean environment.

References

Aronin, L. (2006). Dominant language constellations: An approach to multilingualism studies. In M. Ó. Laoire (Ed.), Multilingualism in educational settings, Schneider Publications, 140–159.

Aronin, L. (2016). Multicompetence and dominant language constellation. In Vivian Cook & Li Wei (Eds.), The Cambridge handbook of linguistic multicompetence, Cambridge University Press, 142–163.

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.

Goldberg, A.E. (2003). Constructions: A new theoretical approach to language. Trends in Cognitive Sciences, 7(5), 219-224.

Höder, S. (2012). Multilingual constructions: a diasystematic approach. Multilingual Individuals and Multilingual Societies (Hamburg Studies on Multilingualism 13), 241–257.

Höder, S. (2014). Constructing diasystems. Grammatical organization in bilingual groups. In Tor A. Åfarli & Brit Mæhlum (Eds.), The Sociolinguistics of Grammar, John Benjamins, 137-152.

Höder, S., Prentice, J., & Tingsell, S. (2021). Additional language acquisition as emerging multilingualism. A Construction Grammar Approach. In Hans C. Boas & Steffen Höder (Eds.), Constructions in Contact 2. Language change, multilingual practices, and additional language acquisition, John Benjamins, 310-337.

Mair, C. (2012). Progressive and imperfective Aspect. In Robert Binnick (Ed.), The Oxford Handbook of Tense and aspect, 803–827.

Discussing the categorization of speakers' language background: implicit assumptions and methodological challenges for Learner Corpus Research

Lopopolo, Olga (Eurac Research, University For Foreigners of Perugia), Arianna Bienati (Eurac Research, Università di Modena e Reggio Emilia), Jennifer-Carmen Frey (Eurac Research), Aivars Glaznieks (Eurac Research) and Stefania Spina (University for Foreigners of Perugia)

The categorization of speakers' language background is one of the core types of information needed in Learner Corpus Research (LCR). The choice of terms like 'L1', 'native speaker' and 'mother-tongue' in the different corpus research projects leads to a complex and sometimes controversial web of linguistic inquiry. The adoption of specific terms, subsequently integrated in the form of speakers' metadata in learner corpora, may be rooted in underlying theoretical paradigms not consistently clarified in corpus description papers. Consequently, the adoption of specific terms often carries implicit meanings, leading to a lack of shared understanding among researchers. This lack of consensus regarding the intended meanings of these labels can result in diverse interpretations, affecting not only the coherence and reliability of learner corpus studies, but also the comparability of the results when adopting different categorization types.

In the present contribution we aim at addressing three key issues: (1) the impact of implicit assumptions and methodological choices taken in learner corpus design when categorizing speakers' language background, (2) the consequences that different conceptualizations of speakers' language background might have on results and (3) the integration of alternative perspectives on speakers' language categorization in LCR. Through a comprehensive review of prominent learner corpora, we (1) identify the most common operationalizations of the learner_L1 metadata (Paquot et al. 2023) and scrutinize the underlying theoretical assumptions. Our exploration then extends to (2) the Italian and German subsections of LEONIDE (Glaznieks et al. 2022). We analyzed the learner texts with a range of complexity measures obtained with CTAP (Chen & Meurers 2016; Okinina et al. 2020) and compared group effects on the results by varying the constitution of learner and reference groups according to the operationalization found in (1). Drawing on research on plurilingualism, we (3) problematize the practice of classifying speakers' language backgrounds in ways that may not align with the complex reality of multilingual societies. We therefore propose a re-evaluation of language classification systems for LEONIDE that might better reflect speakers' language experiences in multilingual environments.

References

Chen, X.B. & Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In Proceedings of The Workshop on Computational Linguistics for Linguistic Complexity. Osaka, Japan. The International Committee on Computational Linguistics.

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.

Okinina, N., Frey, J.-C, & Weiss, Z. (2020). CTAP for Italian: Integrating Components for the Analysis of Italian into a Multilingual Linguistic Complexity Analysis Tool. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 7123–7131). European Language Resources Association.

Paquot, M., König, A., Stemle, E. & Frey, J.-C. (2023). Core Metadata Schema for Learner Corpora, https://doi.org/10.14428/DVN/4CDX3P

Exploring Formulaic Language in Student Academic Writing in L1 and L2 German

Lösel, Andrea (Leipzig University)

This working paper introduces an ongoing research project investigating the lexical resources used by native (L1) and non-native (L2) German-speaking students in their academic writing at German universities. Over the past decades, a growing consensus suggests that (academic) text production heavily relies on the retrieval of stable word combinations (Brommer, 2018; Bubenhofer, 2009; Steyer, 2000). This perspective is reflected in propaedeutic textbooks, which often present prospective students with prefabricated and exemplary word combinations (e.g., Buchner, 2015). However, limited research has delved into the word combinations actually employed by L1 and L2 students in their academic writing in German.

The research project addresses this gap by exploring the linguistic characteristics of student academic writing, focusing on the use of multi-word units by L1 and L2 German speakers. The annotated StuWiss corpus, created by the author, serves as the data source, encompassing 266 master's theses (6.04 million tokens) from L1 and L2 speakers submitted to German universities in GFL studies, further divided into an L1 and L2 subcorpus (134 and 132 theses, respectively).

The study aims to identify frequent multi-word units in L1 and L2 academic writing, further investigating their formal-morphosyntactic, lexical-semantic, and functional characteristics. Beyond the overall inventory of multi-word units used in the StuWiss corpus, the research also takes a comparative perspective, highlighting the most significant differences between the L1 and L2 subcorpus.

To address non-linear structures in German academic writing, the project introduces syntactic n-grams (Andresen & Zinsmeister, 2017; Sidorov, 2019). Leveraging dependency paths, syntactic n-grams allow to identify syntactically dependent multi-word units. This way, they may capture discontinuous token strings overlooked by traditional linear n-gram analysis. By incorporating both linear and syntactic n-grams, the study seeks a comprehensive understanding of the multi-word units used by L1 and L2 students in their academic writing in German.

References:

Andresen, Melanie; Zinsmeister, Heike (2017). The Benefit of Syntactic vs. Linear N-Grams for Linguistic Description. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). Linköping: Linköping University Electronic Press, pp. 4–14.

Brommer, S. (2018). Sprachliche Muster: Eine induktive korpuslinguistische Analyse wissenschaftlicher Texte. Berlin/Boston: de Gruyter.

Bubenhofer, Noah (2009). Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin/New York: de Gruyter.

Buchner, Patricia (2015). Campus Deutsch: Schreiben. München: Hueber.

Sidorov, Grigori (2019). Syntactic N-Grams in Computational Linguistics. Cham: Springer.

Steyer, Kathrin (2000). Usuelle Wortverbindungen. Linguistisches Konzept und lexikografische Möglichkeiten. In: Deutsche Sprache 28(2), pp. 101–125.

Quoting and Referencing Mastery: ExpoKo’s Insights Into Student Academic Writing Challenges

Lösel, Andrea; Matthias Schwendemann and Franziska Wallner (University of Leipzig)

The ExpoKo project, initiated in November 2022, aims to create an anonymized corpus of student exposés within the DaF/Z (German as a Foreign/Second Language) context. Exposés, giving a brief and structured overview of a proposed research project, are increasingly required by potential supervisors for final theses to assess the feasibility and scholarly value of the proposed research before agreeing to supervise. However, for many students, especially those with German as a foreign or second language, verbalizing their research project in the form of an exposé can be a challenging task (cf. Stezano Cotelo 2008). Despite the widespread need for exposés in the study context, there is a lack of exemplary materials and didactic resources. Due to their brevity, however, exposés display textual characteristics in student writing that can also be observed in longer academic (student) text types such as term papers and theses. Hence, their concise nature makes them conducive to effective didactic preparation.

ExpoKo aims to address the aforementioned gap by creating a freely accessible didactic corpus resource online. The evolving corpus currently comprises 74 exposés, predominantly contributed by L2 writers. All exposés have been anonymized and linguistically processed (tokenization, lemmatization, POS-tagging, and an annotation of structural elements, e. g. text blocks on research questions, methods, etc.).

A sub-project stems from the observation in teaching and in the exposés submitted that many student writers face challenges in the receptive-productive handling of research literature (cf. Mächler 2012, Dirks/Zhou 2019). We explore how students incorporate information from other academic texts into their own writing. Employing a corpus-driven approach, an annotation workflow has been developed to manually annotate quotations and references in all 74 exposés. This annotation facilitates quantitative data analysis, e. g. differentiation between syntactically integrated and syntactically isolated quotations. At the same time, it allows for in-depth qualitative examination, e. g. facilitating an analysis of the adequacy of formulation patterns used to incorporate external information.

The sub-project's overarching aim is to furnish illustrative material on quoting and referencing in student academic writing in German. In our presentation, we will introduce the annotation workflow and shed light on key challenges identified among L2 students when incorporating research literature into their own writings.

References

Dirks, Uni & Zhou, Bingchen: Zitierpraktiken und argumentative Funktionen in den Agrégationsklausuren des Fachs Deutsch. Nouveaux Cahiers d'Allemand: Revue de linguistique et de didactique, 37(3), 2019, pp. 265–278.

Mächler, Lissette: Erwerb des wissenschaftlichen Schreibens in der Fremdsprache Deutsch. Exemplarische Analyse von intertextuellen Prozeduren. Informationen Deutsch als Fremdsprache, 39(5), 2012, pp. 519-539.

Stezano Cotelo, Kristin: Verarbeitung wissenschaftlichen Wissens in Seminararbeiten ausländischer Studierender. Eine empirische Sprachanalyse. München: Iudicium 2008

Designing and compiling a multi-L1 corpus of L2 Spanish: Lessons learnt from CEDEL2

Lozano, Cristóbal (Universidad de Granada)

This talk will discuss the principles behind the design, data collection and compilation of any freely-available learner corpus that intends to be maximally informative not only for SLA researchers but also for a wide range of foreign language users and practitioners. The talk is ultimately intended to problematise the challenges that a solid L2 corpus design has to face with a view to signposting L2 corpus designers to foresee challenges of learner corpora that will likely grow and expand across time.

We will do so by assessing the lessons learnt from the evolution (and errors!) in the design and compilation of CEDEL2. Since its inception in 2004, CEDEL2 started as an L1 English-L2 Spanish written corpus. Over the past 20 years, many challenging decisions (which ultimately affect the overall design and usability of the corpus) had to be taken along the way. CEDEDL2 has grown into a large, multi-L1 corpus of L2 Spanish. In its current online web search & download interface (version 2.0), it contains mainly written (and some spoken) data from 11 typologically (un)related mother tongues (English, German, Dutch, French, Portuguese, Italian, Greek, Russian, Japanese, Chinese, Arabic). Version 3.0 is under development and will incorporate more written and spoken data, new L1s (Estonian, Swedish, Polish, Turkish and Vietnamese) and a new automatized data-collection interface.

The principles we will discuss relate to theoretical and technical decisions that have to be made if the corpus is intended to be used by a wide range of users. We will consider whether SLA-relevant questions should be asked first before deciding which linguistic-profile variables to collect for the corpus to be maximally informative for SLA researchers. We will discuss the convenience (and feasibility) of collecting other (non-)linguistic and cognitive variables that are central in SLA research (e.g., aptitude, motivation, dominance, attrition) and in pscyholinguistic approaches to SLA (e.g., working memory), which would require using additional standardized tests. However, a balance must be struck between the amount of linguistic information to be collected and the length of time participants need to complete the profiles and the actual linguistic data: the task. Task-related variables (from the actual task prompt to the resources, conditions and timing used to do the task) require special attention since these will affect the final output and may often mask the linguistic knowledge the learner is capable of attaining. We will also consider other data-related issues, like the benefits of collecting a native 'control' subcorpus (for every learner subcorpus), written and spoken data (from the same participant), different tasks (produced by the same participant), etc.

Technical aspects may be determined by the previous decisions and will determine the usability of the corpus: the transcription conventions for spoken data (if any), the protocols for massive data collection (and whether to automatize it), and the computational aspects behind an attractive (yet useful) online corpus search and download interface. These (and other) decisions are essential for the design of a valuable L2 corpus, but are vital if the corpus designer wants to go one step ahead and design a twin corpus (or mirror-image corpus). This was the case of COREFL, an L2 English corpus that was designed after CEDEL2 so as to be able to do mirror-image comparisons (e.g., L1 English-L2 Spanish vs L1 Spanish-L2 English, amongst others).

We will end up with a discussion of whether triangulation (i.e., the combination of naturalistic corpus-production data with less naturalistic but more controlled experimental data) is desirable to investigate a given linguistic phenomenon in sciences like SLA and bilingualism.

Resources:

LearnerCorpora website (for data collection and participation): http://learnercorpora.com

CEDEL2 corpus, L2 Spanish (freely searchable and downloadable): http://learnercorpora.com

COREFL corpus, L2 English (freely searchable and downloadable): http://corefl.learnercorpora.com

BilinguaLab (for an overview of the corpora and research aims): http://bilingualab.ugr.es

Georgian English learner corpus and its lexicographic applicability

Makhatadze, Marine (Ivane Javakhishvili Tbilisi State University)

The learner corpus is an important lexicographic deposit by which the linguistic internalization of foreign or second language acquisition is observed. We have created a learner corpus Georgian Learner English and Newspaper Corpus (GLEAN). This abbreviation in fact metaphorically represents the scope of our research, as the verb to glean is defined as “a gather (leftover grain) after a harvest”. A variety of texts (student essays, student newspaper articles, blog posts, etc.) are collected for our learner corpus and are morphologically annotated (POS-tagging).

We perform a contrastive analysis between the corpus we have created and the "control" corpus, of comparable native-speaker linguistic production (in this case - English). In the framework of our research, we use the Michigan Corpus of Upper-Level Student Papers (MICUSP). On the other hand, we use ICNALE (International Corpus Network of Asian Learners of English) as well. The aim of our research is: a) to demonstrate the peculiarities of the use of phraseological units and collocations (adverb + adjective, verb+noun, etc.) in the corpora; b) based on statistical measures, which phraseological units are overused or underused in the writings of Georgian students; c) to present some effective lexicographic ways through which the learner corpus data is applied into the English-Georgian dictionary microstructure, for example some usage notes are modified and included to the dictionary entries.

The study reveals that the semantics of phraseological units are, in many cases, generalized by language scholars and adapted to contexts in which native speakers rarely use, for example, the phrases I can say that and To my mind are popular in the Georgian learner corpus, whereas the examples of these phrases are not found in MICUSP. The result showed the linguistic errors in the texts of Georgian students when they used the phrase what about instead of "as for".

References

Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). #LancsBox v. 6.x. [software package].

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

Gouws, R.H. & Prinsloo, D.J. 2005. Principles and practice of South African lexicography. Stellenbosch: AFRICAN SUN MeDIA. doi: 10.18820/9781919980911. The original publication is available from AFRICAN SUNMeDIA - www.sun-e-shop.co.za

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

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

Hasselgård. 2015a. Learners’ and native speakers’ use of recurrent wordcombinations across disciplines. Bergen Language and Linguistics Studies (BeLLS) vol.6, 87-106.

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.

Hoffmann, S., & Evert, S. (2006). BNCweb (CQP-edition): The marriage of two corpus tools. Corpus technology and language pedagogy: New resources, new tools, new methods, 3, 177-195.

Hoffman, S. (2007). study. (Routledge advances in corpus linguistics.) New York: Routledge.(ix, 214). California Linguistic Notes, 32(2).

Ishikawa, S. I. (2023). The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners’ L2 English. Taylor & Francis.

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.

Römer, U., & Swales, J. M. (2010). The Michigan corpus of upper-level student papers (MICUSP). Journal of English for Academic Purposes, 9(3), 249.

Constructing a Spoken Corpus of Cochlear Implant Patients/Users

Mazaherylaghab, Hamzeh (Hamedan University of Medical Sciences)

The language of children with hearing loss receiving Cochlear Implants (CI) and other hearing aids is distinguished by several prominent attributes that are specific to this group. They receive either a hearing aid or a CI at an early age. It is found that they demonstrate various error patterns when producing language. Hearing loss or impairment is believed to affect their communicative and linguistic development in different ways. It is, therefore, required to conduct up-to-date studies focusing on the linguistic features and the communicative ability of these children through longitudinal and cross-sectional research. Discovering how they perform in their everyday lives, at home or elsewhere, could help researchers come up with solutions to their problems. The scarcity of documented language of the above-mentioned individuals calls for the compilation of a multimodal corpus of language of the speakers with different language backgrounds. Such corpora are thought to help boost our perception of the unique developmental features of the language produced by children with unilateral and bilateral cochlear implants. The present work serves as the initial outline of the corpus collection project to be conducted in speech therapy and rehabilitation centres in Iran where many CI users receive services after receiving CI.

To build such a corpus factors like size, balance, and representativeness need consideration. It is estimated that a small percentage of the transcribed data will be marked up using tags indicating necessary information about text structure, speech events, and other discoursal and grammatical properties. In order to have a corpus where annotations are linked to the audio and video files, the current research project is going to use ELAN to annotate multimodal spoken language corpora.

References

Adolphs, S., & Carter, R. (2013). Spoken Corpus Linguistics: From Monomodal to Multimodal: Routledge.

Knight, D. (2011). Multimodality and Active Listenership: A Corpus Approach: Corpus and Discourse. London: Bloomsbury.

Baker, P. (2006) Using Corpora in Discourse Analysis London: Continuum.

Biber, D., Conrad, S. & Reppen, R. (1998) Corpus Linguistics: Investigating Language Structure and Use. Cambridge: CUP.

Baker, P., Hardie, A. & McEnery, T. (2006) A Glossary of Corpus Linguistics. Edinburgh: Edinburgh University Press.

Banerjee, B., Kapourchali, M. H., Najnin, S., Mendel, L. L., Lee, S., Patro, C., & Pousson, M. (2016). “Inferring hearing loss from learned speech kernels,” in Proceedings of IEEE International Conference on Machine Learning and Applications, pp. 26–31.

Ronald, C., & McCarthy, M. (1997) Exploring spoken English. Cambridge University Press.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming & Danae Paolino. 1993. “Outline of Discourse Transcription”. In Edwards, Jane A. & Martin D.Lampert, eds., Talking Data: Transcription and coding methods for discourse research. 45-89.

Fillmore, Charles J. 1992. ““Corpus linguistics” or “Computer-aided armchair linguistics””. In: Svartvik, Jan (ed.) Directions in Corpus Linguistics. Berlin: de Gruyter. 35-60.

Fenlon, J., Schembri, A., Johnston, T., & Cormier, K. (2015). Documentary and corpus approaches to sign language research. In E. Orfanidou, B. Woll, & G. Morgan (Eds.), The Blackwell guide to research methods in sign language studies (pp. 156-172). Oxford: Blackwell.

Hodge, G., Sekine, K., Schembri, A., & Johnston, T. (2019). Comparing signers and speakers: building a directly comparable corpus of Auslan and Australian English. Corpora, 14(1), 63-76. doi:10.3366/cor.2019.0161

Kortmann, Bernd. 2006. “Syntactic Variation in English: A Global Perspective” in: Aarts, Bas April McMahon (eds.). The Handbook of English Linguistics. Malden, MA: Blackwell Publishers: 602-624.

Kennedy, Graeme. 1998. An Introduction to Corpus Linguistics. London & New York: Longman.

Leech, Geoffrey. 1992. “Corpora and theories of linguistic Performance”. In: Svartvik, Jan (ed.). Directions in Corpus Linguistics. Berlin: de Gruyter. 105-122.

Najnin, S., Banerjee, B., Mendel, L. L., Kapourchali, M. H., Dutta, J. K., Lee, S., Patro, C., and Pousson, M. (2016). “Identifying hearing loss from learned speech kernels,” in Proceedings of INTERSPEECH, pp. 243–247.

Scott, M. (2006) WordSmith Tools [available at http://lexically.net/wordsmith/index.html]

Teubert, Wolfgang & Anna Cermáková. 2007. Corpus linguistics. A Short Introduction. London: Continuum.

Tuohimaa, K., Loukusa, S., Löppönen, H., Välimaa, T., Kunnari S., (2022) Communication abilities in children with hearing loss–views of parents and daycare professionals. Journal of Communication Disorders 99, 106256.

Teubert, Wolfgang & Anna Cermáková. 2007. Corpus linguistics. A Short Introduction. London: Continuum.

Välimaa, T. T., Kunnari, S., Aarnisalo, A. A., Dietz, A., Hyvärinen, A., Laitakari, J., ... & Löppönen, H. (2022). Spoken language skills in children with bilateral hearing aids or bilateral cochlear implants at the age of three years. Ear and Hearing, 43(1), 220.

Thompson, P. (2004) Spoken Language Corpora in Wynne, M. (ed.) Developing Linguistic Corpora: a Guide to Good Practice [http://ahds.ac.uk/creating/guides/linguisticcorpora/index.htm]

Navigating the complexity of causality annotation in learner language

Miki, Nozomi (University of Birmingham) and Akira Murakami (Komazawa University)

While linguistic annotation plays a key role in learner corpus research, its use has been largely confined to lexical and syntactic annotation, likely because their annotation can be automated with fair accuracy (e.g., Huang et al., 2018). Discourse-level features in learner language bring interesting insights into second language development, but their annotation needs to be carried out mostly manually. In this talk, we introduce our annotation of cause-effect relationships in the International Corpus Network of Asian Learners of English (ICNALE, Ishikawa, 2023), focusing on inter-annotator reliability and the challenges in complex, discourse-level annotation tasks in learner language.

Specifically, we developed the coding scheme (eight types of tag pairs) about learners' cause and effect relationships in argumentative essays from the ICNALE. Once causal markers (e.g., because) were tagged with the <CAUSALITY> tag, related tags with the same id attribute were assigned to relevant textual segments; for example, <LOCAL> or <GLOBAL> tags were assigned according to the cohesive function. The following is an example of the annotation of an entire causal event:

<DIRECT id=01><LOCAL id=01><EFFECT id=01>I went home</EFFECT id=01><CAUSALITY id=01>because</CAUSALITY id=01><CAUSE id=01>I was ill.</CAUSE id=01></LOCAL id=01></DIRECT id=01>

Cause-effect annotation involves annotators' decisions, and their reliability should be examined. After comprehensive training, including quizzes, four annotators independently annotated the same subset of the data (106 essays and 24,665 words), representing 5% of the entire data set. We then calculated the pairwise inter-annotator agreement, allowing minor segmentation differences to be overlooked. The overall agreement was notably high. Specifically, the agreement for tags assigned to pre-determined, explicit lexical items (i.e., causal markers), exceeded 80%, while the agreement for tags assigned to non-prefixed segments (i.e., causes and effects) was lower, at 65 to 70%. In the presentation, we will explain the criteria used to define agreement and discuss other challenges encountered throughout the annotation process.

References

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. International Journal of Corpus Linguistics, 23(1), 28-54. doi:10.1075/ijcl.16080.hua

Ishikawa, S. (2023). The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English (Routledge).

Proficiency-dependent factors influencing L2 dative alternation

Murakami, Akira (University of Birmingham), Masato Terai (Aichi University of Technology), Yu Tamura (Kansai University) and Junya Fukuta (Chuo University)

Although factors influencing the speaker's choice between the double-object (DO) construction (e.g., John gave Mary a book.) and the prepositional object (PO) construction (e.g., John gave a book to Mary.) in dative alternation have been oft-studied in both first (e.g., Bresnan et al., 2007) and second language (e.g., Bernaisch et al., 2014; Gries & Deshors, 2020) acquisition research, we still know little about when second language (L2) learners come to be sensitive to those factors in their L2 use in the course of their development.

In order to address this question, we targeted 723,000 L2 English writings in the EF-Cambridge Open Language Database Cleaned Subcorpus (Shatz, 2020) and extracted 5,785 occurrences of DO and PO constructions with 23 verbs across CEFR A1–C1 levels and 10 L1 groups. Based on previous literature, we coded each occurrence for the length, pronominality, and animacy of theme and recipient. Additionally, we calculated the verb-level DO proportion (i.e., DO / (DO + PO)) from the Corpus of Contemporary American English as a measure of statistical preemption (Goldberg, 2011).

A mixed-effects regression model predicting the learner's choice between DO and PO largely confirmed our hypotheses, which were that (i) noun phrase length affects the choice from the early stage of development, (ii) pronominality influences it from the intermediate level, and (iii) animacy plays a role only at high proficiency levels. Contrary to our initial hypothesis and Goldberg's (2019) prediction, which suggested that statistical preemption would be found predominantly at higher proficiency levels, our results demonstrate its presence across various levels of proficiency. This points to the need for a multidimensional analysis of statistical preemption, incorporating further covariates and moderators to examine its effects.

References

Bernaisch, T., Gries, S. T., & Mukherjee, J. (2014). The dative alternation in South Asian English (es): Modelling predictors and predicting prototypes. English World-Wide, 35(1), 7-31.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G., Bouma, I., Kraemer, & J., Zwarts (Eds.), Cognitive foundations of interpretation (pp. 69-94). Amsterdam: KNAW.

Goldberg, A. E. (2011). Corpus evidence of the viability of statistical preemption. Cognitive Linguistics, 22(1), 131-153.

Goldberg, A. E. (2019). Explain me this: Creativity, competition, and the partial productivity of constructions. Princeton University Press.

Gries, S. T., & Deshors, S. C. (2020). There's more to alternations than the main diagonal of a 2× 2 confusion matrix: Improvements of MuPDAR and other classificatory alternation studies. ICAME Journal, 44(1), 69-96.

Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. International Journal of Learner Corpus Research, 6, 220–236.

Building the young learner corpus - A longitudinal, bilingual (English-Indonesian) corpus: Challenges of collecting data from different educational settings.

Mustun, Senora (Lancaster University), Mauselyn Pattikawa (SMK Negeri 4 Ambon, Indonesia), Yosef Tanggu Solo (SMPN 1 Wewewa Barat, Indonesia), Vonny Juliana Ruhulessin (SMAN 4 Maluku Tengah, Indonesia), Yeremina Karolina Ngabalin (SMA Santa Karya Langgur, Indonesia), Theresia Astrie Nunumete (SMK NEGERI 1 Ambon, Indonesia)

The learner corpus in teaching and learning languages has been widely recognised as important. The learner corpus provides insights into learners' language use, such as in writing. Most learner corpora are based on university students' written or spoken language (Gilquin & Granger, 2015; Ishikawa, 2023). Few have focused on multilingual settings and children's writings (Dirdal et al., 2022; Durrant, 2022). There is limited research on the long-term effects of the young learner corpus in the bilingual setting, specifically Indonesian learners of English as a foreign language (EFL) and those whose first language is Indonesian.

This project seeks to design and collect longitudinal data to build a bilingual young learner corpus and analyse learners' written language and vocabulary use. As this project is in its early phase, the current data comprises learners' meta-data: writing tasks written in English and Indonesian by learners aged 12 to 18 from five schools in East Indonesia. The aims of the study are: (i) to present the methodology of compiling the corpus; (ii) to report the challenges related to collecting data; and (iii) to report preliminary results of the analysis of the language and vocabulary used in learners' writing. The poster presents the data collection and corpus annotation processes, lists the challenges faced during data collection and describes how it was tackled. Additionally, LancsBox (Brezina & Platt, 2023) was used to generate a word list and keywords and examine concordance lines.

The preliminary results are grouped into two categories: (i) language use, e.g. grammatical errors; and (ii) vocabulary use, e.g. word choices and vocabulary level. The creation and analysis of the young learner corpus (bilingual [English/Indonesian]) hope to: (i) inspire teachers and learners to use the learner corpus in teaching and learning; and (ii) contribute to teacher training, teaching, learning materials and syllabus development.

References:

Brezina, V. & Platt, W. (2023) #LancsBox X [software], Lancaster University, http://lancsbox.lancs.ac.uk.

Dirdal, H., Hasund, I. K., Drange, E. M. D., Vold, E. T., & Berg, E. M. (2022). Design and construction of the Tracking Written Learner Language (TRAWL) corpus: A longitudinal and multilingual young learner corpus. Nordic Journal of Language Teaching and Learning (NJLTL), [P3]10(2), pp. 115-135.

Durrant, P. (2022). Studying children's writing development with a corpus. Applied Corpus Linguistics, 2(3), p. 100026.

Gilquin, G. & Granger, S. (2015). From design to collection of learner corpora. The Cambridge handbook of learner corpus research, 3(1), pp. 9-34.

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

A call for more data-oriented reporting in Learner Corpus research

Nicolas, Lionel (Institute for Applied Linguistics, Eurac Research, Italy), Egon W. Stemle (Institute for Applied Linguistics, Eurac Research, Italy), Magali Paquot (Institut Langage et Communication, Université catholique de Louvain, Belgium), Hubert Naets (Institut Langage et Communication, Université catholique de Louvain, Belgium) and Alexander König (CLARIN ERIC, The Netherlands)

Scientific communication, including in Learner Corpus research (LCR), has traditionally prioritized the reporting of scientific findings and the methods used to derive them. In recent years, however, there have been several calls for placing greater emphasis on detailing how the data used to derive the findings was handled, especially in fields related to computer science (Mitchell et al., 2019; Gebru et al., 2018; Dodge et al., 2019). These evolving trends align with broader initiatives, driven in great part by institutional policies, aimed at enhancing research outputs by promoting best practices with respect to openness, reusability and reproducibility (among other aspects). The most prominent outcome from these initiatives is certainly the growing popularity of the FAIR principles (Wilkinson et al., 2016).

Our presentation will assess data-oriented reporting practices in LCR, regardless of the primary scientific objectives of the data, and will also detail facts and figures explaining, when reported, how data was collected and handled. In that perspective, we are currently surveying the most recent LCR literature published over the past 5 years, as referenced in the learner corpus bibliography[3]. Our survey relies on the coding of each publication for a specific set of features related to, at present, the reporting of formats, tools and licenses used.

Our preliminary findings suggest that progress can still be made. Building on these findings, we would like to open a discussion on future data-oriented reporting practices in LCR and contextualize our findings with respect to recent developments in Natural Language Processing, a computer-science-oriented research area with well-known ties to LCR. In particular, we will discuss the recent and noteworthy development of the “Responsible NLP Research Checklist”, which represents a promising and relevant basis to build upon for advancing data-reporting practices in LCR.

References

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

---

[3] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpus-bibliography.html

SEEFLEX – The Corpus of Secondary School English as a Foreign Language (EFL) Exams

Pauls, Tobias (RWTH Aachen University)

Learner corpus research has been a well-investigated field for more than thirty years. It is usually described as a discipline located in between corpus linguistics and second language acquisition (Granger, 2009). Over the decades, numerous learner corpora have been compiled, covering a variety of languages (see e.g. Learner Corpora around the World, 2024). For written corpora, university level writing (tertiary education) has usually been the focus.

The SEEFLEX corpus was developed as part of a dissertation project. Pilot studies offered initial insights into secondary school writing in EFL classes in Germany. In their exams, students in the final three grades of secondary school encounter unique yet recurring tasks anchored in curricular requirements regarding targeted text types and register knowledge. Research is needed on a larger scale to understand how students complete these tasks linguistically and whether they meet these requirements.

This corpus includes data from 45 classes across those three grade levels at three schools in North Rhine-Westphalia. Authentic curriculum-based hand-written examinations were scanned, fully transcribed, and POS-tagged. Extended xml-markup was added to account for inter alia language mistakes, quotes, references, and text structure. Students also participated in a 90-minute contact session. The collected data include standardized receptive vocabulary assessments, students' reading habits, a cognition scale, and the participants' language experience and proficiency and social background.

SEEFLEX features more than 620.000 tokens across 1967 writing samples (mean wordcount = 317.83; median = 267; SD = 165.67; range = 17-1552; 3-4 texts per participant from 5 possible curriculum-based tasks) by 572 different students. The corpus will function as a pedagogical resource to analyse the linguistic differences between the curricular tasks as well as the official requirement that these tasks be distinctive from one another. Access to the corpus will be available via CQPweb. An online supplement will provide additional resources.

References:

Centre for English Corpus Linguistics. (2024). Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Granger, Sylviane. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In: Aijmer, K., Corpora and Language Teaching, Benjamins: Amsterdam and Philadelphia, p. 13-32.

The effect of different feedback strategies on lexical sophistication of ESP writing

Pojslová, Blanka (Masaryk University, Brno)

This paper presents a study investigating the effect of two feedback strategies on ESP learners´ writing quality from the perspective of lexical sophistication which together with lexical density and lexical diversity constitutes the construct of lexical complexity in the linguistic complexity component of CAF (Bulté and Housen, 2014). The study takes the form of a pre-test/post-test quasi-experiment with two comparison groups. The first group (33 participants) received teacher-only feedback, while the second group (32 participants) received combined peer-teacher feedback. The pre-test and post-test learner corpora were collected and analysed using TAALES (Tool for Automatic Analysis of Lexical Sophistication) to measure changes in 14 indices of lexical sophistication which have been proven to be indicative of L2 writing quality (Durant et al., 2019; Kyle et al., 2018).

The study yielded four major findings. First, the results indicate improvements in writing quality in both comparison groups, and the majority of these changes were statistically significant. Second, the effectiveness of the feedback strategy measured by the effect size seemed to differ depending on the source of feedback, and the index of lexical sophistication. Third, the teacher-only feedback seemed to contribute to more homogenous writing production than the combined peer-teacher feedback. Fourth, there is no significant difference in post-test writing between the comparison groups.

References:

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

Durrant, P., Moxley, J., & McCallum, L. (2019). Vocabulary sophistication in First-Year Composition assignments. International Journal of Corpus Linguistics, 24(1), 33-66.

Kyle, K., Crossley, S. A., & Berger, C. (2018). The Tool for the Automatic Analysis of Lexical Sophistication Version 2.0. Behavior Research Methods, 50 (3),1030-1046.

Capitalisation and consonant doubling in German as a foreign language – an error analysis of learner texts at different CEFR proficiency levels

Reitbrecht Sandra

Studies emphasise the relevance of orthographic competence for reading and pronunciation learning in a foreign language (Hayes-Harb & Barrios, 2021; Jeon & Yamashita, 2014) additionally to its contribution to accuracy in writing. However, in research on German as a Foreign Language (GFL), the questions how GFL learners (with literal competence in at least one other language) acquire orthographic competence and how they can be supported in this process have so far hardly been addressed (Hirschmann, 2007; Thurmair, 2022).

Thus, the submitted paper focuses on orthographic errors in GFL learners' writing performances and presents results on the orthographic accuracy of learner texts at different proficiency levels of the Common European Framework of Reference for Languages (CEFR). The analysis is based on 785 texts (101.496 tokens) from the MERLIN learner corpus for German (Wisniewski et al., 2018) that are provided in an error-annotated version on the integrated web-based search engine ANNIS (Krause & Zeldes, 2016). The errors in two frequently occuring error areas in the corpus (capitalisation and sharpening/consonant doubling) have been submitted to a detail analysis taking into consideration influencing linguistic factors to error susceptibility. The aim of the study is to answer the following research questions: What kinds of errors do GFL learners make in the selected error fields of German orthography, and do error rates differ between the different CEFR proficiency levels? Furthermore, influencing intra- and interlingual factors on error susceptibility are explored.

The preliminary results show a considerable number of over- und under-capitalisations as well as error-free texts from level A1 on. A Kruskal-Wallis test reveals a significant effect of language proficiency on error rate. The analysis for consonant doubling in German is still work in progress. The paper will present results on both orthographic aspects and discuss them with regard to further studies and practical teaching implications for GFL.

References

Hayes-Harb, R. & Barrios, S. (2021). The influence of orthography in second language phonological acquisition. Language Teaching, 54 (3), 297–326.

Hirschmann, H. (2009). Orthographiefehler bei fortgeschrittenen Lernern des Deutschen als Fremdsprache (Vortragsfolien). https://publikationen.ub.uni-frankfurt.de/opus4/frontdoor/deliver/index/docId/12443/file/hirschmann_fo lien.pdf

Jeon, E. H. & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. Language Learning, 64 (1), 160–212.

Thurmair, M. (2022). Interpunktion – (K)ein Thema für Deutsch als Fremdsprache?. In P. Rössler, P. Besl & A. Saller (Hrsg.), Vergleichende Interpunktion – Comparative Punctuation (S. 317–341). De Gruyter.

Wisniewski, K., Abel, A., Vodičková, K., et al. (2018). MERLIN Written Learner Corpus for Czech, German, Italian 1.1. Eurac Research CLARIN Centre. http://hdl.handle.net/20.500.12124/6

Verb phrase versatility as a syntactic complexity indicator in L2-English written texts

Rényi, Andrea Ágnes (Pázmány Péter Catholic University, Budapest)

A recurring language proficiency assessment issue is how to conceptualise Common European Framework of Reference for Languages (CEFR) proficiency levels in quantifiable features of English grammar and vocabulary. That is, what are the characteristics of a certain CEFR level, in terms of its syntactic/lexical accuracy and complexity? As part of the validation process of a CEFR B2+ language exam for English majors at a Hungarian university, we are working to detect the systematic patterns of syntactic and lexical characteristics of a written corpus (N=500) and their match to B2+ expectations. The project research question is (1) whether that exam measures English language proficiency at the B2+ level in a valid and reliable way, and (2) how to find ways to partly automatically assess those texts. In that framework the present research question is whether verb phrase versatility is a reliable indicator of syntactic complexity.

Verb phrase versatility means the range and types of tense, aspect, voice, modality and finiteness variation of verb phrases in English texts. While in everyday teaching and learning using verb phrases that go beyond the Simple Present and the Simple Past is a certain sign of text production development above the B1/B2 level, none of the multivariate analytical systems seem to grasp this variable phenomenon as a single indicator, including the Biber-tagger/MAT (Nini 2019), the L2SCA (Lu 2017) or the CVLA (Uchida & Negishi 2018). In this project, we have been working with each of these, adding some of our own manually and automatically detectable variables. A statistical meta-analysis across these systems will be suggested in the talk, comparing the results to those on verb phrase versatility.

References

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

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In Berber Sardinha, T. & Veirano Pinto M. (eds), Multi-Dimensional Analysis: Research Methods and Current Issues, Bloomsbury Academic, 67-94.

Uchida, S. & M. Negishi (2018). Assigning CEFR-J levels to English texts based on textual features. In Y. Tono and H. Isahara (eds.) Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC 2018), pp. 463-467.

The development of complexity, accuracy and fluency in L2 English writing at secondary school–group and individual learning profiles.

Rokoszewska, Katarzyna (The Institute of Linguistics Jan Dlugosz University in Czestochowa, Poland)

According to Complex Dynamic Systems Theory (CDST), language development is a complex, dynamic, variable and individually owned process. CDST proponents claim that studies should draw from both group-based and individual-based data since research results should not be generalised from the group to the individual and vice versa unless the group is an ergodic ensemble (Lowie & Verspoor, 2018). Moreover, CDST researchers argue that group-based results represent a hypothetical average learner who in fact does not exist as virtually no learner is exactly average (Hiver & Al-Hoorie, 2020).

The aim of the presentation is to describe a sequential mixed method (MM) study in which panel data were analysed with respect to individual learners. The aim of the study was to examine language development in L2 English writing at secondary school in the case of individual learners in comparison to the whole group. The study focused on eleven variables, such as syntactic complexity, including subordination, coordination, and nominalisation, lexical complexity, including lexical density, sophistication, and variation, accuracy, fluency, and general language development. It was based on The Written English Developmental Corpus of Polish Learners (WEDCPL) which consists of over 1900 essays written by 100 learners during 21 data waves organised over the period of three years at secondary school.

The results of the study indicated that the individual learners' average results rarely differed from the group in statistically significant ways, but their learning profiles connected with progress in syntactic complexity, accuracy, lexical complexity, and fluency (CALF) over time were different than the group profile.

References

Hiver, P. & A. Al-Hoorie (2020). *Research Methods for Complexity Theory in Applied Linguistics*. Bristol: Multilingual Matters.

Lowie, W. & M. Verspoor (2018). Individual differences and the ergodicity problem. *Language Learning, 69,* 184–206.

Syntactic-complexity development in Norwegian learner English

Rørvik, Sylvi (Inland Norway University of Applied Sciences)

The impetus for the present paper came from a longitudinal study of noun-phrase (NP) complexity in texts written by Norwegian learners of English across Years 8-10 (Rørvik 2022), where it was shown that the frequency of complex NP increases over time, but the sophistication in the phrasal modification does not, as measured according to the developmental-stages framework proposed by Biber et al (2011). In line with previous research, however, there were indications of individual developmental trajectories (cf. Kreyer & Schaub 2018; Díez-Bedmar & Pérez-Paredes 2020).

The present study examines Norwegian learners at a slightly later stage in their development, to attempt to pinpoint where the hypothesized increase in complexity sophistication takes place. Thus, pre-tertiary argumentative texts from Years 10 (pupils aged 15-16) and 11 (pupils aged 16-17) (both text categories from the TRAWL corpus (Dirdal et al 2022)) and undergraduate texts (from the Norwegian component of ICLE (Granger et al 2020)) were examined with regard to their use of the nominal and clausal features described by Biber et al (2011). The results resemble those of the initial study in that there is little evidence of a shift towards more sophisticated complexity features, be they nominal or clausal. This is perhaps not surprising, given previous research that indicates that this shift tends to take place between undergraduate and graduate level (Parkinson & Musgrave 2014, Staples et al 2016; Biber et al 2020; Staples et al 2023). However, Gray et al (2019) found increased sophistication in the development of pre-tertiary Chinese L2 learners of English over a time span of 9 months, and in combination with the previous findings indicating the prevalence of individual developmental trajectories, the present paper also investigates a subset of the writers for evidence of these, with a view to providing more focused pedagogical advice to further the writers' syntactic-complexity development.

References

Biber, D., B. Gray, & K. Poonpon. 2011. "Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development?" TESOL Quarterly, 45:1, 5-35.

Biber, D., R. Reppen, S. Staples, & J. Egbert. 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.

Díez-Bedmar, M. B. & P. Pérez-Paredes. 2020. "Noun phrase complexity in young Spanish EFL learners' writing. Complementing syntactic complexity indices with corpus-driven analyses." International Journal of Corpus Linguistics 25:1, 4-35.

Dirdal, H., I. K. Hasund, E.-M. Drange, E. Vold, & E.-M. Berg. 2022. "Design and construction of the Tracking Written Learner Language (TRAWL) corpus: A longitudinal and multilingual young learner corpus." Nordic Journal of Language Teaching and Learning 10(2), 115–135.

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

Gray, B., J. Geluso, & P. Nguyen. 2019. "The Longitudinal Development of Grammatical Complexity at the Phrasal and Clausal Levels in Spoken and Written Responses to the TOEFL iBT® Test (TOEFL Research Report No. RR-90)." Educational Testing Service.

Kreyer, R. & S. Schaub. 2018. "The development of phrasal complexity in German intermediate learners of English." International Journal of Learner Corpus Research 4:1, 82-111.

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

Rørvik, Sylvi. 2022. "Noun-phrase complexity in the texts of intermediate-level Norwegian EFL writers: stasis or development?" Nordic Journal of Language Teaching and Learning 10(2), 298-326.

Staples, S., J. Egbert, D. Biber, & B. Gray. 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., B. Gray, D. Biber, & J. Egbert. 2023. "Writing Trajectories of Grammatical Complexity at the University: Comparing L1 and L2 English Writers in BAWE." Applied Linguistics 44(1), 46-71.

Automatic generation of target hypotheses for learner language

Ruppenhofer, Josef (FernUniversität in Hagen), Torsten Zesch (FernUniversität in Hagen), Katrin Wisniewski (Universität Leipzig) and Anette Portmann (Universität Leipzig)

A key aspect of learner corpus and SLA research consists in identifying and classifying learner errors (Corder 1967). While one cannot be certain about what a learner was trying to say, error analysis presupposes making assumptions about learners' intended productions. In learner corpus research, annotations of learners' aimed-for productions are referred to as target hypotheses (THs) (Ellis, 1994; Lüdeling & Hrischmann, 2015). While THs are needed for error analyses to be comprehensible and reproducible, few corpora actually include THs. The best-known exceptions are the Falko corpora (Lüdeling et al., 2008) and closely related others such as Merlin (Wisniewski et al., 2013).

Manually producing THs is challenging since usually multiple THs are possible for the same sentence (Reznicek et al., 2013). More minimal THs stay closer to the learners' surface forms while more extended THs get closer to smooth L1-appropriate productions. But even with more constraining guidance on constructing THs, researchers often still come up with different THs. Nevertheless, learner corpora often feature only one TH or at most two, one minimal and the other extended.

To address these challenges, we explore the use of Natural Language Processing for the purpose of automatically creating THs. Not only are NLP methods cheaper and faster in producing THs. They also allow us to create multiple versions of THs, which will be similar but also different in some respects, hopefully mirroring human variation.

In our talk, we report on our work on automatically creating THs. We discuss which NLP methods are best suited to create THs that closely mimic manually constructed ones. To evaluate the quality of automatic THs, we compare new automatic THs for the Falko corpora to available manual THs in these corpora. We further analyze the quality of automatic THs with respect to variables like CEFR level or corpus type.

References

Corder, S. P. 1967. "The Significance Of Learner's Errors". International Review of Applied Linguistics in Language Teaching, vol. 5, no. 1-4, pp. 161-170.

Ellis, Rod 1994. The Study of Second Language Acquisition. Oxford: Oxford University Press.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K. & Walter, M. 2008 Das Lernerkorpus Falko. In: Deutsch als Fremdsprache 2(2008), 67-73.

Lüdeling, A., & Hirschmann, H. 2015. Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The Cambridge Handbook of Learner Corpus Research (Cambridge Handbooks in Language and Linguistics, pp. 135-158). Cambridge: Cambridge University Press.

Reznicek, M., Lüdeling, A., & Hirschmann, H. 2013. Competing target hypotheses in the Falko corpus. Automatic treatment and analysis of learner corpus data, 59, 101-123.

The Use of Punctuation in a German-to-Basque LTC

Sanz Villar, Zuriñe (Universidad del País Vasco / Euskal Herriko Unibertsitatea) (UPV/EHU)

Several members of the TRALIMA-ITZULIK research group joined the MUST initiative in 2019, and since then learner translation corpora (LTC) in different language combinations have been compiled and annotated at the University of the Basque Country. In a study of a German-to-Basque LTC (Sanz-Villar, 2024), which contained 1214 annotations at the time of analysis, it was found that linguistic errors (609 in total) — i.e. segments that are incorrect from the perspective of the target language — outnumbered ST-TT transfer errors (436 in total). Of the 609 linguistic errors, 91 were tagged as punctuation errors. In another analysis of an error-tagged LTC from Czech into English, it was highlighted that “determiners and punctuation seem to be global problems and need to be addressed in the curriculum” (Fictumova *et al.*, 2017: 225). However, as mentioned by Malmkjær (1997), Rodríguez-Castro (2011: 43) and Wang (2018), although punctuation is an important part of translation, it has not received much attention from professionals and researchers in the field of Translation Studies.

The aim of this study is to analyse the behaviour of trainee translators with regard to punctuation. To this end, a parallel corpus of German literary and instructional texts and translations into Basque will be compiled and annotated using version two of the Translation-oriented Annotation System (Granger & Lefer, 2021). Punctuation will be collected from both STs and TTs, and behaviour regarding the use of punctuation marks will be analysed in both student and professional translators’ output, as the corpus will contain a published translation for each of the STs. The identification of errors will help to identify difficulties and to create materials to improve the contrastive competence of translation students in the German-Basque translation classroom. A descriptive analysis of the use of punctuation will shed light on the influence that the ST can have on translations.

References

Fictumova, J., Obrusnik, A., & Stepankova, K. (2017). Teaching Specialized Translation. Error-tagged Translation Learner Corpora. Sendebar, 28, 209–241.

Granger, S. & Lefer, M.-A. (2021). Translation-oriented Annotation System manual (Version 2.0). CECL Papers 3. Centre for English Corpus Linguistics/Université catholique de Louvain.

Malmkjær, K. (1997). Punctuation in Hans Christian Andersen’s stories and in their translations into English. In F. Poyatos (Ed.), Nonverbal Communication and Translation (151–162). John Benjamins.

Rodríguez-Castro, M. (2011). Translationese and punctuation. An empirical study of translated and non-translated international newspaper articles (English and Spanish). Translation and Interpreting Studies, 6(1), 40–61.

Sanz-Villar, Z. (2024). German-to-Basque Translation Analysis of Multiword Expressions in a Learner Translation Corpus. Íkala, Revista De Lenguaje Y Cultura, 29(1), 1–21.

Wang, C. (2018). Decoding and Encoding the Discourse meaning of Punctuation: a Perspective from English-to-Chinese Translation. Babel, 64(2), 225–249.

Compiling a multi-corpus database for automatically annotating developmental stages in L2 German

Schwendemann, Matthias (Leipzig University), Katrin Wisniewski (Leipzig University), Torsten Zesch (FernUniversität Hagen), Lisa Lenort (Herder Institute, Leipzig University)

The step-wise acquisition of verb placement in developmental stages has received much attention in SLA research (e.g., Bettoni & Di Biase, 2015; Lenzing et al., 2019; Pienemann, 1998, 2005). Recent research, particularly from the Complex Dynamic Systems approach (e.g., Verspoor & Lowie, 2021), questions the robustness of prior findings, showing considerable learner language variation across, but importantly also inside stages. In addition, and in contrast to theoretical assumptions, the acquisition of verb placement might depend on a number of influencing factors (e.g., age, L1). This has led to a considerable tension between theoretical frameworks and to uncertainty regarding the validity of the developmental stages approach. However, as yet, most studies rely on small data bases and do not work with publicly accessible learner corpora, which makes in-depth research and replication efforts difficult.

In our presentation, we introduce our interdisciplinary current research project (2022-2025) which, first, on an infrastructure level aims at compiling a large database consisting of more than 40 German learner corpora that we are converting into a common data format (CAS) and for which we designed a comprehensive metadata scheme. This diverse database will be made publicly available. This will allow to take into account potential factors for variation in L2 acquisitional trajectories of developmental stages in cross-corpus analyses. Secondly, the project aims at developing tools to automatically analyze the developmental stages in the learner data. The project adopts an iterative and flexible workflow including quantitative and qualitative linguistic checks and feedback loops to approach this explorative and demanding task.

In our talk, we will summarize the project design, its aims and the results achieved thus far. The main focus of the talk will be on challenges regarding the theory-based computational linguistic operationalization of developmental stages and on methodological issues of automatic learner data analysis.

References

Bettoni, C., & Biase, B. D. (eds). (2015). Grammatical development in second languages. Exploring the boundaries of Processability Theory. Lulu.

Lenzing, A., Nicholas, H., & Roos, J. (eds.). (2019). Widening contexts for Processability Theory: Theories and issues. Benjamins.

Pienemann, M. (1998). Language processing and second language development. Processability theory. Benjamins.

Pienemann, M. (ed.). (2005). Cross-linguistic aspects of processability theory. Benjamins.

Verspoor, M., & Lowie, W. (2021). Complex Dynamic Systems Theory and Second Language Development. In H. Mohebbi & C. Coombe (Hrsg.), Research Questions in Language Education and Applied Linguistics: A Reference Guide (S. 799–803). Springer.

Motion Verbs in the Second/Foreign Language Acquisition of Czech: a corpus-based study on non-native speakers of Czech with Chinese L1

Škodová, Svatava (Faculty of Arts, Charles University) and Melissa Shih-hui Lin (Department of Slavic Languages and Literatures National Chengchi University)

The study aims to uncover how Czech native speakers (CzNS) differ from non-native speakers of Czech with Chinese L1 (CzNNS) in terms of the lexicalization of motion events.

Czech is an inflected language with rich morphology, it belongs to the West Slavic languages (Sussex, 2011). Chinese is an analytic language (Croft, 1990), with mostly invariant monosyllabic words; it uses grammatical words and word order to express word relations in a sentence (Tai, 1985; Liu, 2015). Semantically, Chinese is considered a verb-serializing and Czech a satellite-framed language (Slobin, 2004).

Our analysis examines the range and types of verbs and their grammatical characteristics employed to express motion events: static and dynamic verb ratio, serial verb constructions, verbal prefixation, and the manifestation of satellite-framing vs. verb-serializing language characteristics in the narratives.

The quantitative and qualitative analyses are based on the acquisition patterns by NS and NNS, resp. L2 learners, studied in Czech and Chinese learner corpora, CzeSL and LCSL.

We relied on the data elicited through the "Frog story" (Mayer 1969) from comparable groups of 45 CzNS, 26 ChNS and 45 CzNNS.

Texts produced by CzNS exhibit a notably higher degree of dynamism. Utilization of static and dynamic motion verbs is 2.5 times higher among CzNS than among CzNNS. CzNS encode the direction of motion directly in the verb, employing prefixes.

ChNS express the motion within the motion verb clusters, as a serial verb construction; the verb-external participants or motion-related components are ordered in encoding a single or multiple components of motion.

CzNNS significantly coordinate motion verbs compared to CzNS. They employ reduplication coordination to convey the intensity of an action, which CzNS code on the prefix, and use contact verbal coordination, a schema not employed by CzNS. The prefixation of verbs and subsequently the specification of motion direction is underrepresented.

References:

Croft, W. (1990). Typology and Universals. Cambridge, England: Cambridge University Press.
Liu, M. C. (2015). How to make sense of Chinese? – Start with Chinese Function Words. Journal of Chinese Language Teaching, 12(3), 31-52.
Mayer, D. (1969). Frog, where are you? New York: Dial Books for Young Readers.
Slobin, D. I. "The many ways to search for a frog: Linguistic typology and the expression of motion events". Typological and contextual perspectives, Psychology Press, 2004. pp. 219-257. Web. Retrieved from https://www.taylorfrancis.com/chapters/edit/10.4324/9781410609694-12
Sussex, R. & Cubberley, P. (2011). The Slavic Languages. Cambridge Language Surveys.
Tai, James H-Y. (1985). Temporal sequence and Chinese word order. In J. Haiman (Ed.),Iconicity in Syntax (pp. 49-72). Amsterdam/Philadelphia: John Benjamins Publishing Co.
Corpora:
CzeSL (Czech as a Second Language) Available on-line: http://www.korpus.cz
Šebesta, K. – Bedřichová, Z. – Šormová, K. – Štindlová, B. – Hrdlička, M. – Hrdličková, T. – Hana, J. – Petkevič, V. – Jelínek, T. – Škodová, S. – Poláčková, M. – Janeš, P. – Lundáková, K. – Skoumalová, H. – Sládek, Š. – Pierscieniak, P. – Toufarová, D. – Richter, M. – Straka, M. – Rosen, A.: CzeSL-SGT: CzeSL-SGT – a corpus of non-native speakers' Czech with automatic annotation, version 2 from 28 Sep 2014. Ústav Českého národního korpusu FF UK, Praha 2014.
LCSL (NCCU Learner Corpus Of Slavic Languages) Available on-line: https://lcsl.nccu.edu.tw
Lin, M. S.-H., Yeh, H.-L. & Lin, Y.-L. (2021. 08) NCCU Learner Corpus of Slavic Languages (LCSL). Paper presented at the X. International Symposium on Czech as a Foreign Language, Institute of Czech Studies, Charles University.

Triangulation with learner translation corpora

Skogmo, Siri Fürst and Susan Nacey

This paper discusses the utility of combining learner corpus data with other data types to gain a fuller picture of how L2 language learners novice translators handle translation challenges. Our learner corpus data consists of multiple translations of an especially metaphor-dense source text (horoscopes) from the learners' L1 Norwegian into L2 English. In addition, we collected individual student reflection notes about their translations, and observed their classroom discussion about the different translation solutions.

We discuss the translation of one particularly culture-specific metaphor as an illustrative example to show which conclusions may be drawn on the basis of the three individual data sources. First, the corpus data allows us to analyze the learners' translation strategies (i.e. what they did). Second, individual students' retrospective reflections provide insight into their motivations for translation choices and perceived challenges (i.e. why they did what they did). Third, observation of classroom discussion (transcribed into a collaborative translation protocol) allows for insight into their mutual understanding of successful translation solutions (i.e. how they evaluate what they did).

Our study highlights the added value of adopting methodological triangulation, thus allowing for the consolidation of findings from learner corpus research with findings from other data types. Our findings are discussed in terms of validity (where results based on different data types corroborate each other), completeness (where results from different data types create a full(er) picture), and complementarity (where findings from one data type raises hypotheses answered by findings from another). The findings from three individual data sources thus complement each other to form a fuller picture of the translation process and what may be regarded as a 'successful' translation. Through this triangulation, we gain a multifaceted picture of language learners' decision-making in translating in general, and translation of culture-specific metaphors in particular.

Intensification of adjectives in young L2 learners of German and Italian

Spina, Stefania (Università di Stranieri di Perugia), Aivars Glaznieks (Eurac Research Bolzano) and Andrea Abel (Eurac Research Bolzano/ Free University of Bolzano)

Intensification is the use of any linguistic device that scales a quality, by establishing different degrees of that given quality (Bolinger 1972). Previous studies in second language acquisition investigated adjective intensification in advanced learners of English (e.g., Lorenz, 1998), while few have focused on other languages (Hendrikx et al. 2019), and on younger learners (e.g., Hasselgård 2022). Some of their most relevant results highlight learners' overuse of all-purpose, delexicalized adverbial intensifiers (e.g., very).

We will present a comparative study in which we analyzed the use of the adjective intensification construction [[X]int [Y]ADJ]AP 'very Y' in the Italian and German subcorpora of Kolipsi (Glaznieks et al. 2023), a learner corpus collection for L2 Italian and German. The corpus consists of written essays (with about 570,000 tokens) from the multilingual Italian province of South Tyrol. Applying a Diasystematic Construction Grammar approach (Höder et al. 2021), which allows us to distinguish idio- from diaconstructions of Italian and German, we replicated previous research from a different multilingual area (Van Goethem & Hendrikx 2021) and analysed the occurrence of the [[X]int [Y]ADJ]AP 'very Y' construction on different levels of schematicity using mixed-effect models.

Our research question was: Are there differences in the way L2 Italian and L2 German young learners from the multilingual province of South Tyrol use adjective intensification in written essays?

We found the main difference between learners of Italian and German on the most abstract level of analysis, where we investigated the effect of different variables (e.g., L1, linguistic environment) on the choice of morphological ([[X]AFFIX [Y]ADJ]ADJ) or syntactic intensification types (e.g., [[X]ADV [Y]ADJ]AP). For L2 Italian learners, the linguistic environment is a significant predictor for their choice (Spina et al., forthcoming). L2 German learners prefer intensifying adverb constructions regardless of their L1 or linguistic environment.

References:

Bolinger, D. (1972). Degree Words. Berlin, Boston : De Gruyter Mouton.

Glaznieks, A., Frey, J.-C., Abel, A., Nicolas, L. & Vettori, C. (2023). The Kolipsi Corpus Family. Resources for learner corpus research in Italian and German. Italian Journal of Computational Linguistics 9(2). https://doi.org/10.4000/ijcol.1210

Hasselgård, H. (2022). Adverb-adjective combinations in young writers' English (EL1 and EL2). Nordic Journal of Language Teaching and Learning, 10(2).

Hendrikx, I., Van Goethem, K. & Wulff, S. (2019). Intensifying constructions in French- speaking L2 learners of English and Dutch: cross-linguistic influence and exposure effects. International Journal of Learner Corpus Research, Vol. 5:1, pp. 63–103.

Höder, S., Prentice, J. & Tingsell, S. (2021). Additional language acquisition as emerging multilingualism. In: H. C. Boas & S. Höder (eds.): Constructions in Contact 2: Language change, multilingual practices, and additional language acquisition. Amsterdam, Philadelphia: Benjamins, pp. 309-337.

Lorenz, G. (1999). Adjective intensification - learners versus native speakers: A corpus study of argumentative writing. Amsterdam: Rodopi.

Pérez-Paredes, P. & Díez-Bedmar M.B. (2012). The Use of Intensifying Adverbs in Learner Writing. In Y. Tono, Y. Kawaguchi & M. Minegishi (eds), Developmental and Crosslinguistic Perspectives in Learner Corpus Research, Amsterdam: Benjamins, pp. 105–124.

Spina, S., Glaznieks, A. & Abel, A. (forthcoming): Intensification in written L2 Italian: Insights from the multilingual region of South Tyrol. International Journal of Learner Corpus Research.

Van Goethem, K. & Hendrikx, I. (2021). Intensifying constructions in second language acquisition. A diasystematic-constructionist approach. In: H. C. Boas & S. Höder (eds.): Constructions in Contact 2: Language change, multilingual practices, and additional language acquisition. Amsterdam, Philadelphia: Benjamins, pp. 376–428.

Reporting verbs in Norwegian undergraduate learner English

Thormodsæter, Øyvind (Oslo Metropolitan University) (OsloMet)

Although reporting verbs have been investigated from various theoretical and taxonomical angles, there is, to the best of my knowledge, little specifically on Norwegian learner use of English reporting verbs. Moreover, Kwon et al. (2018: 1) point out that "[…], relatively few quantitative studies have focused on the use of reporting verbs by undergraduate students […]".

Aiming to diminish these gaps, the current presentation reports on the use of reporting verbs in English texts written by L1 Norwegian learners, as well as the phraseological patterns in which the verbs occur. More specifically, the study investigates the following research questions:

1. Which reporting verbs do L1 Norwegian undergraduate students use, and what phraseological patterns do the reporting verbs occur in?
2. How are the reporting verbs distributed across different text types and semantic categories?

Theoretical framework

Inspired by Kwon et al. (2018), the framework in the current study is based on Charles' (2006, p. 319) adaptation of Francis et al.'s (1996, p. 97-101) taxonomy for the classification of reporting verbs, viz. ARGUE, THINK, SHOW and FIND. However, since the investigation is bottom-up, the taxonomy may need further adaptation.

Material and method

The Oslo Corpus of Learner English Texts (OCLET) contains 282 texts belonging to different genres, written by Norwegian undergraduate students (age 17-19) in their final two years of upper secondary education. Texts are POS-tagged using TagAnt and investigated in AntConc, where all examples of a lexical verb followed by 'that' are extracted and manually sorted.

Preliminary findings

Preliminary searches support other findings suggesting that learners tend to favour more general reporting verbs, such as 'say', 'think' or 'state', but the material contains a variety of reporting verbs, as well as some arguably unidiomatic/creative use, such as 'The book tries to say that' or 'It is important to state that'.

References

Anthony, L. (2022). TagAnt (Version 2.0.5) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Charles, M. (2006). "Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines". English for Specific Purposes, 25(3), 310-331. https://doi.org/10.1016/j.esp.2005.053.

Francis, G., S. Hunston, E. Manning. Collins COBUILD grammar patterns 1: Verbs. Harper Collins, London (1996)

Kwon, M. H., S. Staples & R. S. Partridge (2018). "Source work in the first-year L2 writing classroom: Undergraduate L2 writers' use of reporting verbs". Journal of English for Academic Purposes 34, 86-96. https://doi.org/10.1016/j.jeap.2018.04.001

The use of English articles in essays written as part of the Estonian National Examination of English

Torn-Leesik, Reeli and Liina Tammekänd (University of Tartu)

Several studies (Miller 2005, Narita 2013) have shown that the English article system poses challenges to learners of English. Those learners of English whose L1 lacks articles often misuse English articles (Masters 1997, Barrett and Chen 2011). Earlier studies on Estonian learners of English confirm this finding. For instance, both Torn and Kährik (2004) and Konso (2023) have found that Estonian learners of English tend to overuse the definite article with both non-count abstract and generic nouns. In addition, Konso (2023) highlighted issues related to cataphoric reference by means of the definite article and nonreferring uses of the indefinite article. Both studies focussed on upper-intermediate and advanced students but differed in their methodology. While Torn and Kährik (2004) used a gap filling test, Konso (2023) conducted a corpus study of university entrance examination essays constructed as a reading-to-write task.

The present study will explore the use of English articles in essays written for the Estonian National Examination of English. Such essays are normally at CEFR levels B1-C1. Unlike the material studied by Torn and Kährik (2004) and Konso (2023), the national examination essays are written without any base text. The aim of the study is to explore whether similar patterns of article misuse occur in essays that are not of a reading-to-write task type and whether the national examination task elicits other types of article misuse. For the analysis, 250 essays will be tagged using the CLAWS tagger. AntConc will be used to compile concordances and frequency lists for automatic identification of relevant grammatical structures. To ensure correct analysis, manual identification will also be implemented.

References:

Barrett, Neil Edward and Li-mei Chen. 2011. English Article Errors in Taiwanese College Students' EFL Writing. International Journal of Computational Linguistics & Chinese Language Processing, 16: 3–4, 1–20.

Konso, Johanna, 2023. Estonian EFL learners' misuse of the English article system: A corpus-based study. Unpublished MA thesis. Department of English, University of Tartu, Tartu, Estonia.

Master, Peter. 1997. The English article system: Acquisition, function, and pedagogy. System, 25: 2, 215–232.

Miller, Julia. 2005. Most of ESL students have trouble with the articles. International Education Journal, 5: 5, 80–88.

Narita, Masumi. 2013. The use of articles in Japanese EFL learners' essays. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier (eds). Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead, 357–366. Louvain-la-Neuve: Presses universitaires de Louvain.

Torn, Reeli and Kaja Kährik. 2004. Lost in world of the articles. A study of the use of articles by Estonian learners of English. In Neil Murray and Tony Thorne (eds). Multicultural Perspectives on English Language and Literature: International conference; Tallinn; 22-23.05.2003, 179−185. Tallinn, London: Tallinna Pedagoogikaülikooli Kirjastus.

Compiling oral learner corpora: Is automatic transcription really worth it?

Vandeweerd, Nathan (Radboud University)

Until recently, the cost of manual transcription has made it prohibitively expensive to compile large corpora of spoken L2 data. Recent advances in automated transcription however, may offer a solution to this problem. These tools have been shown to have very high levels of reliability, sometimes even outperforming human transcribers (Xiong et al., 2018). Yet, research has also shown that factors such as L1-background or age of acquisition can affect transcription quality (Chan et al., 2022), which means that automatically generated transcriptions still need to be verified and manually corrected before being used for research purposes.

This presentation reports on a study investigating the practical usefulness of using Sonix software to transcribe L2 French oral exams. As part of a larger project, a subset of exams (N = 45) were automatically transcribed using Sonix and then manually corrected and annotated for fluency phenomena (e.g., filled pauses, repetitions, etc.). A smaller subset of exams (N = 12) were then manually transcribed and annotated by the same human transcriber. The time per text (normalized per 100 words) was compared between the two methods as well as the correlation between each method and the proficiency scores given to the exams to determine the relationship between L2 proficiency and transcription time.

The results revealed that using Sonix to transcribe texts saved, on average, 2.34 (95%CI: 1.61, 3.07) minutes per 100 words as compared to manual transcription ($p < 0.001$). Moreover, whereas manual transcription was found to have a significant negative correlation with proficiency scores ($r = -0.59$, $p < 0.05$), no such significant correlation was found for the Sonix method ($r = -0.2$, $p = 0.195$), suggesting that this method may be cheaper than manual transcription even for low-proficiency learners and even when considering the time investment required to manually correct the transcripts.

References

Chan, M. P. Y., Choe, J., Li, A., Chen, Y., Gao, X., & Holliday, N. (2022, September). Training and typological bias in ASR performance for world Englishes. Interspeech.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The microsoft 2017 conversational speech recognition system. 5934–5938. http://arxiv.org/abs/1708.06073

Using crowdsourced comparative judgement and rubric-based rating to grade texts in the ICLE corpus: a report on reliability and validity

Vandeweerd, Nathan (Radboud University), Peter Thwaites (UCLouvain), Magali Paquot (UCLouvain) and Jiacheng Shen (Ghent University)

Comparative judgement (CJ), an assessment method in which judges are shown pairs of texts side-by-side and asked to choose which is "better", has recently been introduced as a new method to generate reliable and valid proficiency scores for texts in learner corpora (Paquot et al., 2022). Recent (small-scale) studies have shown this approach to be effective for evaluating argumentative essays of varying lengths, even when texts cover a narrow proficiency span (e.g. CEFR B2-C1) or diverse essay prompts (Authors, submitted). They have also found that CJ assessments made by judges recruited through a crowdsourcing platform have similar validity and reliability to those made by linguists recruited through a community-driven approach (Authors, resubmitted).

This presentation reports on an ongoing large-scale study investigating the extent to which rubric-based judges and CJ raters focus on the same linguistic features when assessing texts. A CJ task was created in which users of the Prolific crowdsourcing platform assessed a representative sample of 1300 texts from the ICLE corpus. In tandem, 500 of these texts were triple rubric-rated by a team of experienced graders. Text-based measures representing the main rubric constructs (e.g., lexical complexity, accuracy) were then calculated on a subset of these texts (N=222) using a combination of automated tools and manual annotation to determine the extent to which these measures were predictive of rubric and rank scores.

The results of the study will benefit the learner corpus research community in three ways. Firstly, they will provide the first evidence of CJ's potential for generating assessment of L2 writing samples at the scale required in LCR. Second, insights into the textual features predictive of CJ scores will provide much-needed evidence of the method's validity. Lastly, both the rubric-based CEFR scores and the ranking scale generated in the study will be made publicly available, allowing researchers to benefit from a large, reliably graded sub-section of the ICLE corpus for the first time.

References

Paquot, M., Rubin, R., & Vandeweerd, N. (2022). Crowdsourced Adaptive Comparative Judgment: A community-based solution for proficiency rating. Language Learning, 72(3), 853–885.

Writing argumentative essays: Discourse strategies in L1- and L2-authored versus ChatGPT-generated text

Wan, Shujun (Shanghai International Studies University), Julián Moreno-Schneider (German Research Center for Artificial Intelligence) and Georg Rehm (German Research Center for Artificial Intelligence)

The arrival of ChatGPT has sparked extensive user engagement and has become a focal point of interest in academic research circles. This study explores the discourse features of German argumentative essays generated by ChatGPT and compares them with those written by native speakers (L1) and Chinese learners of German (L2). The decision to separate L1 and L2 for comparison with ChatGPT texts, rather than grouping them together as human texts, is twofold. Firstly, our preliminary research reveals notable differences in rhetorical strategies between Chinese L2-Texts and German L1-Texts. Secondly, such a comparison holds undeniable importance, especially in discussions about potential applications in language learning and teaching. Our key questions are:

1. How is the discourse structure between different text regions shaped?
2. What rhetorical strategies are used in constructing the texts?
3. Which discourse connectives are used to link different arguments together?

To address the questions, we built a corpus consisting of 90 German argumentative texts, all of which are on the same topic. To describe the rhetorical construction of the texts, we employed rhetorical structure theory (RST) (Mann & Thompson, 1988). The annotation work was carried out independently by two linguistic experts, and the inter-annotator agreement is $\kappa = 0.85$.

The findings reveal that: a) GPT-Texts show lower diversity in discourse structure compared to L1- and L2-Texts, b) GPT-Texts share more similarities in rhetorical strategies with Chinese L2-Texts than with L1-Texts, c) GPT-Texts show fewer variations in connectives for linking arguments compared to L2-Texts, and both GPT-Texts and L2-Texts exhibit a lower diversity than L1-Texts. Based on the findings, we conclude by presenting the insights gained from this research in using ChatGPT for language learning and teaching purposes.

References

Mann, W. C. and S. A. Thompson (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization." Text 8(3), 243–281.

The influence of L1 Dutch on cohesion in L2 German academic writing: A contrastive corpus-based analysis

Wedig, Helena; Carola Strobl, Jim Ureel, Tanja Mortelmans (University of Antwerp)

Novice L2 writers tend to rely on L1 strategies to create cohesive texts, which may differ from strategies used in the L2 (Roberts et al., 2008). One of these strategies is coreference (Halliday & Hasan, 1976). Research on coreference in L2 German has been scarce to date in stark contrast to L2 English (e.g., Grüter et al., 2017; He, 2020). The few existing studies on L2 German focus on specific coreference types, such as possessives (e.g., Fabricius-Hansen et al., 2021) or pronominal adverbs (Belz, 2005; Strobl, 2019) and on heterogeneous learner groups. To date, there is no comprehensive study on L2 German coreference use available.

Our study aims to fill this gap by comparing coreference use of L2 writers with L1 Dutch, L2 writers with heterogenous L1s other than Dutch and L1 German writers. The analysis is based on the Belgisches Deutschkorpus (Beldeko) and the two subcorpora of the German summary corpus (GerSumCo L1 & L2). Coreference was manually annotated with the help of a newly developed annotation system that combines categories of different frameworks (e.g., Becher, 2011; Kunz, 2010; Reznicek, 2013), such as antecedent types, coreferential expression, degree of coreference explicitness, and coreferential relations.

The analyses via R revealed differences in coreference use, which distinguish L2 German writers with heterogeneous L1s from those with L1 Dutch and L1 German: They use fewer proper nouns (23% vs. 31% and 34%) and more pronouns (40% vs. 34% and 27%) than the two other groups. They also use less repetitions (21% vs. 28% and 33%) and more personal pronouns (40% vs. 34% and 27%) for coreference. Additionally, we find more intra-sentential relations (29%) when compared to the L1 group (24%) and the L2 group with L1 Dutch (23%). In the presentation, we will discuss these results in light of L1 Dutch influence.

References

Becher, V. (2011). Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts. Unpublished Doctoral dissertation, Universität Hamburg.

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

Fabricius-Hansen, C., Pitz, A. P., & Torgersen, H. A. T. (2021). Lexical interference in non-native resolution of possessives? Oslo Studies in Language, 12(2), 25–63. https://doi.org/10.5617/osla.8955

Grüter, T., Rohde, H., & Schafer, A. J. (2017). Coreference and discourse coherence in L2: The roles of grammatical aspect and referential form. Linguistic Approaches to Bilingualism, 7(2), 199–229. https://doi.org/https://doi.org/10.1075/lab.15011.gru

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. Longman.

He, Z. (2020). Cohesion in Academic Writing: A Comparison of Essays in English Written by L1 and L2 University Students. Theory and Practice in Language Studies, 10(7), 761–770. http://dx.doi.org/10.17507/tpls.1007.06

Kunz, K. (2010). Variation in English and German nominal coreference: a study of political essays. Peter Lang.

Reznicek, M (2013). Linguistische Annotation von Nichtstandardvarietäten — Guidelines und „Best Practices": Guidelines Koreferenz: Version 1.01. F-AG 7: Angewandte Sprachwissenschaft, Computerlinguistik Kurationsprojekt 2. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-cor-1.1

Roberts, L., Gullberg, M., & Indefrey, P. (2008). Online pronoun resolution in L2 discourse - L1 influence and general learner effects. Studies in Second Language Acquisition, 30, 333–357. https://doi.org/10.1017/S0272263108080480

Strobl, C. (2019). Darum sind Pronominaladverbien eine Herausforderung für Deutschlerner. Germanistische Mitteilungen, 45(1&2). https://doi.org/10.33675/GM/2019/1&2/11

Frequency vs. accuracy in learner Englishes: A study on tense and aspect

Werner, Valentin (University of Bamberg) and Robert Fuchs (University of Bonn)

This study expands on previous work in the area of morphosyntactic tense-aspect (TA) expression (see, e.g., Deshors, 2021; Fuchs & Werner, 2020) with a view to testing and refining established SLA principles on the acquisition of TA markers. Specifically, we consider (i) the order of acquisition of tense and aspect (OATA) and (ii) the Default Past Tense Hypothesis (DPTH; Salaberry, 2008). To date, these hypotheses have been put to the test only in smaller learner groups, mainly applying experimental SLA approaches (see, e.g., Bardovi-Harlig, 2000; Svalberg, 2018). In this study, we test the predictions of the OATA and DPTH on data from learners of English as a Foreign Language at school and university level. The central issue in focus is to what extent an increase in the frequency of usage corresponds to an increase in accuracy.

Accordingly, we use a quasi-longitudinal research design and measure both the frequency and the accuracy of usage of TA markers, using multi-layer error annotations to explore whether and to what extent an increase in the frequency of usage corresponds to an increase in accuracy. Data is drawn from the International Corpus of Crosslinguistic Interlanguage (Tono & Díez-Bedmar, 2014) and the International Corpus of Learner English (Granger et al., 2009) to assess TA acquisition in (tutored) learner writing from the beginning to the advanced level in four typologically different L1 backgrounds (German, Chinese, Polish, Spanish). Based on the categories established in Dagneaux et al. (2005), error ratings of more than 4,000 data points (verb tokens) were provided by two native speakers, with disagreements between these raters being resolved by a third native-speaker rater.

Overall, our results confirm the predictions of the OATA and the DPTH. Findings indicate that simple forms are used (i) earlier and more frequently and (ii) more accurately than complex forms at any stage in the acquisition process. However, the data also are also suggestive of nuanced patterns, indicating that accuracy of usage does not linearly increase with frequency of usage or proficiency. In addition, the manual accuracy ratings allow us to assess (i) accuracy of usage in terms of "false negatives" (e.g. using a present simple where a present progressive is required) and (ii) particular error types (functional errors, i.e. confusion of TA forms, and formal errors, e.g. omission of 3rd person singular -s in the present). The analysis also yields various types of interaction across the learner samples with different L1s and individual TA forms studied.

References

Bardovi-Harlig, K. (2000). Tense and aspect in second language acquisition: Form, meaning, and use. Malden: Blackwell.

Dagneaux, E., Dennes, S., Granger, S., Meunier, F., Neff, J., & Thewissen, J. (2005). Error tagging manual version 1.2. Louvain: Centre for English Corpus Linguistics.

Deshors, S. C. (2021). Contextualizing past tenses in L2: Combined effects and interactions in the present perfect versus simple past alternation. Applied Linguistics, 42(2), 269–291.

Fuchs, R., & Werner, V. (2020). Tense and aspect in second language acquisition and learner corpus research. Amsterdam: Benjamins.

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

Salaberry, R. (2008). Marking past tense in second language acquisition: A theoretical model. London: Continuum.

Svalberg, A. M.-L. (2018). Mapping tense form and meaning for L2 learning – from theory to practice. International Review of Applied Linguistics in Language Teaching, 57(4), 417–445.

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.

The Corpus of Young German Learner English

Werner, Valentin (University of Bamberg), Robert Fuchs (University of Bonn), Anna Rosen (University of Freiburg), Lea Bracke (University of Bamberg) and Bethany Stoddard (University of Bonn).

The full potential of Learner Corpus Research (LCR) for the analysis of interlanguage (Selinker 1972, 1992) has yet to be realized due to several major challenges. These include (i) a neglect of task effects, as well as underrepresentation or lack of: (ii) beginner and lower-intermediate learners, (iii) spoken material and bi-modal data, (iv) metadata and (v) longitudinal or quasi-longitudinal perspectives (e.g. Myles 2021; Tracy-Ventura et al. 2021).

The project presented will address these limitations by compiling and analyzing a corpus of Young German Learner English (YGLE). The project aims to complement the extensive body of work on highly advanced, university-level L1 German EFL learners (e.g. Fuchs et al. 2016; Römer et al. 2020) by creating a database on the production of beginning to intermediate L1 German EFL learners (grades 5–11) in institutional contexts.

The tasks administered to a target sample of approximately 700 students include an array of traditional and innovative task types with varying degrees of planning and interactivity, such as an argumentative essay, a group discussion, and a simulated text chat. An extensive set of metadata is collected, in accordance with procedures proposed by Möller (2017) and Frey et al. (2023). This comprises established test batteries assessing socioeconomic and linguistic background, language use in social contexts, motivation (standardized tests FLM 3–6 R, FLM 7–13; Lohbeck & Petermann 2019; Petermann & Winkel 2015), as well as cognitive abilities (standardized test AID-G; Kubinger & Hagenmüller 2019).

After data transcription and annotation, a first focus will lie on the triad complexity-accuracy-fluency (CAF) and the influence of contextual and learner variables will be assessed using mixed-effects regression modeling. YGLE will eventually be made available to the LCR community, allowing (i) the exploration of areas beyond CAF and potentially (ii) comparison with data from beginner and intermediate learners of English worldwide.

References

Frey, Jennifer-Carmen, Alexander König, Egon Stemle and Magali Paquot, "A core metadata schema for L2 data" (Paper presented at EuroSLA 32: Conference of the European Second Language Association. Birmingham, UK, August 30-September 2, 2023).

Fuchs, Robert, Sandra Götz and Valentin Werner. 2016. The present perfect in learner Englishes: A corpus-based case study on L1 German intermediate and advanced speech and writing. In Re-assessing the Present Perfect, edited by V. Werner, E. Seoane and C. Suárez-Gómez, 297–338. Berlin: Mouton de Gruyter.

Kubinger, Klaus and Bettina Hagenmüller. 2019. Gruppentest zur Erfassung der Intelligenz auf Basis des AID. Göttingen: Hogrefe.

Lohbeck, Annette and Franz Petermann. 2019. Fragebogen zur Leistungsmotivation für Schülerinnen und Schüler der 3. bis 6. Klasse – Revision. Göttingen: Hogrefe.

Möller, Verena. 2017. Language Acquisition in CLIL and Non-CLIL Settings: Learner Corpus and Experimental Evidence on Passive Constructions. Amsterdam: Benjamins.

Myles, Florence. 2021. Commentary: An SLA perspective on Learner Corpus Research. In Learner Corpus Research Meets Second Language Acquisition, edited by B. Le Bruyn and M. Paquot, 258–273. Cambridge: Cambridge University Press.

Petermann, Franz and Sandra Achtergarde. 2015. Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13. Klasse. Frankfurt: Pearson Harcourt.

Römer, Ute, Stephen C. Salicky and Nick C. Ellis. 2020. Verb-argument constructions in advanced L2 English learner production: Insights from corpora and verbal fluency tasks. Corpus Linguistics and Linguistic Theory, 16 (2): 303–331.

Selinker, Larry. 1972. Interlanguage. International Review of Applied Linguistics in Language Teaching, 10 (1/4): 209–232.

Selinker, Larry. 1992. Rediscovering Interlanguage. London: Longman.

Tracy-Ventura, Nicole, Magali Paquot and Florence Myles. 2021. The future of corpora in SLA. In The Routledge Handbook of Second Language Acquisition and Corpora

L1 Influence on Chinese English Learners’ Use of Individual Senses of “IN”

Xu, LingLing (University of Birmingham)

This study examines how a learner's first language (L1) impacts the accuracy of Chinese English learners’ use of the preposition "IN" which encompasses over 10 senses (i Ferrando, 2000; Evans & Tyler, 2004), with substantial semantic differences from its Chinese counterpart, creating great challenges for L1 Chinese English learners. By investigating both oft-studied spatial senses and under-studied metaphorical senses, I aim to discern the L1 influence on different senses of “IN”, contributing insights to the prototypicality hypothesis (Kellerman, 1979; Tanaka, 1983, Davy, 2000) which postulates that prototypical senses of English spatial prepositions are less susceptible to L1 influence than non-prototypical ones.

Error-annotated data from the EF-Cambridge Open Language Database (EFCAMDAT: Geertzen et al., 2013) were utilized. 400 obligatory contexts of “IN” (200 accurate cases and 200 inaccurate ones) were identified in the Chinese learner group, and another 400 were identified in the aggregated group of non-Chinese learners. It is worth noting that including learners of different L1s is imperative as L1 influence can be identified only when learners of different L1s differ in their use of the same L2 feature. Each case was coded in terms of “accuracy”, “sense” (e.g., “Temporal” sense: *“in May*”), “sense type” (i.e., spatial vs metaphorical) and whether the Chinese equivalent is used in the corresponding translated context. In addition, the same data was coded by another PhD student for inter-rater reliability.

A mixed-effects model indicated that Chinese English learners’ accuracy was higher when the Chinese equivalent is also used in the corresponding translated context than when it is not.

Furthermore, the results revealed no significant difference in the strength of L1 influence between senses of “IN”. This finding indicates that L1 influence may not vary across different senses of “IN”.

References

Davy, B. L. (2000). A cognitive-semantic approach to the acquisition of English prepositions. University of Oregon.

Evans, V., & Tyler, A. (2004). Spatial experience, lexical structure and motivation: The case of in. In G. Radden & K. U. Panther (Eds.), Studies in Linguistic Motivation (pp. 157-192). Walter de Gruyter.

i Ferrando, I. N. (2000). A cognitive-semantic analysis of the English lexical unit" in". Cuadernos de investigación filológica, 26, 189-220.

Kellerman, E. (1979). Transfer and non-transfer: Where we are now. Studies in second language acquisition, 2(1), 37-57.

Tanaka, S. (1983). LANGUAGE TRANSFER AS A CONSTRAINT ON LEXICO-SEMANTIC DEVELOPMENT IN ADULTS LEARNING A SECOND LANGUAGE IN ACQUISITION-POOR ENVIRONMENTS (LOCATIVES; JAPANESE, ENGLISH). Teachers College, Columbia University.

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 (pp. 240-254).

Wordless: An integrated corpus tool with multilingual support for the study of language acquisition, pedagogy, and assessment

Ye, Lei (Shanghai International Studies University)

This paper presents Wordless (Ye, 2023), an integrated corpus tool with multilingual support for the study of language, literature, and translation. Wordless is free, cross-platform, and open-source. It aims to eliminate the barriers to the utilization of bleeding-edge technologies by language researchers and provide a better alternative to other corpus tools currently available such as WordSmith, AntConc, Sketch Engine, and ParaConc.

Wordless features a user-friendly graphical interface, catering specifically to the needs of non-technical users. It adopts a modular design and each of the 9 main modules provides functionalities for corpus profiling, concordancing, parallel concordancing, dependency parsing, wordlist generation, n-gram generation, collocation extraction, colligation extraction, and keyword extraction respectively.

Wordless is "batteries-included": it has built-in multilingual support for multiple natural language processing tasks including sentence/word/syllable tokenization, part-of-speech tagging, lemmatization, stop word filtering, dependency parsing, sentiment analysis, and more. It can process and analyze corpora in more than 120 languages across the globe, including many non-Indo-European languages as well as those spoken in some Southeast Asia countries which require special tokenization handling such as Burma, Chinese, Khmer, Lao, Japanese, Thai, Tibetan, and Vietnamese.

Wordless is capable of calculating a wide range of statistical measures, including indicators of readability and lexical diversity of the whole text, measures of dispersion and adjusted frequency indicating the distributive evenness of tokens/n-grams in the corpus, and methods of Bayes factor and effect size used to complement tests of statistical significance in collocation/colligation/keyword extraction. Wordless also comes with a variety of data visualization options including dispersion plots, line charts, word clouds, network graphs, and dependency graphs. Other convenience features include zapping options for quick creation of cloze tests to be used in classroom settings and functionalities for parallel concordancing.

References

Ye, L. (2023). Wordless (Version 3.4.0) [Computer software]. Github. https://github.com/BLKSerene/Wordless

How does the first language influence the shape of text? A Comparison of Korean and Polish non-native Czech texts

Zasina. Adrian Jan (Charles University)

The influence of the first language on a second/foreign language has been confirmed by numerous studies (Figueredo, 2006), especially in the case of closely related languages where the degree of negative interference is high (Romaševská, 2018). However, up to the time of writing, there has been no extensive research conducted on the Czech language that considers the more complex characteristics involved in text creation. An opportunity to analyse a text based on its complexity in terms of using language features, which are used in specific communication situations, comes with multidimensional analysis (Biber, 1988). Multidimensional analysis captures the functional variability in a text that is related to a concrete situation. Considering many language features, it is possible to compare texts with an existing model of language. Therefore, the Czech language model (Cvrček et al., 2018) may be compared with any other dataset to reveal similarities and dissimilarities.

This study attempts, for the very first time, to compare two datasets of learners' texts in the background of the multidimensional model of Czech. The analysis considers learner texts from Polish speakers at level A2 (62 texts from 16 students) and texts from Korean speakers at level A2 (32 texts from 8 students). All students wrote the text based on four tasks with strictly given instructions, encompassing an informal letter, a description of a place, an argumentative essay, and storytelling. Subsequently, both sets were projected onto the multidimensional space to provide a comparison.

The preliminary results reveal that Korean texts are more dynamic and less cohesive than Polish texts, which may be caused by learners’ language competence. In terms of language features, however, it seems that Korean informal texts use fewer demonstrative pronouns and polyfunctional pronoun “to” ‘it’, and Korean formal texts are poorer in verbal predicates complemented by a clause in comparison with Polish texts. These remarks indicate the areas that should be considered in teaching materials focusing on the different needs of Korean and Polish students.

References:

Biber, D. (1988). Variation Across Speech and Writing. Cambridge University Press.

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018). Variabilita češtiny: Multidimenzionální analýza. Slovo a slovesnost, 79(4), 293–321.

Figueredo, L. (2006). Using the known to chart the unknown: A review of first language influence on the development of English-as a-second-language spelling skill. Reading and Writing, 19, 873–905. https://doi.org/10.1007/s11145-006-9014-1

Romaševská, K. (2018). Specifika osvojování českých reflexivních sloves polskými a ruskojazyčnými mluvčími [Characteristics of Acquisition of Czech Reflexive Verbs by Speakers of Polish and Russian]. Slavia, 87(4), 409–428.