Analyzing learner language in task contexts: A study case of linguistic complexity and accuracy in EFCAMDAT

Akira Murakami[1], **Marije Michel**[2], **Theodora Alexopoulou**[1], **Detmar Meurers**[3]

[1]Cambridge University, United Kingdom
[2]Lancaster University, United Kingdom
[3]Tuebingen University, Germany

Large learner corpora from online foreign language tasks, e.g., EF Cambridge Open Language Database (EFCAMDAT, Alexopoulou et al. 2015) provide opportunities to analyze second language (L2) learner data at an unprecedented scale. EFCAMDAT covers all CEFR levels with 33 million words based on 128 writing tasks by 85,000 learners with varying L1s.

Interpreting learner language in such corpora necessitates understanding of task effects: How do the visual and textual task prompt as well as instructional focus influence the L2 writings that build the corpus? This question is vital for modelling the developmental L2 trajectory in large corpora. A key challenge is the considerable variation in responses, for example due to natural variation in the linguistic means learners use to meet task requirements.

We present a study case of three task types (narrative, descriptive, professional) using a subcorpus of EFCAMDAT (4.3 million words, 51,439 writings by 22,954 learners). First, tasks were characterised using the TBLT framework (e.g., regarding cognitive task complexity, Robinson, 1995, and code complexity, Skehan, 2001) and predictions about expected language elicited by instruction and task prompt were formulated. Second, linguistic complexity and accuracy of learner writings was analysed using computational linguistic methods, exploring a range of 45 lexical, syntactic, and discourse measures. Third, randomly selected individual task responses were inspected qualitatively.

Findings show that linguistic complexity can characterise the language elicited by different task types in line with our predictions. We were able to confirm that code complexity supported higher syntactic complexity (e.g., dependent clauses). By contrast, accuracy shows a weaker relationship with tasks. Importantly, there is an emerging classification of complexity measures, those aligning with proficiency and those that are more task-dependent.

The work demonstrates the fruitfulness of combining computational linguistic analysis of large corpora with the TLBT perspective for SLA research.

Keywords: learner corpora, computational linguistics, L2 writing, TBLT, CAF.