Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language Processing techniques

Theodora Alexopoulou
University of Cambridge

Marije Michel
Lancaster University

Akira Murakami
University of Cambridge

Detmar Meurers
University of Tübingen

To appear 2017 in *Language Learning*, Special Issue on Language learning research
at the intersection of experimental, corpus-based and computational methods

Abstract

Large-scale learner corpora collected from online language learning platforms, such as EF-Cambridge Open Language Database (EFCAMDAT), provide opportunities to analyze learner data at an unprecedented scale. However, interpreting the learner language in such corpora requires a precise understanding of tasks: How does the prompt and input of a task and its functional requirements influence task-based linguistic performance? This question is vital for making large-scale task-based corpora fruitful for SLA research. We explore the issue through an analysis of selected tasks in EFCAMDAT and the complexity and accuracy of the language they elicit.

1 Introduction and Motivation 2

2 Combining TBLT with Learner Corpus Research 3

3 Research Question 4

4 Data and Methods 4
4.1 The EFCAMDAT Corpus . . . 4
4.2 Methods . . . 5
4.2.1 Analyzing accuracy . . . 5
4.2.2 Analyzing complexity . . . 6

5 Development of Complexity and Accuracy Across Proficiency 6
5.1 Results of Complexity Analysis . . . 7
5.2 Results of Accuracy Analysis . . . 7

6 Effect of Task Design Features and Instructional Context 10
6.1 Three Pairs of Tasks: Narrative, Descriptive and Professional Tasks . . . 10
6.2 Predictions based on task design features . . . 12
6.3 Predictions based on task type and instructional focus . . . 13
6.4 Results . . . 13
6.4.1 Global Measures of Complexity and Accuracy per Task . . . 13
6.4.2 Effects of task design features on linguistic complexity and accuracy . . . 14
6.5 Effects of Task Type and Instruction . . . 15

7 General Discussion and Conclusions 17

1 Introduction and Motivation

Learner corpus research has primarily relied on collecting and analysing second language (L2) learner writings such as essays (Granger, 2008). The increasing use of online language learning platforms creates opportunities for collecting L2 data at an unprecedented scale, covering a wide range of contexts e.g., people all over the world working on tasks on their computer or tablet at home or elsewhere. The EF-Cambridge Open Language Database (EFCAMDAT; Geertzen et al., 2014) as the data underlying the research presented in this article gives access to 1.2 million learner writings at all proficiency levels of English, and the corpus continues to grow. Learner corpora can, therefore, provide Second Language Acquisition (SLA) research with a potential empirical treasure trove. However, learner corpora like EFCAMDAT that are collected in an educational context also raise the fundamental question of how to identify and interpret the data given the many interacting linguistic, instructional and learner factors.

In this paper we argue that a fruitful step forward is the collaboration between corpus linguists, SLA researchers, and computational linguists. We will exemplify this synergy through investigating task effects on learner language by applying Natural Language Processing and corpus-linguistic tools to test hypotheses formulated within a task-based approach to language teaching (TBLT). Evaluating hypotheses in learner corpora introduces challenges absent from typical experimental paradigms used within TBLT research. Unlike experiments, corpora typically do not follow a research design dedicated to the evaluation of a specific hypothesis, since, by their very conception, corpora are resources to be used for a wide range of research questions. The interpretation, therefore, needs to tease apart the different linguistic, instructional and learner factors that interact in shaping the corpus data.

In any learner corpus, the prompts and topics used to elicit the L2 samples shape the language that is represented in the corpus. Indeed, there is a recognition in SLA literature that corpus-based developmental investigations need to control for task effects to ensure that the developmental trajectory is not skewed. Tracy-Ventura and Myles (2015) show how task effects can affect generalisations regarding the acquisition of imperfective forms if tasks do not provide sufficient opportunity for use of a variety of forms. Vyatkina (2012, p. 595) warns that task effects pose a “particularly severe threat to validity in longitudinal designs”.

While task effects are widely recognised as an important aspect of learner language analysis – for L1 writing, see Huot (1990), for language assessment: Bachman (1990); Biber and Conrad (2009); Biber et al. (2014); Hinkel (2009); Weigle (2002), for instructed SLA: Kormos (2011); Kuiken and Vedder (2008); Way et al. (2000), and for CALL: Quixal and Meurers (2016) – learner corpus research has generally not been linked to research investigating effects of task on writing (but cf. Ott et al., 2012; Gablasova et al., 2015).

The first aim of this paper is to provide a conceptual and methodological example of how to connect the analysis of task effects in learner corpora with insights from Task-based Language Teaching (TBLT). We adopt task-based frameworks (e.g., the Cognition Hypothesis by Robinson, 1995) to characterize tasks used in EFCAMDAT so as to separate task effects from developmental trajectories.

Our second goal is to illustrate the relevance of large-scale learner corpora as an empirical test-bed for complementing the experimental TBLT research, so as to empirically broaden and strengthen findings and situate results within a proficiency trajectory of data from large numbers of learners from around the globe with different language backgrounds working on a large number of tasks.

To achieve these goals we first examine how cognitive task complexity might affect the global complexity and accuracy of the elicited language in line with earlier work on Complexity, Accuracy, and Fluency (CAF; cf. Housen and Kuiken, 2009; Michel, 2017, we here leave aside fluency, given our focus on writing). We then investigate individual language features that may be elicited by the task and the instructional focus, such as the vocabulary and grammatical features given in the task prompt or the focus of the teaching unit leading up to the task. This perspective relates to Loschky and Bley-Vroman's (1993) distinction between 'natural', 'useful' and 'essential' structures. It also means we draw on learner corpora that allow to reliably evaluate L2 knowledge as opposed to L2 use (Tracy-Ventura and Myles, 2015).

Last, but not least, we aim to demonstrate the necessity of combining corpus techniques with computational linguistics in order to enable the linguistic analysis (complexity, accuracy, lexical and grammatical features) that is needed for modelling task effects and developmental trajectories in large corpora. Natural Language Processing (NLP) tools are needed to automatically extract the relevant linguistic features, structures and patterns (Granger et al., 2007; Meurers, 2012, 2015). While the NLP analysis of learner language presents challenges waiting to be addressed, requiring more interdisciplinary collaboration (Meurers and Dickinson, 2017), state-of-the art NLP for native language can provide a first approximation of the potential usefulness in terms of supporting the effective identification of relevant subsets of data and extending the standard inventory of complexity measures currently used in CAF analyses.

In sum, this paper will argue for and exemplify the fruitful triangulation of corpus linguistic, computational linguistic, and task-based approaches to SLA research.

2 Combining TBLT with Learner Corpus Research

Among the many definitions of a task within TBLT (e.g., Ellis, 2003; Skehan, 1998), we here follow Samuda and Bygate (2008) who define a task as "a holistic activity, which engages language use in order to achieve some non-linguistic outcome while meeting a linguistic challenge, with the overall aim of promoting language learning, through process or product or both" (p.69). This definition is well suited for the EFCAMDAT corpus as a collection of instructed writing activities.

One productive strand of task-based research is interested in how different design features such as task complexity affects linguistic performance. A central hypothesis of the Limited Attentional Capacity Model (Skehan, 1998) and the Cognition Hypothesis (Robinson, 1995) is that the cognitive complexity of a task will impact on the complexity and accuracy of the language people use to meet the cognitive/communicative requirements of the task. Skehan distinguishes factors of code complexity (e.g., vocabulary load and variety), cognitive complexity (e.g., clarity and structure of information to process) and communicative stress (e.g., time pressure) of a task that will affect the linguistic complexity and accuracy of the elicited language. Specifically, he predicts that limitations in attentional resources will lead to competition between complexity and accuracy. Unlike Skehan, Robinson's Cognition Hypothesis assumes that learners can access multiple attentional pools. His triadic framework (Robinson and Gilabert, 2007) distinguishes task complexity features (e.g., number and similarity of elements to deal with, whether a task plays in the here-and-now or not, whether it involves reasoning and perspective taking) from task condition features (e.g., gender of conversational partners) and task difficulty (e.g., perceived task demands, related to individual differences such as aptitude). In contrast to Skehan, Robinson claims that some aspects of task complexity (e.g., higher reasoning demands) will promote both high linguistic complexity and accuracy because the higher cognitive load will trigger learners to activate and allocate attentional resources to the linguistic form of task performance. Despite a rich body of empirical research (Jackson and Suethanapornkul, 2013; Robinson, 2011), to date no overarching outcomes can be presented because the different studies have used a plethora of task complexity manipulations and an even larger number of complexity and accuracy measures hindering comparisons (Long, 2016; Jackson and Suethanapornkul, 2013). On the one hand, the meta-analysis based on nine comparable studies by Jackson and Suethanapornkul (2013) attests small increases in accuracy and a decrease in fluency, which is in line with the Cognition Hypothesis. On the other hand, complexity measures did neither lend support nor disconfirm Robinson's (1995) account.

In addition to task complexity, task type (e.g., narrative, argumentative) has been shown to impact in particular on linguistic complexity (Foster and Skehan, 1996; Vyatkina, 2012; Yoon and Polio, 2016; Bouwer et al., 2015; Lu, 2011). Yoon and Polio (2016) suggest that the functional differences between task types are stronger than differences in cognitive complexity.

The role of focused instruction has not been addressed systematically in learner corpus research. Yet, Vyatkina (2012) demonstrates that when a unit leading up to a task targets a specific structure, it is likely that a task prompt at the end of this unit will elicit that target structure when in another instructional setting, the same prompt might elicit a different structure.

Our goal is to demonstrate that corpus-based developmental SLA research can significantly benefit from insights from TBLT providing a deeper understanding of how different task design features affect the language elicited. At the same time, corpora can complement experimental paradigms in TBLT. Experiments allow researchers to manipulate fine aspects of task design while controlling for others in order to test their hypotheses. However, they are constrained by limited data size. Conversely, corpus investigations are constrained by the fact that the task design does not follow the specific manipulations that would be needed for testing specific hypotheses since most learner corpus developers aim for resources that can be of more general use. Yet, they offer a much richer data source in terms of data size as well as diversity (proficiency, L1 backgrounds, etc.) and, thus, allow us to investigate how core design features of a task (e.g., task type) impact on the language used by learners (Plonsky and Kim, 2016). Importantly, corpora allow us to situate task effects within the proficiency trajectory and, therefore, better understand their impact on second language development (Jackson and Suethanapornkul, 2013; Norris and Ortega, 2009).

In this paper, we will test the specific predictions made based on task-based insights of the EFCAMDAT data for selected tasks.

3 Research Question

Our goal is to investigate the impact of task design on elicited language. In particular:

RQ How do task design features and instructional focus affect the written language used by second language learners when they try to meet the non-linguistic goal of a task? Specifically, how does task complexity, task type, and/or instructional focus impact on the complexity and accuracy of the language use in global as well as specific features or structures?

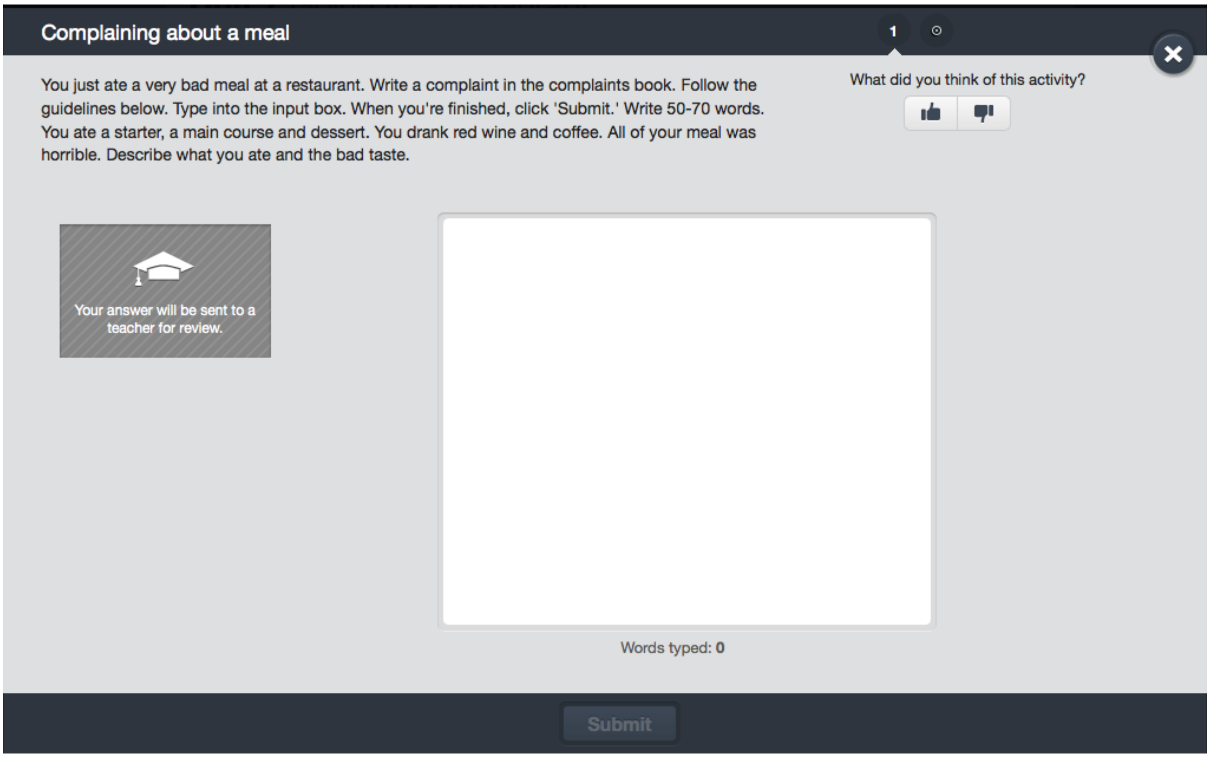
We embed our main research question in the more general question of how learner language develops in a longitudinal corpus, grounding the task-based analysis of linguistic complexity and accuracy on a longitudinal analysis of linguistic complexity across proficiency levels in EFCAMDAT. We finally ask if the observed task effects can be accounted for by the Limited Attentional Capacity Model (Skehan, 1998) and/or the Cognition Hypothesis (Robinson, 1995).

4 Data and Methods

4.1 The EFCAMDAT Corpus

We drew on EFCAMDAT, an open access corpus available at `http://corpus.mml.cam.ac.uk/efcamdat`. The corpus consists of writings submitted to *Englishtown*, the online school of EF Education First. The Englishtown curriculum contains 16 levels, from A1 to C2 of the Common European

Framework of Reference (CEFR). Each level consists of 8 units, each including a free writing task such as the one illustrated in Figure 1, summing to 128 distinct tasks.



Complaining about a meal

You just ate a very bad meal at a restaurant. Write a complaint in the complaints book. Follow the guidelines below. Type into the input box. When you're finished, click 'Submit.' Write 50-70 words. You ate a starter, a main course and dessert. You drank red wine and coffee. All of your meal was horrible. Describe what you ate and the bad taste.

Your answer will be sent to a teacher for review.

Figure 1: Englishtown Task 6.7: Complaining about a meal

The pre-release second version we used for this study contains 1,180,543 individual scripts. Script length ranges from 20-40 words at lower levels to 150-180 words at higher levels. There are 174,771 learners; 98,686 (56.5%) learners have written at least three scripts. Two thirds of scripts (787,010; 66.7%) include teacher corrections using a list of 24 error labels. National language background is used as an approximation of L1 (Geertzen et al., 2014). For further details of EFCAMDAT and Englishtown, readers are referred to an online supporting document.

4.2 Methods

4.2.1 Analyzing accuracy

We exploit the 24 error labels in EFCAMDAT provided by EF teachers (cf. Figure 6) to identify learner errors. Murakami (2014) used error labels in EFCAMDAT and successfully replicated findings on morpheme accuracy from the Cambridge Learner Corpus, which is manually error-tagged using the sophisticated error annotation system of Nicholls (2003).

Among the many indices of accuracy (e.g., Foster and Wigglesworth, 2016; Polio, 1997; Polio and Shea, 2014), we calculate relative error frequency which has been shown to successfully discriminate learners of different proficiency levels (Hawkins and Filipović, 2012). Since error frequency is different from accuracy (Schachter and Celce-Murcia, 1977), we further calculated accuracy for two features, prepositions and past tense verbs. As shown in the next section, the use of prepositions increases across proficiency while the use of past tense verb forms decreases. These two features then allow us to investigate possible interactions between accuracy and use. To measure accuracy we calculated target like use in obligatory contexts (Pica, 1983; cf. Crosthwaite, 2016; Murakami and Alexopoulou, 2016).

First, we derived corrected texts in which incorrect portions were replaced with the corresponding corrected forms based on teacher corrections. We then annotated the original and corrected learner writings with part-of-speech tags using TreeTagger (Schmid, 1994) with the provided English model. Next, we targeted “IN” and “VBD” in the corrected scripts for prepositions and past tense forms respectively. Finally, correct suppliances were counted by subtracting the number of omission and misformation errors from that of obligatory contexts. Capitalization errors and spelling errors were excluded.

4.2.2 Analyzing complexity

While recent work suggests that more finegrained measures might be more appropriate (Inoue, 2016; Lambert and Kormos, 2014), as a first step for establishing a common set of measures to capture global syntactic complexity, we adopted the suggestions in Norris and Ortega (2009) and used the following subconstructs:

1. An overall length-based metric: average sentence length (in words)
2. Sub-clausal complexity: mean length of clause (in words)
3. Subordination: subordinate clause to all clause ratio

Average sentence length targets potentially multi-clausal complexity of different types (e.g., through subordination, coordination, modification), whereas mean clausal length targets sub-clausal complexity at the phrasal level. We do not report a measure of sentential coordination (suggested by Norris and Ortega, 2009) since automatic identification of such coordinations in EFCAMDAT was not reliable enough due to the multi-faceted nature of coordination and its interplay with ellipsis.

To measure global lexical complexity, we use the Measure of Textual Lexical Diversity (MTLD; McCarthy and Jarvis, 2010):

4. Measure of Textual Lexical Diversity (MTLD)

MTLD represents the mean number of sequential words that maintains a given threshold of type-token ratio in a text and was chosen over basic type-token measures to avoid their sensitivity to text length.

The complexity measures are computed using the freely available linguistic complexity code described in Vajjala and Meurers (2012), integrating the lexical and syntactic features of Lu (2010, 2012). We also computed a range of specific measures, which we will introduce in the analysis of task-type effects below. To date, we have not carried out a formal evaluation of the validity of individual measures, which also depend on the validity of the linguistic analyses provided by the native language NLP tools and the related conceptual issues (Meurers and Dickinson, 2017). However, prior studies have shown that native language taggers and parsers perform fairly well on the learner data in EFCAMDAT (Geertzen et al., 2014). For example, Alexopoulou et al. (2015) evaluated the accuracy of extraction of relative clauses, reporting an F-score of 83.9%, and found that state-of-the art NLP tools provide reasonable quality.

5 Development of Complexity and Accuracy Across Proficiency

Increased complexity is often seen as an indication of the internalisation of new structures, that is, new structural representations, while accuracy indicates acquisition of finer elements of newly acquired structures and their representations (Housen and Kuiken, 2009; Larsen-Freeman, 2006; Skehan, 2003). Documenting complexity and accuracy across the learning trajectory therefore to a certain extent can model the developmental trajectory of L2 acquisition.

5.1 Results of Complexity Analysis

We first consider the profile of our global measures of linguistic complexity and the patterns of accuracy across proficiency. Starting with average sentence length as overall length-based metric, Figure 2 shows the development across the 16 Englishtown levels and for six CEFR levels resulting from grouping Englishtown levels as follows: 1–3 (A1),4–6 (A2), 7–9 (B1), 10–12 (B2), 13–15 (C1), 16 (C2).

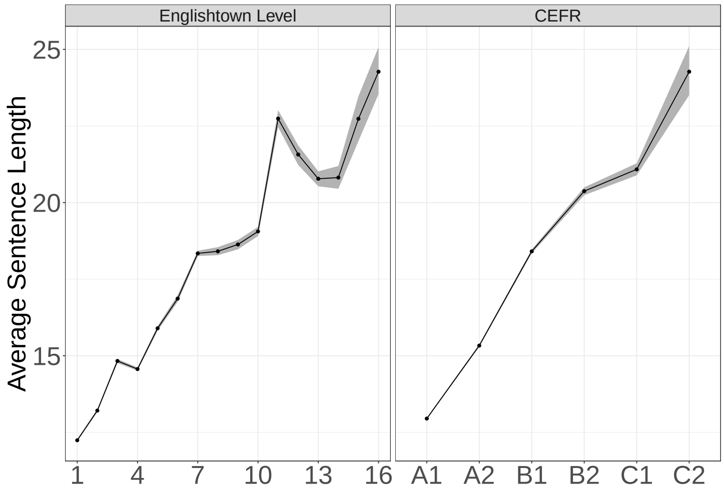


Figure 2: Average Sentence Length in words per Englishtown level (left) and CEFR aligned (right)

Here and throughout the paper, the grey shade indicates the 95% confidence intervals of the mean. The intervals are based on nonparametric bootstrapping with 1,000 bootstrap samples since Anderson-Darling tests implemented in the nortest package (Gross and Ligges, 2015) in R (R Core Team, 2016) suggested that the values did not normally distribute. The confidence intervals are generally wider at higher proficiency levels in Figures 2 through 9 because the number of writings is smaller at those levels and we can be less confident of the positions of the true values.

As expected, overall sentence length increases from beginner to advanced levels. Tapping into phrasal complexity (Norris and Ortega, 2009), Mean Length of Clause in Figure 3 takes off at A2 and increases until B2 in the CEFR-aligned graph. The subordinate clause ratio in Figure 4 shows a sharp increase from beginner levels continuing until intermediate levels where it levels off (at B2), confirming earlier findings (Perkins, 1980; Scott, 1988; Bardovi-Harlig and Bofman, 1989). Lexical diversity measured by MTLD (Figure 5) grows steadily from the very early to the very advanced levels, but, acquisition slows after B1. Generally, the individual Englishtown levels show more fluctuation, in particular for Mean Length of Clause.

In sum, all global complexity measures increase across proficiency, though some differences in phasing can be observed.

5.2 Results of Accuracy Analysis

Figure 6 shows the mean overall relative error frequency (left) and the mean relative error frequency across teacher error labels (right). The error rate generally drops as learners' proficiency advances. However, errors like phraseology (PH), possessive (PO), or verb tense (VT) show more fluctuation.

As shown in Figure 7, the use of prepositions consistently increases from early beginner to advanced levels while the accuracy shown in Figure 9 suggests a U-shape pattern, dropping at level 5 to then increase until late intermediate (level 13). In other words, the initial increase in use gives rise to a drop

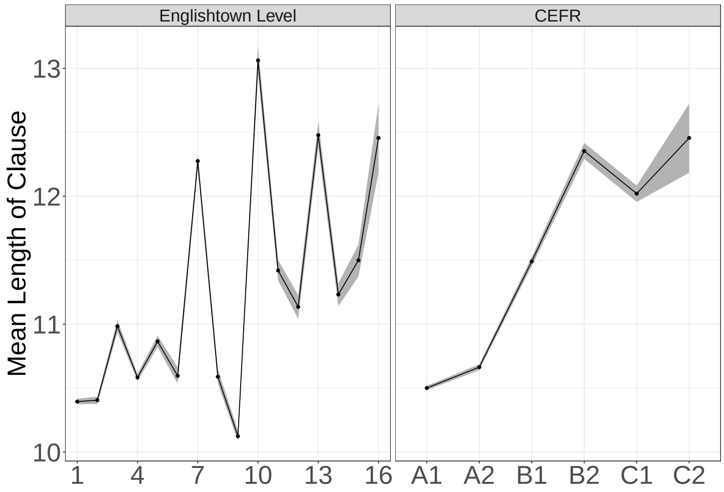


Figure 3: Mean Length of Clause in words across Englishtown (left) and CEFR-aligned levels (right)

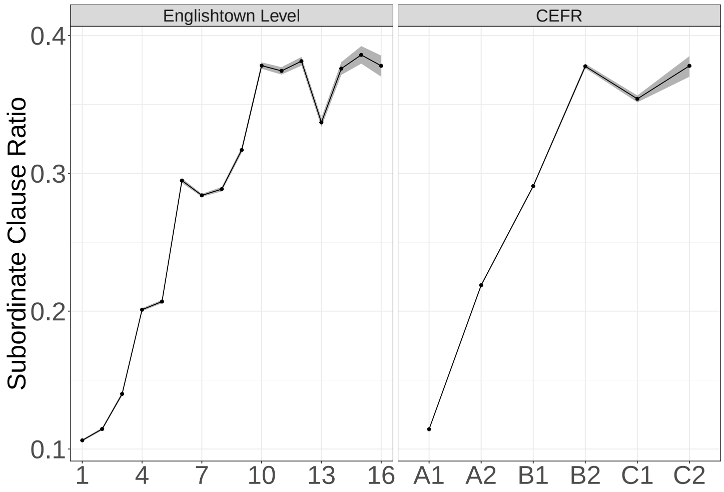


Figure 4: Subordinate Clause ratio across Englishtown (left) and CEFR-aligned levels (right)

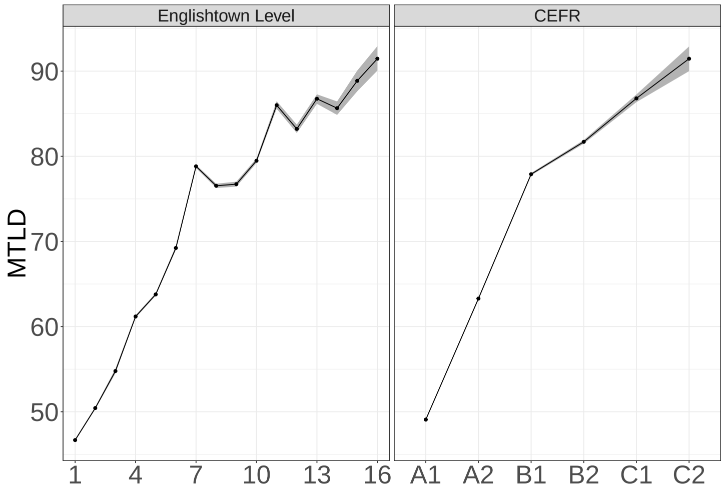


Figure 5: MTLD across Englishtown (left) and CEFR-aligned levels (right)

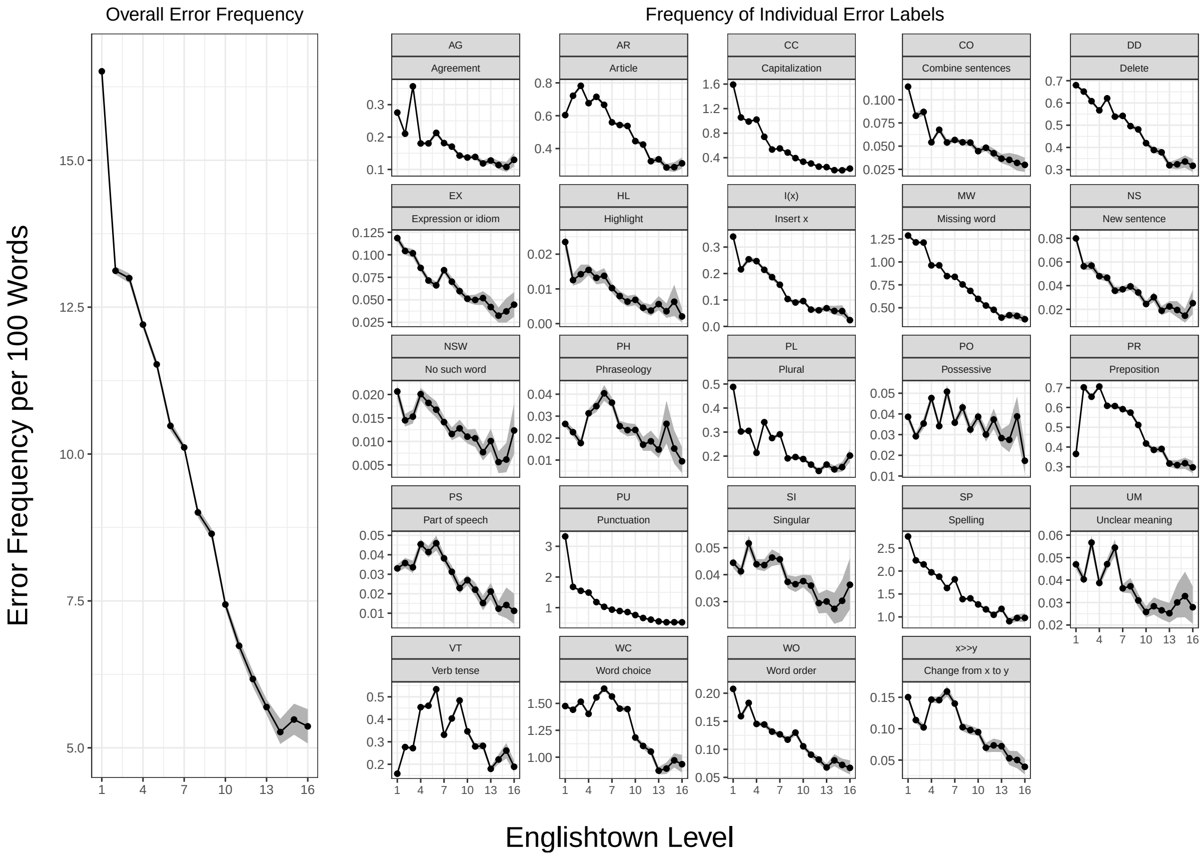


Figure 6: Error frequency across Englishtown levels

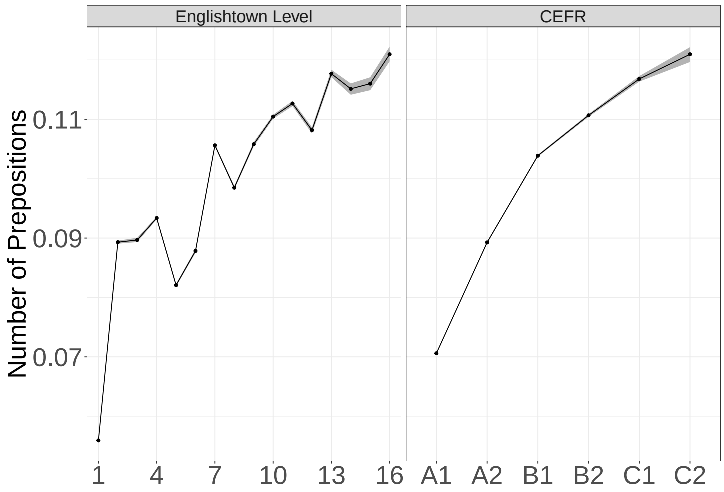


Figure 7: Number of prepositions across Englishtown (left) and CEFR-aligned levels (right)

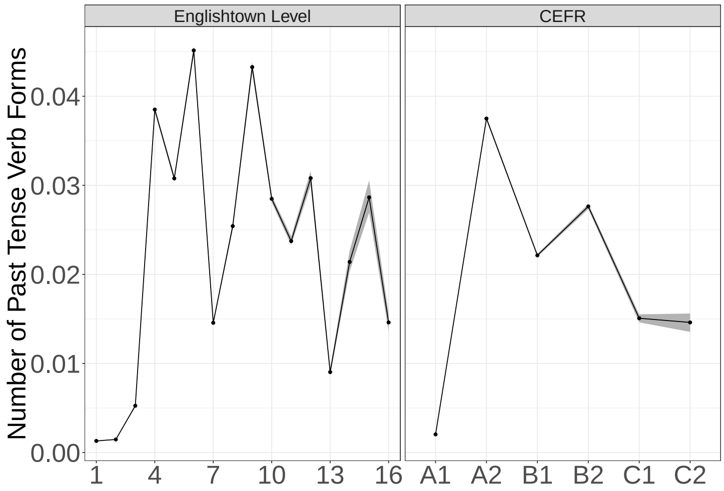


Figure 8: Number of past tense verb forms across Englishtown (left) and CEFR-aligned levels (right)

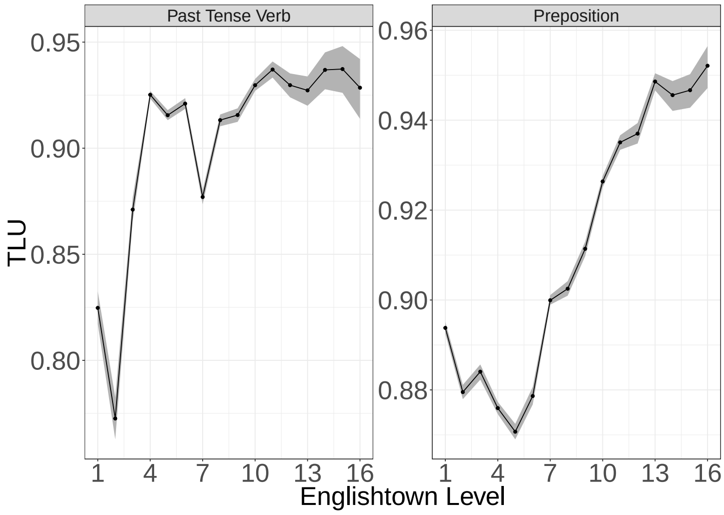


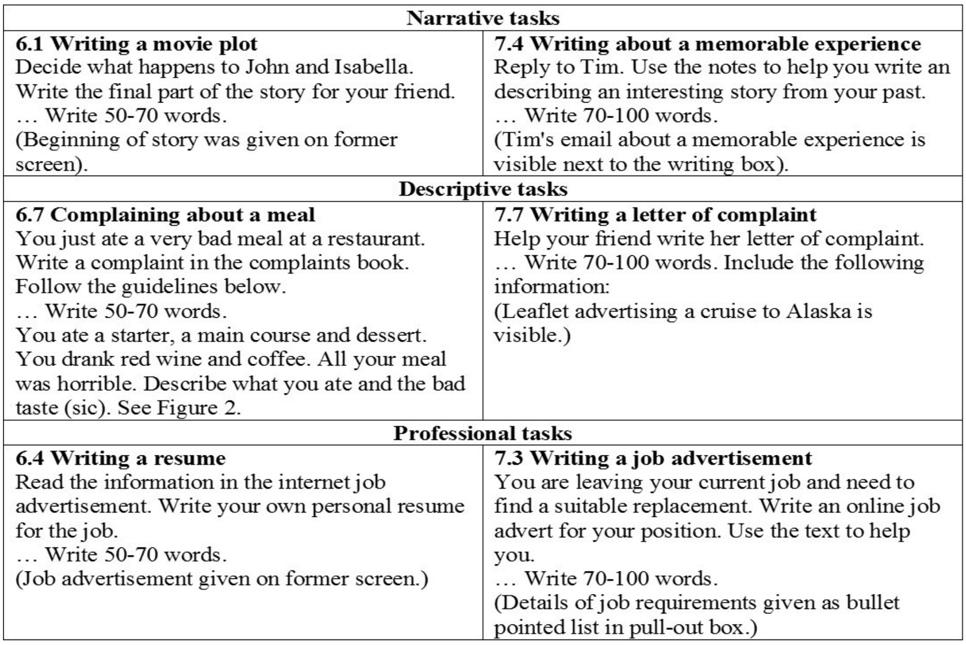
Figure 9: Accuracy of past tense verbs (left) and prepositions (right) across Englishtown levels

in accuracy which reaches ceiling at around 95% at level 13. Past tense use increases from Levels 2 through 4 (Figure 8), which is matched by the rise of accuracy (Figure 9). The comparison of these two features thus demonstrates that the extent to which relative frequency and accuracy correlate varies across linguistic features.

6 Effect of Task Design Features and Instructional Context

6.1 Three Pairs of Tasks: Narrative, Descriptive and Professional Tasks

Broadly speaking, the writing activities contained in EFCAMDAT can be characterized as tasks, since L2 writers work towards a non-linguistic outcome, e.g., write a complaint or apply for a job, and are engaged in language use to achieve that goal (Samuda and Bygate, 2008). In this sense, EFCAMDAT can be seen as a task-based corpus. However, task design is not linked to any theoretical framework. We cannot,therefore, draw directly from EFCAMDAT tasks varying in specific parameters in a controlled and minimal way as done typically in experimental designs. Nonetheless, the task-based frameworks provide us with categories and features that allow us to characterize EFCAMDAT tasks and evaluate a range of predictions regarding the influence of task design features on the elicited language. Figure 10 presents the six tasks from intermediate learners (CEFR B1, Englishtown 6 and 7), we selected exemplifying three task types: narrative, descriptive, and professional. See the supplementary material for a more elaborate description of the tasks.



Narrative tasks	
6.1 Writing a movie plot Decide what happens to John and Isabella. Write the final part of the story for your friend. … Write 50-70 words. (Beginning of story was given on former screen).	**7.4 Writing about a memorable experience** Reply to Tim. Use the notes to help you write an describing an interesting story from your past. … Write 70-100 words. (Tim's email about a memorable experience is visible next to the writing box).
Descriptive tasks	
6.7 Complaining about a meal You just ate a very bad meal at a restaurant. Write a complaint in the complaints book. Follow the guidelines below. … Write 50-70 words. You ate a starter, a main course and dessert. You drank red wine and coffee. All your meal was horrible. Describe what you ate and the bad taste (sic). See Figure 2.	**7.7 Writing a letter of complaint** Help your friend write her letter of complaint. … Write 70-100 words. Include the following information: (Leaflet advertising a cruise to Alaska is visible.)
Professional tasks	
6.4 Writing a resume Read the information in the internet job advertisement. Write your own personal resume for the job. … Write 50-70 words. (Job advertisement given on former screen.)	**7.3 Writing a job advertisement** You are leaving your current job and need to find a suitable replacement. Write an online job advert for your position. Use the text to help you. … Write 70-100 words. (Details of job requirements given as bullet pointed list in pull-out box.)

Figure 10: Task Prompts

Table 1 summarizes the characteristics of the six tasks with reference to the Limited Attentional Capacity Model (Skehan, 1998) and the Cognition Hypothesis (Robinson and Gilabert, 2007). We explain the meaning of the labels used in Table 1 as we present the first narrative task, the "movie plot" (6.1), in comparison with the other tasks as needed.

*** INSERT TABLE 1 HERE ***

In terms of Skehan's code complexity, both narratives create a high vocabulary load because all lexis needs to come from the learner. In contrast, the elaborate prompt of the holiday complaint (7.7) provides most of the vocabulary needed. Regarding cognitive complexity, the two narratives stand out for their fairly low structure and low to medium clarity since they are free writing tasks. By contrast, the professional tasks involve high structure because expectations about how a good resume or a job ad looks like in terms of form and layout are genre-specific (e.g., a bullet pointed list). No vocabulary information is provided for the CV (6.4), but for the job ad (7.3) most of the relevant lexis is given by the pull-out prompt, adding also to clarity and structure. As we do not hold information of how familiar writers are with a given task and under which conditions they were working, no characterization regarding cognitive familiarity and communicative stress is made here.

In light of Robinson's Cognition Hypothesis most tasks involve a high number of elements except for the job advertisement, which seems to be the simplest task in this respect. Regarding "here & now" vs. "there & then", the movie plot is interesting because it plays in the there, but the first part of the story presented to learners uses the present tense, which could be interpreted as "now".

Telling a story (narratives) may well involve reasoning, while no reasoning is expected in the professional tasks since they primarily consist of factual statements (6.4) or the description of desirable applicants (7.3). The factor "perspective taking" is important for the continuation of the movie plot, because the point of view of the different characters in the story needs to be taken into account. In contrast, the

Table 1: Overview of Task Characteristics

	Narrative		Descriptive		Professional	
Framework/ Dimension of Task Complexity	**6.1 Continuing movie plot**	**7.4 Telling memorable experience**	**6.7 Complaint restaurant**	**7.7 Complaint cruise**	**6.4 Write resume**	**7.3 Write job advertisement**
Limited Attentional Capacity Model (Skehan, 1998)						
Code complexity	High voc.load	High voc.load	Medium voc.load	Low voc.load	High voc.load	Low voc.load
Cognitive Complexity	Low structure and clarity	Low structure, medium clarity	Medium structure, high clarity	Medium structure, high clarity	High structure, low clarity	Medium structure, high clarity
Cognition Hypothesis (Robinson and Gilabert, 2007) resource-directing variables						
Number of Elements	Many - different	Many - different	Medium - similar	Many - similar	Many - similar	Medium - different
Here & now	There&now	There&then	Here&now	There&then	Here&now	Here&now
Reasoning	Probably	Likely	No	Likely	No	No
Perspective taking	Yes	No	No	Yes	No	Yes

memorable experience is a story told from the writer's perspective. The cruise complaint needs to be written for a friend and the job advertisement needs to take up the perspective of a job hunter.

Unlike experimental studies, this characterization of tasks is post-hoc and therefore lacks the fine manipulations that would ideally be built on the theoretical underpinnings presented in the former sections. Moreover, many environmental factors, e.g., the fact that learners worked in a digitally mediated context (Ziegler, 2016) could not be controlled for. However, these characterizations do capture some core properties of the tasks in question which are bound to impact on language use and allow us to form predictions regarding the expected language.

6.2 Predictions based on task design features

Skehan's (1998) **Code Complexity** singles out the cruise complaint (7.7) and the job advert (7.3) as two tasks with low vocabulary load, keeping cognitive resources available to focus on form (complexity and accuracy). This prediction might hold within (e.g., meal vs. cruise complaint) or across task types (e.g., job ad vs. movie plot).

Prediction 1: Higher accuracy and syntactic complexity for cruise complaint and job ad.

The high **Clarity and Structure** of the professional tasks are expected to increase accuracy because these characteristics ease cognitive demands and reduce processing pressure. The two narrative tasks with low clarity and structure are expected to show lower accuracy.

Prediction 2: High accuracy for professional tasks, low accuracy for narratives.

In terms of Robinson's Cognition Hypothesis, the cruise complaint letter is the most complex task (Table 1), therefore, it is predicted to elicit the most complex (syntax and lexis) and accurate language in contrast to the job ad and the restaurant complaint.

Prediction 3: Higher linguistic complexity and accuracy for cruise complaint, lower for restaurant complaint and job ad.

If we compare prediction 1 with prediction 3 we notice that they both predict high syntactic complexity for the cruise complaint, in particular in comparison to the restaurant complaint, but diverge regarding the job ad.

6.3 Predictions based on task type and instructional focus

Narratives require writers to introduce entities, events, location (including ways to characterize them, modifiers, relative clauses) and linguistic means to link them to one another (incl. adverbs, e.g., firstly, secondly). The movie plot narrative asks for a story about different people, and will be mostly told in 3rd person, which is likely to cause more agreement errors.

Prediction 4a: Narratives will elicit referential phrases, modifiers, relative clauses, temporal and locative adverbs, subordination. High numbers of agreement errors are likely.

The instructional focus of the preceding units and the respective prompt of the movie plot will elicit (mostly) present tense, while the memorable experience is more likely to elicit past tense (simple past vs. past continuous).

Prediction 4b: The movie plot will elicit (mostly) present tense, while the memorable experience mostly past (simple and continuous).

The **Descriptive tasks** (meal complaint, cruise complaint) are expected to elicit many copulas and predicative adjectives. The preceding instruction for both tasks focuses on the use of adjectives and, therefore, a high number of adjectives can be expected. The prompts are likely to elicit past tense. In terms of syntactic complexity, descriptive tasks can be expected to require simple (rather than complex) sentences linked through coordination rather than subordination.

Prediction 5: Descriptive tasks will elicit copulas, adjectives, low syntactic complexity and past tense.

Given the formal setting of the **Professional tasks** (resume, job ad) we expect high levels of lexical sophistication, that is, more infrequent words. A nominal style (lexical density, e.g., noun/verb ratio might be high) is likely, including complex NPs, while syntactic complexity outside nominals is predicted to be low due to list-like writings. Furthermore, we expect low numbers of agreement errors, due to absence of many different persons acting.

Prediction 6: Professional tasks will elicit sophisticated lexis, nominal style and complexity, low syntactic complexity outside nominals, low number of agreement errors.

6.4 Results

6.4.1 Global Measures of Complexity and Accuracy per Task

Figure 11 shows global complexity measures for our six tasks, along with their bootstrap-based 95% confidence intervals. There are task effects with task type emerging as a key factor in some cases: e.g. professional tasks pattern together for Mean Length of T-Unit and Mean Length of Clause while narratives pair at the higher end of subordinate clauses per T-Unit. There is also variation cutting across task-type as depicted in the graphs for MTLD and Subordinate clauses.

Figure 12 depicts the mean relative error frequency of each error type (left), mean overall relative error frequency (top right), and the mean accuracy of prepositions and past tense verbs (bottom right) for each

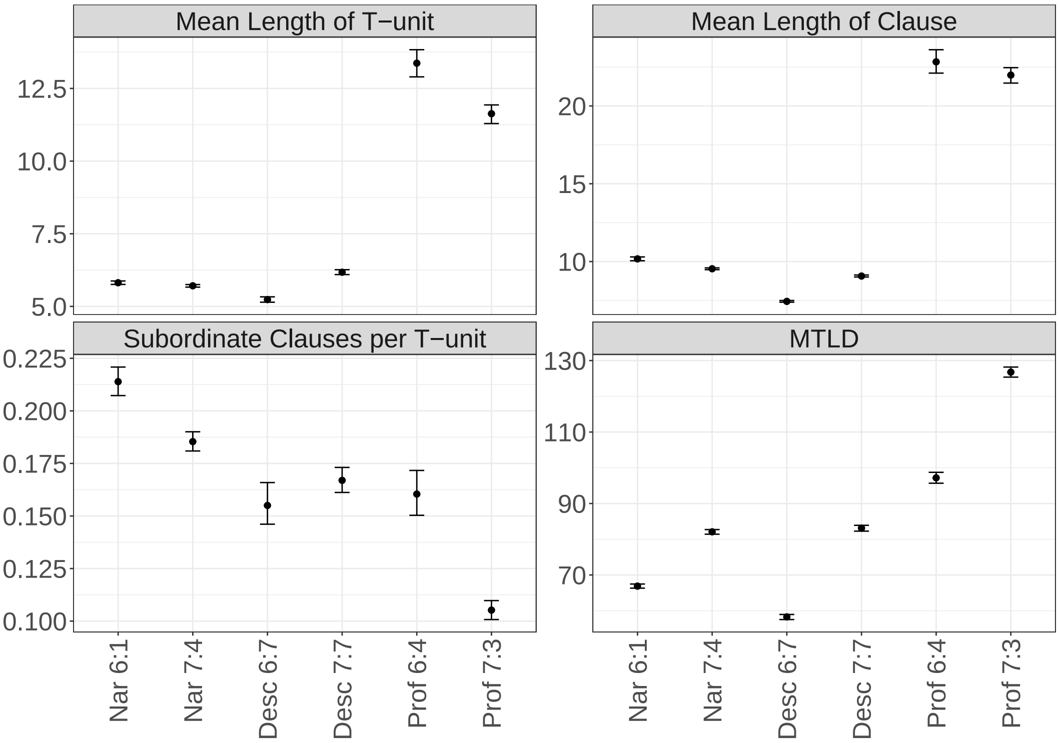


Figure 11: Global complexity measures per task: Mean Length of T-unit (top left), Mean Length of Clause (top right), Subordinate Clauses per T-unit (bottom left) and MTLD (bottom right).

of the target tasks. Again, we can visually grasp task effects on the rate of individual errors as well as the accuracy of prepositions and past tense.

6.4.2 Effects of task design features on linguistic complexity and accuracy

Due to their low code complexity (Skehan, 1998; cf. Prediction 1), the cruise complaint (7.7) and the job ad (7.3) should show relatively high syntactic complexity and accuracy. In Figure 11, task 7.3 is indeed eliciting highly complex language, with the highest score for lexical complexity and consistently high scores for the other complexity measures. The only exception is subordination, which was expected for a list-like writings (Prediction 6). Prediction 1 is not confirmed, though, for the cruise complaint, which shows medium to low linguistic complexity. This fact also disconfirms Prediction 3, where we predicted high linguistic complexity for task 7.7 based on high cognitive complexity. When comparing code-complexity effects within task-type, we see that the cruise complaint (7.7) indeed elicits more complex language than the restaurant complaint (6.7), confirming the prediction. Note that there is no increase in the relevant measures between levels 6 and 7 (Figures 2-5), indicating that the difference between 6.7 and 7.7 is not due to the higher proficiency of the latter. For 7.3, we might argue for lower overall error frequency, but 7.7 does not stand out in terms of accuracy.

The accuracy figures do not confirm Prediction 2. Even though the two professional tasks have low error rates and one of the two narratives stands out for its high error frequency, the memorable experience and the CV give rise to similar numbers.

Prediction 3, based on Robinson and Gilabert (2007), is confirmed for the meal complaint (6.7) and the the job ad (7.3) which consistently elicit language of low complexity (except for subordination). Parallel effects on accuracy are visible for the job ad only in terms of overall error frequency and the use of prepositions. In contrast, Prediction 3 is mostly disconfirmed for task 7.7 (Prediction 1). As the cruise

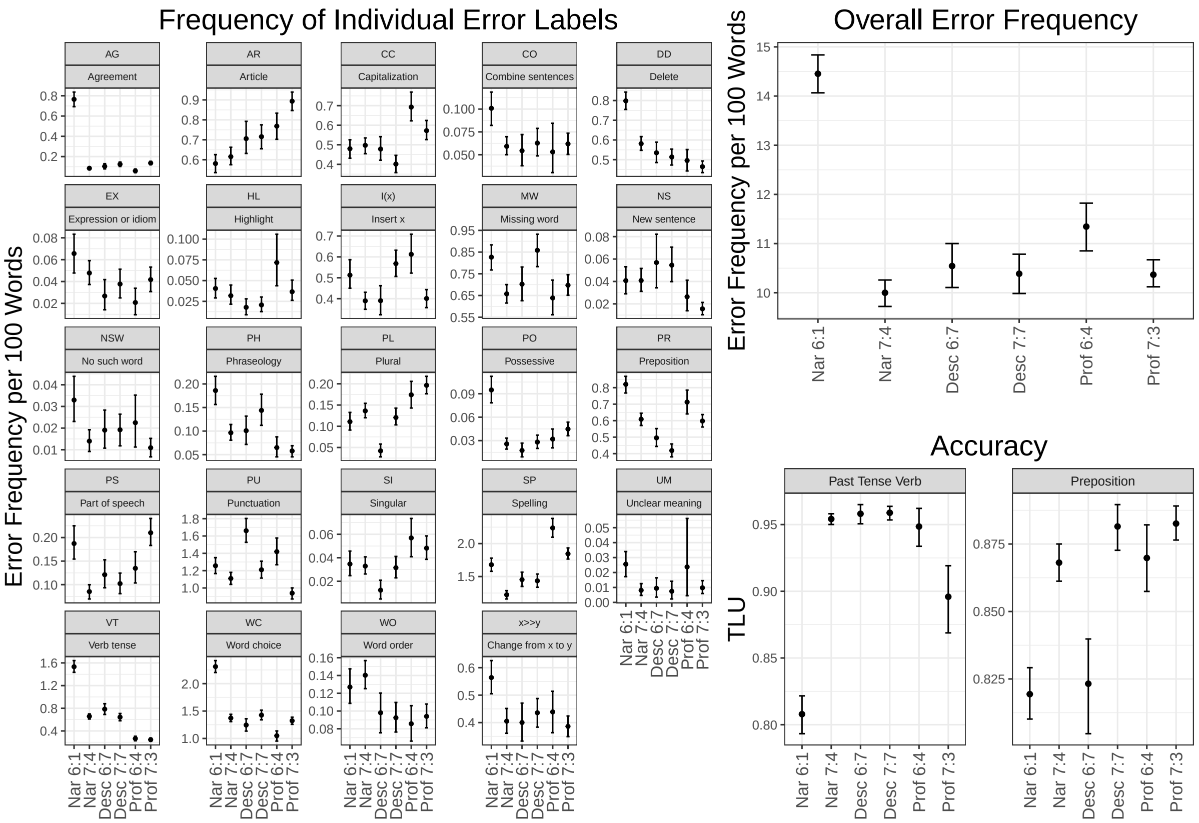


Figure 12: Error frequency and accuracy per task: mean relative error frequencies (left), mean overall relative error frequency (top right), mean accuracy of past tense verb and prepositions (bottom right).

complaint does indeed elicit more complex language than the meal complaint, Prediction 3 is confirmed within task type.

In sum, predictions for linguistic complexity were only partially confirmed and most apparent in relation to the simplest task (meal complaint/job ad) in terms of cognitive task complexity (Robinson and Gilabert, 2007), particularly, when comparing within – but not across – task-types. Accuracy effects were less pronounced overall and could not be accounted for by task design features in this study.

6.5 Effects of Task Type and Instruction

We focus on narrative tasks for effects of task type and instruction, while only exemplary findings for descriptive and professional tasks are presented.

Two shortened random excerpts in (1 a./b.) from the movie plot (6.1) and memorable experience (7.7) task illustrate Prediction 4a: high use of demonstratives, pronouns, possessives and connectives to temporally link events; a variety of subordinate clauses to identify or describe referents and causation; the use of present and past tense (task 6.1). Writers in task 7.7 use aspectual distinctions (simple past vs. past continuous) and because the story is told from the writer's perspective there are fewer anaphoric expressions.

(1) a. **Narrative 6.1: Movie plot** Writing ID: 122619, L1: Brazilian
Nothing was stronger than John's love. This poor boy was persuaded for Isabella that actually, was lesbian. [...] At the end, Isabella kills John using a poisoned sardine and finally,

lives happily ever after with Sara, the one that she really loves. ;)

b. **Narrative 7.7: Memorable experience** Writing ID: 271872, L1: Brazilian
Dear Tim, When I was 10, in a weekend, I went with my family to my uncle's farm. In that place there was a big artificial lake. My family and relatives were in around it, while my brothers, cousins and I were swimming [...]

To capture referential cohesion in the six selected tasks, we used measures of Local and Global Argument Overlap (McNamara et al., 2014, p. 65), capturing that the same noun or pronoun occurs in two sentences, which in the local variant must be adjacent. As can be seen in the top graphs of Figure 13 both narratives score high in these measures. In accordance with Prediction 4, on syntactic measures narratives show high numbers of subordinate clauses (cf. Figure 11) and Wh-phrases but low noun phrase complexity (cf. bottom graphs of Figure 13).

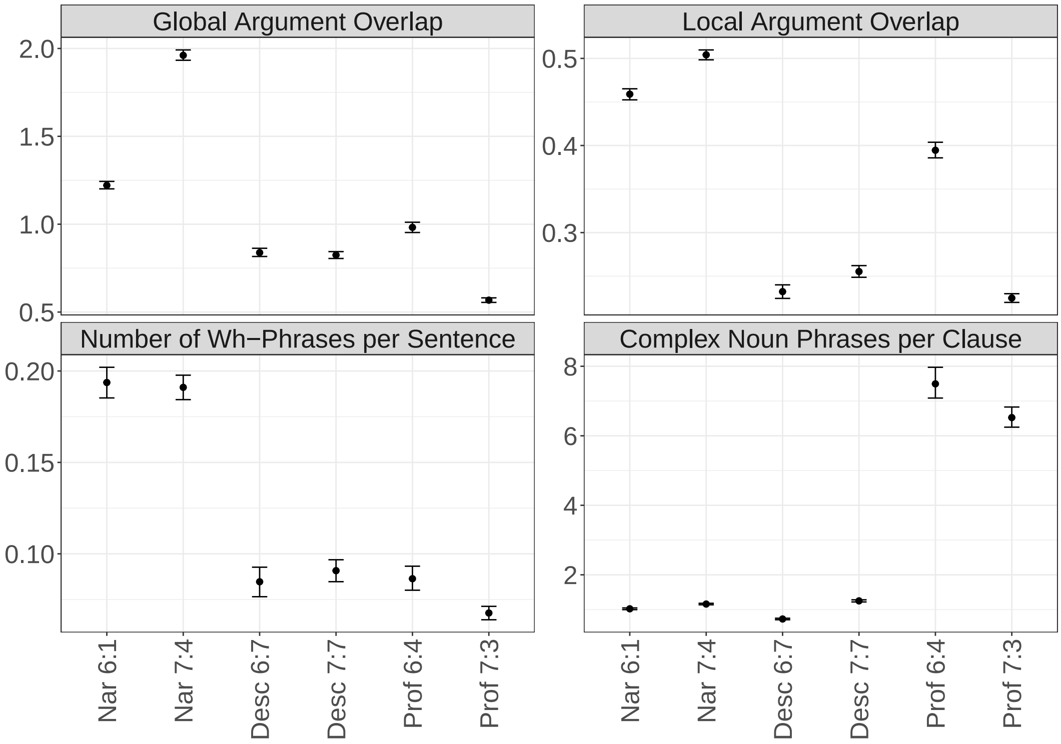


Figure 13: Top: Nominal discourse cohesion: global (left) and local (right) argument overlap count. Bottom: Number of Wh-Phrases per Sentence (left) and Complex Noun Phrases per Clause (right).

In accordance with Prediction (4a,b), we also found the "movie plot" to have high numbers of 3rd person simple present forms, while the "memorable experience" elicited past tense (Figure 14). Confirming our expectation regarding third person *-s*, the movie plot has by far the largest relative frequency of agreement errors (see left panel in Figure 12).

Confirming Prediction 5, descriptive tasks elicit high numbers of adjectives, while scoring low on global (Figure 11) and specific (bottom graphs Figure 13) complexity measures. Figure 14 confirms that they elicit past tense.

Finally, Prediction 6 for the professional tasks is largely confirmed. We see high numbers of complex nominals (bottom graphs Figure 13) and high global complexity except for subordinate clauses (Figure 11). The high number of gerundial forms seen in Figure 14 is linked to the nominal style of this genre. Similarly, these tasks showed high scores on global (MTLD in Figure 11).

In sum, the core predictions relating to effects of task type and instructional focus are confirmed.

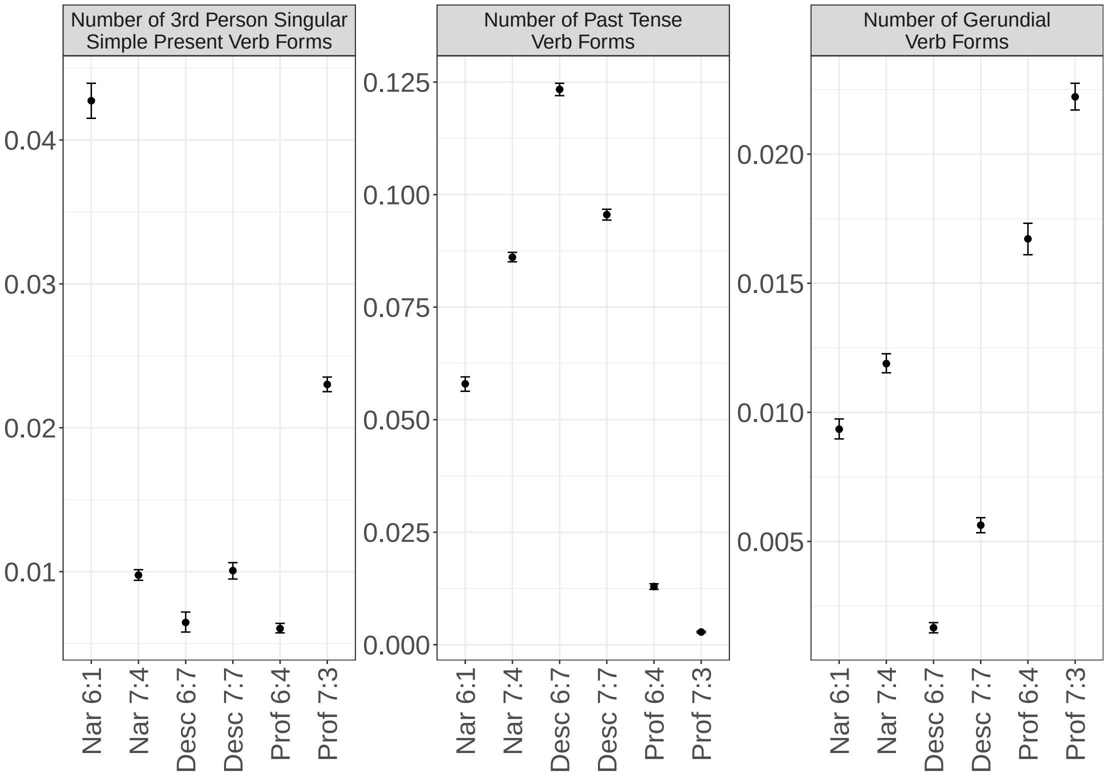


Figure 14: Number of 3rd person singular simple present verb forms (left), past tense verb forms (center), and gerundial *-ing* forms (right)

7 General Discussion and Conclusions

The research background of this paper is to develop and evaluate approaches and methods for exploiting learner corpora, in particular, the large amounts of data that online language learning technology is already making available to SLA research. The collaboration between corpus linguists, SLA researchers and computational linguists is a necessary condition for this research program. In this paper we explored this synergy through investigating task effects on learner language, as conceived within task-based frameworks. Our aim was twofold: first, to show that understanding of task effects and related TBLT insights is indispensable if we are to model learner language in large learner corpora; second, to demonstrate that big corpora can contribute to TBLT research and complement the standard experimental paradigm. We focused on linguistic complexity and accuracy of language use.

The crucial contribution of a corpus like EFCAMDAT is the possibility of an analysis across proficiency *and* across tasks in *a single design*. We presented such an analysis and were able to show that linguistic complexity and accuracy can characterise both development from early to advanced proficiency and tasks. Though not in itself surprising, this result is meant to demonstrate how the richness of learner corpora can enable multi-factorial designs of a large scope. Our exploratory study is viewed as the first step to an extended investigation to better understand how the general picture obtained here is shaped by the interplay of varying factors such as L1, variability in the writings of individual learners across tasks and proficiency, variability in the effects of same task type (e.g. narrative) across proficiency etc.

We showed that linguistic complexity can model the language elicited by our three task types – narrative, descriptive and professional – but that such modeling needs to employ a rich and diverse set of global and specific measures, e.g., gerundial forms in Figure 14. Interestingly, some task-dependent features were not predicted. For example, the restaurant complaint task elicited a large number of irregular verbs, a probable consequence of the irregularity of verbs relating to food (e.g., eat, drink) and the copula. This is an example where corpus explorations can help identify naturally emerging structures in a task, potentially revealing essential, natural and useful structures (Loschky and Bley-Vroman, 1993).

Task-based insights from this exploration are directly beneficial to researchers developing tasks targeting the elicitation of specific features or a variety of forms in order to evaluate the acquisition of specific grammatical phenomena (Tracy-Ventura and Myles, 2015). Identifying language forms elicited by specific task types can also highlight gaps in the features elicited in current corpora as well as complement experimental research (Gilquin and Gries, 2009).

The contribution of NLP for an enriched inventory of reliable complexity measures is vital. For instance, the nominal discourse cohesion measures of local and global argument overlap distinguished the two narrative tasks (Figure 14) while we found that measures like Age of Acquisition also distinguish sharply between the three task-types, an interesting and unexpected finding.

Our results bear on the ongoing debate regarding the relation between task type and task complexity, indicating that task complexity does indeed impact on linguistic complexity, but its effect is mainly visible within task type (e.g., simple vs. complex descriptive). When global complexity measures are considered (Figure 11), it is task type rather than task complexity that affects linguistic complexity, which confirms Yoon and Polio (2016). Yet, it may well be that the impact of task type is more readily detected in corpora exactly because finer aspects of task-design features cannot be controlled for.

Error rate generally did not discriminate between tasks (again mirroring Yoon and Polio, 2016), but accuracy of specific features did - be it not in a predictable way. In the future, the interplay between complexity and accuracy (Skehan, 1998) could be clarified through an analysis of the variability of individual learner writings.

To conclude, this has been an exploratory investigation and comes with its limitations. As such, some of our results may be tentative or inconclusive. We have, nevertheless, demonstrated the kind of insights that can be gained through combining the developmental with the task-based perspective in the study of complexity and accuracy and the possibilities learner corpora and big learner data more generally open for wide scope multi-factorial designs that could tease apart the impact of distinct factors and yield rich inventories of features modelling development as well as tasks.

References

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

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press, Oxford.

Bardovi-Harlig, K. and Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11(1):17–34.

Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Biber, D., Gray, B., and Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*.

Bouwer, R., Béguin, A., Sanders, T., and van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, 32(1):83–100.

Crosthwaite, P. (2016). L2 English article use by L1 speakers of article-less languages: A learner corpus study . *International Journal of Learner Corpus Research*, 2(1):68–100.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press, Oxford.

Foster, P. and Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18(3):299–323.

Foster, P. and Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, 36:98–116.

Gablasova, D., Brezina, V., McEnery, T., and Boyd, E. (2015). Epistemic stance in spoken l2 english: The effect of task and speaker style. *Applied Linguistics*.

Geertzen, J., Alexopoulou, T., and Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In Millar, R. T., Martin, K. I., Eddington, C. M., Henery, A., Miguel, N. M., Tseng, A., Tuninetti, A., and Walter, D., editors, *Selected proceedings of the 2012 Second Language Research Forum. Building bridges between disciplines*, pages 240–254. Cascadilla Proceedings Project.

Gilquin, G. and Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1):1–26.

Granger, S. (2008). Learner corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus linguistics. An international handbook*, pages 259–275. Walter de Gruyter, Berlin / New York, NY.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G., and Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL*, 19(3):252–268.

Gross, J. and Ligges, U. (2015). *nortest: Tests for Normality*. R package version 1.0-4.

Hawkins, J. A. and Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge University Press, Cambridge.

Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics*, 41(4):667–683.

Housen, A. and Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, 60(2):237–263.

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *The Language Learning Journal*, pages 1–19.

Jackson, D. O. and Suethanapornkul, S. (2013). The Cognition Hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning*, 63(2):330–367.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, 20(2):148–161.

Kuiken, F. and Vedder, I. (2008). Cognitive task complexity and written output in italian and french as a foreign language. *Journal of Second Language Writing*, 17(1):48–60.

Lambert, C. and Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, 35(5):607–614.

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

Long, M. H. (2016). In defense of tasks and TBLT: Nonissues and real issues. *Annual Review of Applied Linguistics*, 36:5–33.

Loschky, L. and Bley-Vroman, R. (1993). Grammar and task-based methodology. In Crookes, G. and Gass, S., editors, *Tasks and Language Learning: Integrating Theory and Practice*, pages 123–167. Multilingual Matters, Philadelphia.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

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

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Languages Journal*, pages 190–208.

McCarthy, P. M. and 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.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Meurers, D. (2012). Natural language processing and language learning. In Chapelle, C. A., editor, *Encyclopedia of Applied Linguistics*, pages 4193–4205. Wiley, Oxford.

Meurers, D. (2015). Learner corpora and natural language processing. In Granger, S., Gilquin, G., and Meunier, F., editors, *The Cambridge handbook of learner corpus research*, pages 537–566. Cambridge University Press.

Meurers, D. and Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning, Special Issue on Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and Interpretation*. this volume.

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. In Sato, M. and Loewen, S., editors, *Hanbook of instructed Second Language Acquisition*, pages 50–68. Routledge, London and New York.

Murakami, A. (2014). *Individual variation and the role of L1 in the L2 development of English grammatical morphemes: Insights from learner corpora*. PhD thesis, University of Cambridge.

Murakami, A. and Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3):365–401.

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003*, pages 572–581, Lancaster.

Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4):555–578.

Ott, N., Ziai, R., and Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Schmidt, T. and Wörner, K., editors, *Multilingual Corpora and Multilingual Corpus Analysis*, pages 47–69. John Benjamins, Amsterdam.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly*, 14(1):61–69.

Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition*, 6(1):69–78.

Plonsky, L. and Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, 36:73–97.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47:101–143.

Polio, C. and Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, 26:10–27.

Quixal, M. and Meurers, D. (2016). How can writing tasks be characterized in a way serving pedagogical goals and automatic analysis needs? *CALICO Journal*, 33(1).

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Learning Language*, 45(1):99–140.

Robinson, P. (2011). *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*. John Benjamins, Amsterdam.

Robinson, P. and Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *IRAL-International Review of Applied Linguistics in Language Teaching*, 45(3):161–176.

Samuda, V. and Bygate, M. (2008). *Tasks in second language learning*. Palgrave Macmillan, New York, NY.

Schachter, J. and Celce-Murcia, M. (1977). Some reservations concerning error analysis. *TESOL Quarterly*, 11(4):441–451.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of Inter-*

national Conference on New Methods in Language Processing, pages 44–49, Manchester, UK.

Scott, C. M. (1988). Spoken and written syntax. In Nippold, M. A., editor, *Later Language Development: Ages Nine through to Nineteen*, pages 41–91. Little Brown, Boston, MA.

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

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

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

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173.

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.

Way, D. P., Joiner, E. G., and Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of french. *Modern Language Journal*, pages 171–184.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press, Cambridge.

Yoon, H.-J. and Polio, C. (2016). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*.

Ziegler, N. (2016). Taking technology to task: Technology-mediated TBLT, performance, and production. *Annual Review of Applied Linguistics*, 36:136–163.