Instructed
Second Language
Acquisition

ISLA (PRINT) ISSN 2398–4155
ISLA (ONLINE) ISSN 2398–4163

Article

Effects of task type on morphosyntactic complexity across proficiency: evidence from a large learner corpus of A1 to C2 writings

Marije Michel, Akira Murakami, Theodora Alexopoulou and Detmar Meurers

Abstract

This study investigates the effect of instructional design on (morpho)syntactic complexity in second language (L2) writing development. We operationalised instructional design in terms of task type and empirically based the investigation on a large subcorpus (669,876 writings by 119,960 learners from 128 tasks at all Common European Framework of Reference for Languages levels) of the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen, Alexopoulou and Korhonen 2014).

First, the 128 task prompts were manually categorised for task type (e.g. argumentation, description). Next, developmental trajectories of syntactic complexity from A1 to C2 were established using a variety of global (e.g. mean length of clause) and specific (e.g. non-third person singular present tense verbs) measures extracted using natural language processing techniques. The effects

Affiliations

Marije Michel: University of Groningen, NL
email: m.c.michel@rug.nl
Akira Murakami: University of Birmingham, UK
email: a.murakami@bham.ac.uk
Theodora Alexopoulou: Cambridge University, UK
email: ta259@cam.ac.uk
Detmar Meurers: Tübingen University, Germany
email: dm@sfs.uni-tuebingen.de

ISLA VOL 3.2 2019 124–152 doi: https://doi.org/10.1558/isla.38248

of task type were analysed using the categorisation from the first step. Finally, tasks that showed atypical behaviour for a measure given their task type were explored qualitatively.

Our results partially confirm earlier experimental and corpus-based studies (e.g. subordination associated with argumentative tasks). Going beyond, our large-scale data-driven analysis made it possible to identify specific measures that were naturally prompted by instructional design (e.g. narrations eliciting wh-*phrases). We discuss which measures typically align with certain task types and highlight how instructional design relates to L2 developmental trajectories over time.*

KEYWORDS: TASK-BASED LANGUAGE TEACHING; LEARNER CORPUS; MORPHOSYNTACTIC COMPLEXITY; NATURAL LANGUAGE PROCESSING; COMPUTATIONAL LINGUISTICS

Introduction

In recent years, second language (L2) learners worldwide have started using online language learning materials. Both public and commercial language schools are nowadays providing L2 lessons via digital tools and online learning platforms (Benson and Reinders 2011). Often, L2 learners using these online resources will submit their written/spoken task performances to receive feedback. Combined into a learner corpus, these L2 samples can become treasure troves for researchers interested in Instructed Second Language Acquisition (ISLA; Loewen 2015), given that these samples come from learners all over the world who have been working on the same set of learning activities at different proficiency levels. Drawing on natural language processing (NLP) techniques, analyses of these large corpora (e.g. Alexopoulou *et al.* 2017) can provide insights into the interplay of language development and instructional design over time, overcoming the limitations of smaller scale, one-off empirical investigations typically used in ISLA research (Meurers and Dickinson 2017). In addition, taking instructional design (e.g. features of the tasks L2 users were working on) as a starting point when exploring a wide range of measures in these kinds of corpora can target limitations in learner corpus research (LCR), where corpora typically consist of a large set of similar texts (e.g. argumentative essays), and where investigations often focus on error analyses.

The current study aims to address the shortcomings of both ISLA and LCR, and merge their strengths by investigating written production in the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen, Alexopoulou and Korhonen 2014). EFCAMDAT gives access to 1.2 million scripts written by more than 120,000 learners of English working on 128 different writing tasks spread over sixteen proficiency levels from A1 to C2 according to the Common European Framework of Reference for Languages (CEFR). Using NLP analyses integrated into the Common Text Analysis Platform (CTAP) system (Chen and Meurers 2016) to extract a broad set of measures of linguistic complexity, we explore how instructional design, here operationalised as the effect of task type, might explain variability in (morpho)syntactic complexity identified in the EFCAMDAT.

Integrating ISLA, L2 writing research, LCR and NLP

The field of ISLA has started to emancipate itself from SLA, as it specifically wishes to 'understand how the systematic manipulation of the mechanisms of learning and/or the conditions under which they occur enable or facilitate the development and acquisition of an additional language' (Loewen 2015:2). Several outlets of academic work in ISLA have emerged, including textbooks (e.g. Loewen 2015), handbooks (e.g. Loewen and Sato 2017; Sato and Loewen 2019) and the current journal.

The present study forms part of this joint endeavour to give insights into how instructional design might shape L2 development with a specific focus on syntactic complexity. We use task type as a proxy to study instructional design. Specifically, we aim to investigate which task type characteristics prompt what kind of language use in writing as gauged using measures of (morpho)syntactic complexity.

Investigating effects of task type on learner language has been a fruitful line of research, given that it has the potential to inform language pedagogy and assessment alike. Specifically, achieving a better understanding of the complex relationship between task type (input) and learner language (output) can give important insights into how individual task design features and sequencing tasks according to developmental patterns might support language learning (Baralt, Gilabert and Robinson 2014). It is also important in terms of assessment to understand what language a specific task type might typically elicit in order to allow learners to demonstrate their full potential in their L2 and to increase the validity of assessment tasks.

Most research into L2 instructional design has drawn on smaller scale empirical studies, often comparing one or more experimental groups to a control group, with a limited number of participants (see studies reported by Nassaji 2016; Wolfe-Quintero, Inagaki and Kim 1998). Usually, these investigations focus on learners at a single proficiency level or compared one or two levels and targeted either a mix of L1 backgrounds or one or two different groups of L1 speakers (see Ferris 1994, for an early exception).

Yet, concerning effects of task manipulations, these size restrictions do not allow us to 'situate task effects within the proficiency trajectory and therefore better understand their impact on L2 development' (Alexopoulou *et al.* 2017:5). In line with Vyatkina, Hirschmann and Golcher (2015), we would argue that investigating large-scale corpora of learner language using NLP can overcome size and conceptual challenges (Meurers and Dickinson 2017), in order to make it possible to investigate how instructional design might affect the language used by learners. To build on the different strands of research, we briefly review each of them (i.e. ISLA, LCR, NLP) with a focus on task type effects in writing, given the focus of this paper.

The role of instructional design for L2 writing

Manchón (2011) sees writing as an important site for L2 learning, given that composing a text pushes learners to produce output in their L2 (Swain and Lapkin 2002) and thus allows learners to practise and promote their grammar and vocabulary knowledge. The combination of a slower speed of processing and the permanence of the written output enables learners to draw their attention to language form, which in turn helps them to build refined form–meaning connections and, as a consequence, supports L2 development (Cumming 1990; Williams 2012). Crucially, as Lee and Polio (2017:303) argue: 'If writing facilitates SLA, we should understand how writing prompts or tasks affect written language production.' Over the years, a body of research has increased our understanding of the link between instructional manipulations and learner texts (e.g. Byrnes, Maxim and Norris 2010; see also Lee and Polio 2017, for a recent review).

In particular, task type has been demonstrated to affect learner language, mostly linguistic complexity, in the context of ISLA (e.g. Kormos 2011; Kuiken and Vedder 2008; Way, Joiner and Seaman 2000; Yoon and Polio 2017), computer-assisted language learning (e.g. Quixal and Meurers 2016) and language assessment (e.g. Biber, Gray and Staples 2014; Hinkel 2009; Weigle 2002). For example, argumentative tasks have been found to elicit syntactically more complex language than narratives (Lu 2011; Polio and Yoon 2018; Vyatkina 2012), while descriptive tasks favour the use of simple *that*-complement structures (Crossley and McNamara 2014; see studies in a special issue edited by Connor-Linton and Polio 2014). Using more specific measures, Frear and Bitchener (2015) revealed more adverbial clauses (but not adjectival or nominal clauses) in argumentative than in narrative texts. In line with this earlier work, we aim to substantiate findings on task type effects on written learner language, drawing on a large learner corpus.

Corpus-based studies into L2 writing

Similar to ISLA, learner corpus research (LCR) nowadays has its own journal, handbooks by major publishers (Granger, Gilquin and Meunier 2015; Tracy-Ventura and Paquot, in preparation) and several dedicated edited volumes and special issues (e.g. Brezina and Flowerdew 2017; Rebuschat, Meurers and McEnery 2017). One current limitation of corpus-based work is that only a limited number of studies has taken ISLA into account. For example, many studies using corpora of L2 writing rely on essays of one specific genre (Bulté and Housen 2014; Connor-Linton and Polio 2014; Crossley and McNamara 2014; focusing on descriptive essays). However, not exploring differences between text genres, effects of task type or other aspects of the instructional prompt is likely to give a biased picture of what L2 writers can do and/or how they develop, given that the corpus may not be representative (Gablasova, Brezina and McEnery 2017a). In this line of study, Gablasova *et al.* (2017a:140) and Lu (2011) call for more attention to 'context-related factors', such as focus of task instructions or task input that affect the occurrence and frequency of a specific target feature or syntactic complexity score in a corpus.

Only a handful of corpus-based works acknowledge that instructional design – for example, a task learners received when producing language samples that eventually entered a corpus – may impact the type of language that can be examined using corpus techniques (e.g. Gablasova *et al.* 2017b; Meurers 2015; Ott, Ziai and Meurers 2012; Tracy-Ventura and Myles 2015; Vyatkina 2012). For L2 writing, Lu (2011) investigated different syntactic complexity measures in argumentative vs narrative essays and found generally higher complexity scores in argumentative writing. Within argumentative essays, topics requiring justification elicited more subordination and global sentence complexity (Yang, Lu and Weigle 2015), while essays on a topic not asking for causal reasoning (cf. Robinson 2001) were characterised by elaboration within the clause (e.g. complex noun phrases). More recently, Alexopoulou and co-workers (2017) showed that task type (narration vs description) influences global as well as specific measures of syntactic, lexical and discourse complexity. For example, narrative tasks at the B1 level of the CEFR yielded higher numbers of subordination and *wh*-phrases, and higher local and global argument overlap than descriptive tasks. By contrast, descriptive tasks elicited many adjectives and past tense verb forms. Within narratives, an instructional unit focusing on past tense unsurprisingly triggered high numbers of past tense verbs, while a prompt focusing on the present tense yielded a high frequency of third person *-s*. Similarly, Alexopoulou *et al.* (2015) could identify a high number of formulaic sequences that were lifted from the task prompts.

While more work is needed on task effects in L2 corpora, controlling for different contextual factors when extracting measures of linguistic complexity from large corpora can prove challenging. In particular, corpora that are collected in an educational context, such as EFCAMDAT, raise the fundamental question of how to identify and interpret the data, given the many interacting linguistic, instructional and learner factors (see Alexopoulou *et al.* 2017, for a discussion).

Measuring linguistic complexity of L2 writing using NLP methods

The complexity of the language produced by second language learners has systematically been analysed since the 1990s as part of the triad complexity, accuracy, and fluency (CAF) to capture the development of second language proficiency (Bulté and Housen 2012, 2018; Michel 2017; Norris and Ortega 2009; Pallotti 2015; Polio and Park 2016; Skehan 2009; Wolfe-Quintero, Inagaki and Kim 1998; and references therein). Recently, the automatic analysis of written language using NLP methods has made it possible to compute many of the complexity measures discussed in the literature automatically (Graesser *et al.* 2004; Kyle 2016; Lu 2010; Vajjala and Meurers 2012, 2014), with several tools being made freely available to researchers without requiring programming expertise, such as Coh-Metrix (Graesser *et al.* 2004),[1] the L2 Syntactic Complexity Analyzer (Lu 2010),[2] the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle 2016)[3] and the Common Text Analysis Platform (CTAP; Chen and Meurers 2016)[4] for English and other languages (e.g. T-Scan for Dutch; Pander Maat *et al.* 2014).[5]

At the conceptual level, the valid interpretation of NLP results still requires a sound understanding of the underlying analyses and models (see Alexopoulou *et al.* 2017; Meurers and Dickinson 2017), also taking into account the fact that the NLP tools are all based on models of well-formed, native English texts. A first analysis of the reliability and validity of such automated analysis is provided by Polio and Yoon (2018). The present study builds on these earlier achievements.

Research questions

Given the nature and limitations of the earlier research discussed above, this study pursues the following broad research question:

> What effects of task type can we identify in a large corpus of learner writings across proficiency levels?

More specifically, we examine how a large set of almost 600 measures of complexity might pattern and develop over proficiency levels in the EFCAMDAT corpus by assigning the full set of 128 tasks to six different task types. We thereby extend the work of Alexopoulou *et al.* (2017), who inspected a smaller set of six tasks (two narratives, descriptions and professional tasks each) within the same corpus using a smaller set of global and specific measures. To remain within the scope of the current special issue, this paper will focus on (morpho)syntactic measures only.

Method and design

Data

The current study draws on learner data from EFCAMDAT, an open-access corpus available at http://corpus.mml.cam.ac.uk/efcamdat2 consisting of written assignments submitted to Englishtown, the online school of EF Education First. The curriculum of Englishtown covers all proficiency levels from A1 to C2 spread over sixteen EF levels. Each EF level contains eight instructional units, at the end of which learners perform a free writing task where they respond to a task prompt without being instructed on the linguistic form(s) to use.[6] EFCAMDAT is a collection of the writings submitted to those prompts ranging from A1 to C2 level. Accordingly, there are 128 distinct writing tasks and the second version of EFCAMDAT contains 1,180,309 individual scripts written by 174,743 learners. For the current study, we worked with a subcorpus that consisted of 669,876 writings by 119,960 learners.[7] Texts varied in length from 20–40 words (lower levels) to 150–180 words (higher levels).

Task type

The 128 writing tasks were coded by a linguistically trained research assistant for task type. The categorisation was theoretically informed by earlier research into task genre and task type (e.g. Byrnes, Maxim and Norris 2010; Foster and Skehan 1996; Yoon and Polio 2017) and extended by emerging categories. This resulted in the following types: argumentation, description, instruction, narrative; and two emergent categories: comparison and list/form. The emergent category of 'comparison' was added, given that the instructional prompts of these tasks in EFCAMDAT explicitly asked to *compare* or *evaluate* two or more options (see example, Table 1). These tasks were deemed different from, for example, an argumentative task, where a comparison is not always needed. Similarly, the category 'list/form' was added, because some task prompts specifically prompted learners to

Table 1: Task types.

Task type	Example instruction/prompt
Argumentation	The manager [...] has asked you to write him an email outlining why you think you are suitable for the job. Write a convincing email as to why you are the perfect candidate.
Description	Read the notes on a survey. Write up the findings of the survey.
Instruction	Jane is lost and needs more directions to get to your house. Use the map to reply to her text message.
Narrative	You decide to enter a story-writing competition. Read about the details on the website and write your entry.
Comparison	Write a short report comparing the three robots for the web-based magazine.
List/form	You want to plan your future. […] Make a list of things you should do to improve your CV.

make a list or *fill in a form* and the language produced upon such a prompt consisted indeed of a (bullet-pointed) list or single-word items.

The first author checked all the codings and made adjustments as necessary. Accordingly, the corpus consists of twenty-one argumentative, sixty-nine descriptive, fourteen instructive, ten narrative and five comparative tasks, respectively, plus nine tasks asking for a list or form to be filled out. See Table 1 for example prompts for each task type.

Measures of complexity and data analysis

Based on the NLP methods computing a wide range of complexity measures originally developed by Vajjala and Meurers (2012, 2014) and extended for CTAP (Chen and Meurers 2016), we extracted 571 measures of linguistic complexity covering elaborateness and variedness of the linguistic system at all levels of modelling, language use and human sentence processing complexity.

Next, we performed a quantitative analysis of the corpus data and identified those complexity measures whose values differed across task types. More specifically, we first excluded the twenty complexity measures where the most frequent value, typically a 0, occupied 95% or more cases. For each remaining complexity measure, we then quantified the effect of task type on the complexity value by comparing two generalised additive mixed models (GAMM; Murakami 2016; Wieling 2018) built with the 'mgcv' package (Wood 2017) in R; one with task type as a predictor and one without. The observational unit was the individual writing in the corpus, and the baseline model predicted the values of the target complexity measure as a function of non-linear proficiency smooth, by-nationality random intercepts and by-task random intercepts. Proficiency was operationalised as the Englishtown unit (1–128), and the smooth term was specified with a thin plate regression spline.

We then built another model that included all the variables above and task type, which allowed the absolute value of each complexity measure to vary across task types. In order to quantify the importance of task type, the standard deviation (*SD*) of by-task random intercepts (i.e. remaining variability across tasks) was estimated using the getSD.gam function[8] (Wieling *et al*. 2014). Next, we examined the extent to which task type decreases the *SD* of the by-task random-intercepts by comparing the baseline model and the model with task type. This is akin to looking into the change in the coefficient of determination (R^2) in multiple regressions as predictors are added to the model (although we only focus on variability across tasks rather than variability across writings). For instance, the *SD* of by-task random intercepts of the 'type frequency of past tense verb forms' in the baseline model was 2.21, and it decreased to 1.68 when task type was entered into the model. This suggests that task type explained 24% ((2.21 – 1.68) / 2.21 = 0.24) of the variability across tasks. Finally, we excluded measures that correlated highly (i.e. $r \geq 0.80$) with another measure with larger explained variance because they presumably measure the same construct.

Qualitative explorations

After identifying the general patterns in the data quantitatively, we closely examined atypical tasks to better understand why some individual tasks might behave differently than the general pattern of their task type would predict. More specifically, we first computed the mean residual (i.e. the difference between the predicted and observed values) in each of the twenty complexity measures where task type explained the largest variance. Next, we identified five tasks with the smallest mean absolute residuals (i.e. tasks that behave as predicted by the model) and five tasks with the smallest/largest mean residuals, respectively (i.e. tasks with atypically low or high scores).

Results

To allow direct comparisons with earlier work on EFCAMDAT, we first review the global and specific measures explored by Alexopoulou *et al.* (2017) before we report on data-driven measures that emerged from the current analyses and focus on writings elicited by atypical tasks (see Appendix 1 for further illustrations). We thereby limit ourselves to measures of (morpho)syntactic complexity, in order to stay within the scope of this thematic issue.

Table 2: Morphosyntactic measures taken from Alexopoulou *et al*. (2017).

Global measures	Specific measures
• Mean length of T-unit (MLT)	• Number of complex noun phrases/clauses
• Mean length of clause (MLC)	• Number of *wh*-phrases per sentence
• Number of dependent clauses/T-unit	• Use of past tense verbs
	• 3rd person singular present tense verbs

Measures identified earlier for EFCAMDAT

Following Alexopoulou and associaties (2017), we explored three global and four specific measures, which are shown in Table 2.

Global measures

Figure 1(A) shows the developmental trajectory of the three global measures for the different task types from Englishtown levels 1 (A1) to 16 (C2). Figure 1(B) shows the relative value of complexity measures in each task type after partialling out proficiency and nationality effects. Table 3 shows the results of the pairwise comparison between the five task types in each

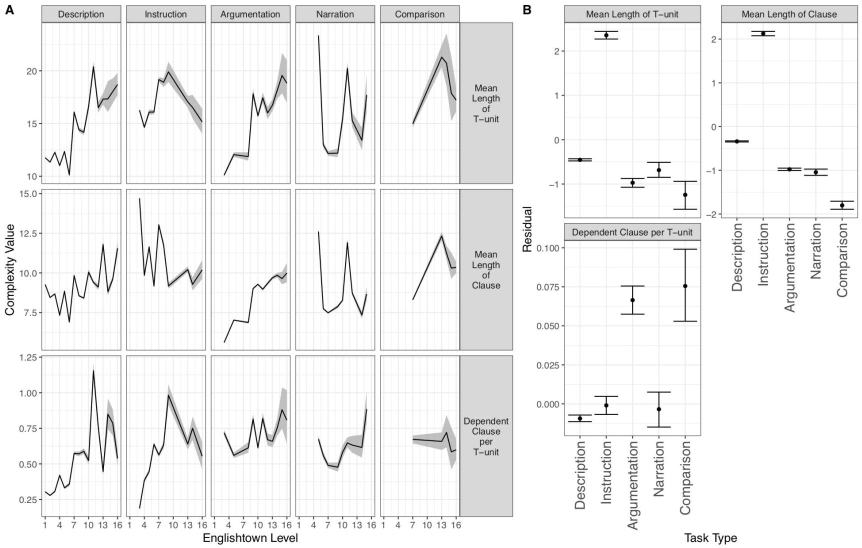

Figure 1: (A) Development of global complexity measures in each task type. The grey band represents bootstrap-based 95% confidence intervals of the mean. CEFR levels to EF mapping: A1 = 1–3, A2 = 4–6, B1 = 7–9, B2 = 10–12, C1 = 13–15, C2 = 16. (B) The mean residual of each task type in each global complexity measure and its bootstrap-based 95% confidence interval. The residuals were calculated based on the model predicting each complexity measure based on non-linear proficiency smooth and nationality as a random-effects factor.

Table 3: Pairwise difference between task types in each measure.

Measure	Difference between task types
Mean length of T-unit (MLT)	Instruction > Description
Mean length of clause (MLC)	Instruction > {Description, Argumentation}
Type frequency of base verb forms	{Instruction, Argumentation} > {Description, Narration, Comparison}
Type frequency of modals	{Instruction, Argumentation} > {Description, Narration, Comparison} Description > Narration
Type frequency of non-3rd person singular present verb forms	{Instruction, Description, Argumentation, Comparison} > Narration Argumentation > Description
Number of past tense verb forms	Narration > {Description, Instruction, Argumentation, Comparison}
Type frequency of past tense verb forms	Narration > {Description, Instruction, Argumentation, Comparison}
Number of *wh*-phrases	Narration > {Description, Argumentation, Comparison}
Number of 3rd person singular simple present verb forms	Comparison > Argumentation

Note: the pairwise comparison was performed by swapping the reference level of GAMMs and examining relevant contrasts between the reference-level task type and the other task types. The resulting *p* values were adjusted for multiple comparisons within each complexity measure using Hommel's (1988) method. No significant difference was observed between any pair of task types in 'dependent clause per T-unit', 'complex noun phrases per clause', '*wh*-phrases per sentence', 'number of yes/no questions', 'mean length of sentences in syllables', 'number of fragment T-units' and 'complex noun phrases per T-unit'.

of the measures we discuss below. In the table, the task type on the left marked a significantly higher value than that on the right. For instance, for mean length of clause, instructive tasks resulted in higher values than descriptive or argumentative tasks. However, there was no significant difference between descriptive and argumentative tasks, or between the task type pairs involving the task types not included there (i.e. narrative and comparison tasks).

Both Figure 1 and Table 3 indicate that, in MLT and MLC, instructive tasks score high while descriptive tasks remain comparatively low, even though there is growth over time. Instructive tasks elicit longer clauses than argumentative tasks as well. Given that comparisons only appear as a task type at the midpoint of the Englishtown levels (B1 level), the trajectory graph shows that they score relatively high on these measures (in contrast to the average, which is low), with a peak around the C1 level.

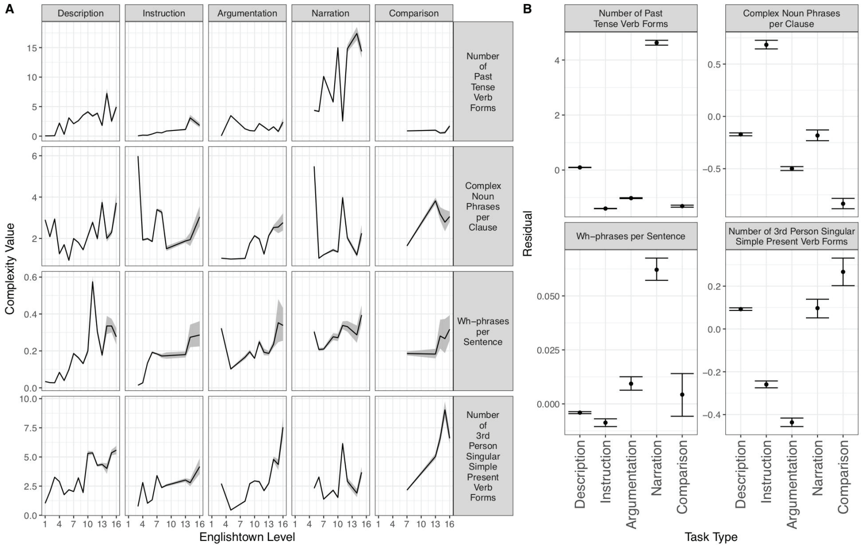

Figure 2: (A) Development of specific complexity measures in each task type. (B) The mean residual of each task type in each specific complexity measure. See the caption to Figure 1 for further details.

Narrative tasks remain low, but overall demonstrate a bumpy trajectory. Towards higher proficiency levels, the lines for the different task types seem to converge.

Looking at the ratio of dependent clauses, argumentative and comparative tasks on average elicit high scores, which seems to be fairly consistent over time. The GAMM, however, does not support any significant difference between task types. From the point at which they appear (level 6), narrative tasks also show higher scores while instructions and descriptions start low and show developmental growth with some peaks and dips at higher proficiency levels.

Specific measures

For specific measures, Figure 2(A) shows the developmental trajectories (A1 to C2), while Figure 2(B) gives the relative position of task types for each specific complexity measure (after controlling for proficiency and nationality effects). The result of pairwise comparisons between task types is presented in Table 3.

Top twenty data-driven measures identified in EFCAMDAT

From the large set of 571 measures available, we investigated for which measure task type would reduce the largest amount of variability across

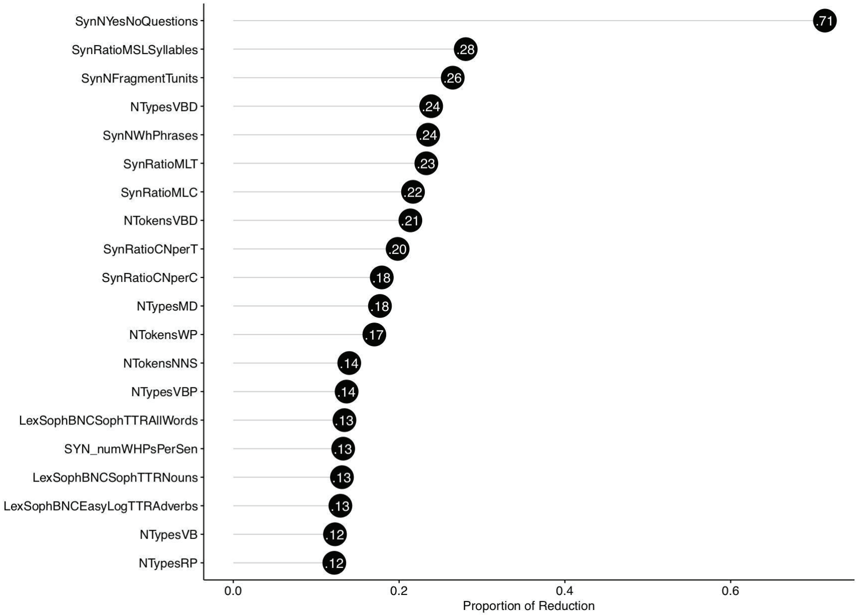

Figure 3: Top 20 complexity measures whose variability was decreased by task type. SynNYesNoQuestions = number of yes/no questions; SynRatioMSLSyllables = mean length of sentences in syllables; SynNFragmentTunits = number of fragment T-units; NTypesVBD = type frequency of past tense verb forms; SynNWhPhrases = number of *wh*-phrases; SynRatioMLT = mean length of T-units; SynRatioMLC = mean length of clauses; NTokensVBD = number of past tense verb forms; SynRatioCNperT = complex noun phrase per T-unit; SynRatioCNperC = complex noun phrase per clause; NTypesMD = type frequency of modals; NTokensWP = number of *wh*-pronouns; NTokensNNS = number of plural nouns; NTypesVBP = type frequency of non-third person singular present verb forms; LexSophBNCSophTTRAllWords = type-token ratio calculated on the words that are not in the most frequent 2000 words of the British National Corpus; SYN_numWHPsPerSen = number of *wh*-phrases per sentence; LexSophBNCSophTTRNouns = type-token ratio calculated on the nouns that are not in the most frequent 2000 words of British National Corpus; LexSophBNCEasyLogTTRAdverbs = log(type frequency) / log(token frequency), calculated on the adverbs that are in the most frequent 1000 lemmas in British National Corpus. NTypesVB = type frequency of base verb forms; NTypesRP = type frequency of particles. Sophisticated words = outside the most frequent 1000 words in British National Corpus; Easy words = within the most frequent 1000 words.

tasks (in comparison to the GAMM without task type). Figure 3 (see caption for details about each measure) visualises the variability reduction scores of the top twenty measures emerging from these data-driven explorations. Given the focus of this article, we will only discuss the (morpho) syntactic measures. Accordingly, task type explained over 20% of the variance in syntactic measures such as the 'number of yes/no questions', the 'mean length of sentences in syllables' and the 'number of *wh*-phrases' as well as the 'type frequency of past tense verb forms'. Close follow-up measures were 'complex noun phrase per T-unit', 'modal verb types', 'number of *wh*-pronouns' and 'type frequency of non-third person singular present verb forms' (i.e. first and second person forms).

Figure 4(B) shows the partial effects of task types on these top twenty measures, while Figure 4(A) provides their trajectory per task type across proficiency levels. In the following, we will highlight a couple of results from either or both parts of Figure 4. We refrain from commenting on those measures that are not (morpho)syntactic in nature, that yielded non-significant differences, or that are difficult to interpret without further (qualitative) explorations. The results of pairwise comparison between task types are shown in Table 3.

From the syntactic measures, the GAMM yields significant results only for the 'number of *wh*-phrases', where narratives score higher than descriptive, argumentative and comparison tasks – a reflection of the specific measure as a ratio per sentence described above. Only some specific descriptive tasks at the B2/C1 level peak above narratives on this measure. The 'number of *wh*-pronouns' shows a similar pattern.

From the verbal measures, 'type frequency of past tense verb forms' aligns with the token-based measure above, showing large fluctuations for narratives despite being always higher than other task types. In both type and token frequency, narrative tasks elicit significantly more frequent use of past tense verb forms than any other task type. In contrast, the number of types of non-third person present tense verbs identifies argumentative tasks in terms of average scores and peaks for instructions at the C1 level. The GAMM suggests that descriptive, instructive, argumentative and comparison tasks elicit a wider variety (i.e. higher type frequency) of 'non-third person singular present verb forms' than narrative tasks, and argumentative tasks elicit the higher type frequency of the feature than descriptive tasks ($t = 2.67$, $p = 0.008$). The 'number of modal verb types' shows steady growth over time and identifies argumentative tasks and, at higher levels, instructions. The GAMM indeed indicates that the 'type frequency of modals' is higher in argumentative and instructive tasks than in comparison, descriptive and narrative tasks. The 'number of types of verbs in their base form' is particularly high on average for argumentative, instructive and also descriptive tasks, which is reflected in the highest scores for instructions across all proficiency levels, while descriptions elicit these mainly at the C1 level. The GAMM shows that instructive and argumentative tasks lead to higher 'type frequency of base verb forms' than the other three task types.

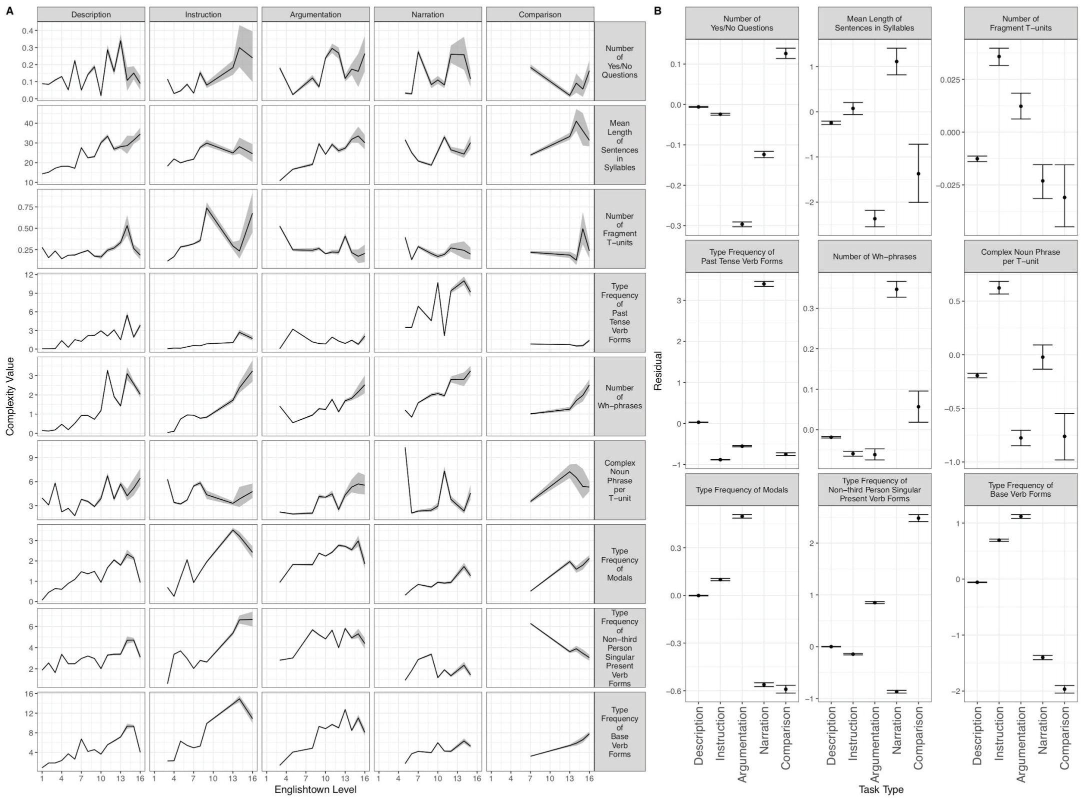

Figure 4: (A) Development of complexity measures in each task type. (B) The mean residual of each task type in each complexity measure. See caption to Figure 1 for details.

Qualitative analyses of atypical tasks

In general, task type explained substantial variance in our data set (e.g. use of past tense verbs – 24%; use of present tense verbs that were not third person forms – 14%). Interestingly, some individual tasks behaved very differently than expected based on their features. Below, we present example scripts from two instructive tasks at level 3/A1 (tasks 4 and 7) that stood out, because they reached much lower scores or much higher scores than their task type would predict from the specific measure of complex noun phrases. On average, instructions elicited high scores on this measure (see Figure 4(B)).

As can be seen, the instructive task on the left is a map task, asking for a text message instruction on how to get from A to B. It is unlikely that such an instruction would elicit complex noun phrases, unless the geographical targets on the map consisted of complex noun phrases (e.g. the yellow-fenced brick house). The task on the right instructs a friend via text message to buy food at a supermarket. The items to buy are (uncountable) nouns with a prenominal modifier giving the size or amount. This results in many complex noun phrases – as is evident from the example. In these cases, the specific task prompt elicited atypical language use. Further illustrative examples for dependent clauses and verb tense are provided in Appendix 1.

Example 1: Scripts of atypical instructive tasks for complex noun phrases (at A1 level).

Low complex noun phrase per T-unit:
Level 3 Unit 4; Writing ID = 162417;
Nationality = Brazilian
Hi Jane! Go straight ahead at Liverpool Road and turn left at Green Ave. My house is opposite the park, between the supermarket and the restaurant. Bye.

High complex noun phrase per T-unit:
Level 3 Unit 7; Writing ID = 344187;
Nationality = Mexican
Hello, can you buy me some ingredients to prepare a special dinner for my husband please. I need a piece of beef, one bottle of red wine, some potatoes, two cans of tomatoes, a packet of noodles and one piece of bread. Thanks.

Discussion

This paper aimed to complement earlier empirical work within ISLA and LCR by investigating the effects of instructional design, operationalised in this article as task type (e.g. description, argumentation), on (morpho) syntactic complexity in a large corpus of L2 writings, that is, EFCAMDAT. Starting with an extensive number of measures (> 500) derived from the corpus by means of NLP (Chen and Meurers 2016), we were able to identify (also unexpected) indices that typically align with a specific task type. Given that EFCAMDAT covers all CEFR levels, these effects could be situated within L2 development from A1 to C2. We discuss our findings in the following paragraphs.

Characteristics of task type

Drawing on the overview supplied in Table 3, descriptive tasks are characterised by a high number of simple present verb forms. This contrasts with the findings of Alexopoulou *et al.* (2017), who identified high use of past tense forms; this may have been due to the atypicality of the two specific tasks at levels 6 and 7 they looked at, whereas the present study covers a range of tasks from levels A1 to C2. Further, in light of the work carried out by Connor-Linton and Polio (2014), which examined a corpus of descriptive tasks, our findings also provide some new insights. In EFCAMDAT, the trajectories from A1 to C2 revealed growth in terms of sentence length and complexity; that is, learners seemed to develop over time when looking at syntactic measures (e.g. subordination) from the sixty-nine descriptive tasks. Interestingly, Bulté and Housen (2014) were not able to detect development using this same measure. It is possible that the different time frames (Bulté and Housen looked at changes within a four-month period) explain the discrepancies, indicating that not only task-related factors but also other contextual variables such as time covered by a developmental corpus, may explain differing results. Importantly, descriptions scored low – often significantly lower than other task types – on most measures. While growth over time indicates learning progress, the low scores of descriptive tasks may imply a predominance of tasks targeting less complex language, given that the vast majority of EFCAMDAT tasks are descriptive. As Lee and Polio (2017) put it: 'If students keep to writing assignments that elicit simple language, they may not have an opportunity to develop their language.' (2017:311). Given that two-thirds of the descriptive tasks in EFCAMDAT were at lower levels (1–8) of the online course, while at higher levels other task types were used more often, other task types seem to fulfil the role of eliciting more advanced structures.

For example, instructive tasks in EFCAMDAT are characterised by long T-units and clauses, and high frequencies of verbs in their base form. Instructions often take a list-like form (do A, then B, then C; cf. Example 1 above) separated by commas, which might explain the long syntactic units. Tasks classified as instructions in EFCAMDAT included those where learners were asked to provide some advice (for example, how to avoid stress), and several online lessons leading up to those tasks highlight formal language use. Prompts and writings typically consist of sentences like 'you can do X, you may try Y', which might explain the verbal measure.

The qualitative explorations related to specific tasks (see the examples above and in Appendix 1) show that even though task type effects seem to be quite strong in EFCAMDAT, individual tasks can still trigger the use of atypical forms, often due to a specific characteristic of the instructional prompt. These qualitative data thus strengthen our call for research that takes into account task effects and other design features when compiling, analysing and interpreting results based on large learner corpora.

In line with Yang *et al.* (2015), our data showed that argumentative tasks are not associated with complex noun phrases. Argumentation apparently triggers the use of modal and base form verbs, and triggers a higher frequency of non-third person singular present tense verbs (i.e. verbs in the first and second person). It seems that argumentative tasks in EFCAMDAT are most successful in eliciting a variety of language forms. This concurs with Crossley and McNamara's (2014) conclusion that tasks that require 'persuasive arguments, or integrating outside information into an essay may be better suited to evaluate developing syntactic proficiency in L2 writers than descriptive writing' (2014:78).

In the current analysis, in particular, narratives complement the other tasks as they are typically set in the past and trigger high numbers of *wh*-phrases. This latter finding may stem from the need to make reference to people and places, which would confirm the findings of Alexopoulou and co-workers (2017). Apart from the *wh*-phrases, the narratives showed a large fluctuation in syntactic indices. This might explain why we, unlike Lu (2011), could not demonstrate many significant differences between argumentative and narrative tasks based on syntactic units. Another reason could be that given the nature of EFCAMDAT we were unable to extract syntactic indices based on co-ordination using NLP, which Lu (2011) and Norris and Ortega (2009) identified as important indices.

Finally, comparisons score high on quite a few measures, but fail to significantly distinguish themselves from other task types with the exception of present tense verb forms (non-third person singular) which seem to identify these tasks. One reason might be that written responses to this task type often included descriptive and argumentative language. Similarly, argumentative tasks regularly included comparative aspects. Given that we identified only five comparison tasks in EFCAMDAT, future explorations could consider incorporating these tasks under the category of argumentation.

Trajectories over time and proficiency: development and fluctuation

When looking at the developmental trajectories from A1 to C2, most measures show that task effects are stronger at initial stages, while from level B1/B2 (level 7/8) onwards lines representing task types seem to converge. This pattern is most clearly visible for global measures (Figure 1(A)), but is also apparent on some specific measures (see 'complex noun phrases per clause', Figure 2(A)). The convergence suggests that at lower proficiency levels, the task-related factors tend to have a strong influence on what language learners will use, while towards higher levels of proficiency learners are less restricted, presumably thanks to their growing linguistic resources. In other words, learners at higher proficiency levels apparently have a wider and less task-specific knowledge, and they therefore seem to be able to demonstrate more variable language for a given task. Indeed, Yang *et al.* (2015) also conclude that 'one essence of linguistic development and L2 writing development is seen in learners' ability to stretch their linguistic repertoire and achieve linguistic complexity in ways not constrained by the task or the topic, shown in the greater linguistic resources and means to attain greater diversity and sophistication in language use' (2015:64). It is likely that at initial stages writers demonstrate the use of a specific feature because it is the target of the lesson or because they could lift a rather formulaic expression from the prompt (e.g. dependent clauses in the descriptive task at level 4, unit 7; cf. Appendix 1, Example 2). From a pedagogic perspective, providing example structures often holds the function of scaffolding, giving learners the possibility to practise and use a form they would not come up with based on their own knowledge and understanding. The language given in the instructional prompt can serve as an example and therefore can determine to a certain extent the output produced by L2 writers, in particular at lower levels of proficiency. For research and teaching practice, this implies that at beginner levels of language learning one needs to be very cautious about task and prompt effects – as was acknowledged by Vyatkina (2012), based on her corpus work of early stages of L2-German.

Some unexpected measures emerged as indicators for task type differences throughout the whole corpus, independent of proficiency. For example, Figure 4 shows that 'type frequency of modals' was consistently high for argumentative tasks and the developmental line never crossed the lines for narratives and descriptives, suggesting it might be an ideal index to distinguish these task types.

Finally, some measures were very difficult to interpret, given that they are characterised by large fluctuations, be this within a single task type or across task types (e.g. most sentence length measures).

Conclusions

In this paper, we present findings based on more than 650,000 writings by over 100,000 learners exploring the effects of instructional design operationalised as task type effects on L2 writing. Our results lend support to earlier experimental and corpus-based studies in confirming measures that are typically associated with a certain task type (e.g. past tense forms for narrative tasks). In addition, through data-driven examination of 571 measures we were able to identify unique task effects. For example, argumentative tasks naturally elicited non-third-person singular present tense verb forms. In terms of Loschky and Bley-Vroman's (1993) work, our analyses extracted those linguistic means that seem to be naturally – or even essentially – emerging in response to a specific task type. Further research is needed to substantiate these findings. The results can serve as a first indication as to what task type might be used in the classroom to elicit a specific structure.

Given the rich data set EFCAMDAT and other large-scale learner corpora provide, there are many different avenues for future work. The current study restricted itself by looking into task type effects on (morpho) syntactic measures only. However, many other task design features (e.g. task complexity, formality, input provided by task prompt) and other measures that emerged for a given task type (e.g. a large diversity of easy adverbs for argumentative tasks) remain to be investigated in greater detail. Similarly, in this paper we statistically controlled for proficiency and between-nationality differences, while future work might specifically target these and other variables which potentially distinguish learner groups. In particular, exploring interrelationships between different task design features and learner variables could be enlightening. Tracing individual learners' writing development from A1 to C2 would be a fascinating endeavour, especially given that the cross-sectional developmental patterns of the kind we looked at in this study can conceal individual learning trajectories (Geertzen, Alexopoulou and Korhonen 2014; Murakami and Alexopoulou 2016).

From a pedagogic perspective, our findings call for language courses that provide learners with writing assignments targeting a wide variety of task types. We demonstrated that each task type has the potential to trigger the use of specific target structures – but at the same time carries the risk of not eliciting other structures. Using mainly descriptive tasks might limit the development of subordination; relying on argumentative essays could hinder the practice of the past tense. Importantly, learning progress requires a good variety of task types at all stages of development.

In the end, it is the diversity of instructional design which ensures that L2 writing may fulfil its full potential as a site for L2 practice and development (Manchón 2011). Not least, task type diversity presumably makes learning more motivating too.

About the authors

Marije Michel is Associate Professor at Groningen University (NL) and Lecturer at Lancaster University (UK). Her research and teaching focuses on second language acquisition, assessment and task-based pedagogy. Her most recent work uses eye-tracking and key-stroke logging to investigate (online) writing processes and alignment.

Akira Murakami is a Birmingham Fellow at the University of Birmingham. His primary research interests include second language acquisition, corpus linguistics, and quantitative data analysis. Prior to joining Birmingham in 2018, he was a post-doctoral researcher at the Universities of Birmingham, Cambridge, and Tübingen.

Theodora Alexopoulou is Principal Research Associate at the Theoretical and Applied Linguistics Section, Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge.

Detmar Meurers is a Professor of Computational Linguistics at the University of Tübingen, Germany. He previously worked as an Associate Professor in the Department of Linguistics at The Ohio State University (USA). As head of his current ICALL-Research.com group, his work focuses on Intelligent Computer-Assisted Language Learning, and computational linguistic methods in second language acquisition research and language teaching.

Acknowledgements

This research was supported by an Aspasia Diversity Fund of Utrecht University awarded to the first author. We would like to thank EF Education First for the sponsorship of the Research Lab for Applied Language Learning at the University of Cambridge which has enabled this research and, specifically, Dr Rachel Baker, Director of Applied Linguistics Research at EF Education First for providing us with information on the task prompts. We also thank Laura Brand for her help with coding.

Notes

[1] http://cohmetrix.com

[2] http://personal.psu.edu/xxl13/downloads/l2sca.html or http://aihaiyang.com/software/l2sca

[3] https://linguisticanalysistools.org/taassc.html

[4] http://ctapweb.com

[5] https://languagelink.let.uu.nl/tscan

[6] This description applies to an earlier version of curriculum and teaching materials which was withdrawn in 2013 and is no longer in use.

[7] For our study we have used a pre-release version of EFCAMDAT, which needed some additional cleaning and checks to ensure matching between topic IDs and writings. On request, we can share the list of writing IDs we used.

[8] http://openscience.uni-leipzig.de/index.php/mr2/article/view/41

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. https://doi.org/10.1075/ijlcr.1.1.04ale

Alexopoulou, T., Michel, M., Murakami, A. and Meurers, D. (2017) Task effects on linguistic complexity and accuracy: a large-scale learner corpus analysis employing natural language processing techniques. *Language Learning* 67(s1): 180–208. https://doi.org/10.1111/lang.12232

Baralt, M., Gilabert, R. and Robinson, P. (eds) (2014) *Task Sequencing and Instructed Second Language Learning*. London: Bloomsbury Publishing.

Benson, P. and Reinders, H. (eds) (2011) *Beyond the Language Classroom*. London: Palgrave Macmillan. https://doi.org/10.1057/9780230306790

Biber, D., Gray, B and Staples, S. (2014) Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics* 37(5): 639–68. https://doi.org/10.1093/applin/amu059

Brezina, V. and Flowerdew, L. (eds) (2017) *Learner Corpus Research: New Perspectives and Applications*. London: Bloomsbury Academic.

Bulté, B. and Housen, A. (2012) Defining and operationalising L2 complexity. In A. Housen, F. Kuiken and I. Vedder (eds) *Dimensions of L2 Performance and Proficiency* 21–46. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.32.02bul

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

Bulté, B. and Housen, A. (2018) Syntactic complexity in L2 writing: individual pathways and emerging group trends. *International Journal of Applied Linguistics* 28(1): 147–64. https://doi.org/10.1111/ijal.12196

Byrnes, H., Maxim, H. H., & Norris, J. M. (eds) (2010) Realizing advanced foreign language writing development in collegiate education: curricular design, pedagogy, assessment [Special issue]. *Modern Language Journal* 94(s1). https://doi.org/10.1111/j.1540-4781.2010.01137.x

Chen, X. and Meurers, D. (2016) CTAP: a web-based tool supporting automatic complexity analysis. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity* 113–19. Osaka, Japan. http://aclweb.org/anthology/W16-4113

Connor-Linton, J. and Polio, C. (2014) Comparing perspectives on L2 writing: multiple analyses of a common corpus. *Journal of Second Language Writing* 26: 1–9. https://doi.org/10.1016/j.jslw.2014.09.002

Crossley, S. A. and McNamara, D. S. (2014) Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Jour-*

nal of Second Language Writing 26: 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Cumming, A. (1990) Expertise in evaluating second language compositions. *Language Testing* 7(1): 31–51. https://doi.org/10.1177/026553229000700104

Ferris, D. R. (1994) Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly* 28: 414–20. https://doi.org/10.2307/3587446

Foster, P. and Skehan, P. (1996) The influence of planning and task type on second language performance. *Studies in Second Language Acquisition* 18: 299–323. https://doi.org/10.1017/S0272263100015047

Frear, M. W. and Bitchener, J. (2015) The effects of cognitive task complexity on writing complexity. *Journal of Second Language Writing* 20: 45–57. https://doi.org/10.1016/j.jslw.2015.08.009

Gablasova, D., Brezina, V. and McEnery, T. (2017a) Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning* 67(s1): 130–54. https://doi.org/10.1111/lang.12226

Gablasova, D., Brezina, V., McEnery, T. and Boyd, E. (2017b) Epistemic stance in spoken L2 English: the effect of task and speaker style. *Applied Linguistics* 38(5): 613–37. https://doi.org/10.1093/applin/amv055

Geertzen, J., Alexopoulou, T. and Korhonen, A. (2014) Automatic linguistic annotation of large scale L2 databases: the EF-Cambridge Open Language Database (EFCAMDAT). *Proceedings of the 31st Second Language Research Forum (SLRF)*. Pittsburgh, PA: Cascadilla Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. and Cai, Z. (2004) Coh-Metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers* 36: 193–202. https://doi.org/10.3758/BF03195564

Granger, S., Gilquin, G. and Meunier, F. (2015) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Hinkel, E. (2009) The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics* 41(4): 667–83. https://doi.org/10.1016/j.pragma.2008.09.029

Hommel, G. (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika 75*(2): 383–6. https://doi.org/10.1093/biomet/75.2.383

Kormos, J. (2011) Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing* 20(2): 148–61. https://doi.org/10.1016/j.jslw.2011.02.001

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. https://doi.org/10.1016/j.jslw.2007.08.003

Kyle, K. (2016) Measuring syntactic development in L2 writing: fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. Doctoral dissertation, Georgia State University. Retrieved on 19 September 2019 from http://scholarworks.gsu.edu/alesl_diss/35

Lee, C. and Polio, J. (2017) Written language learning. In S. Loewen and M. Sato (eds) *The Routledge Handbook of Instructed Second Language Acquisition* 299–318. New York: Routledge. https://doi.org/10.1080/15475441.2016.1180983

Loewen, S. (2015) *Introduction to Instructed Second Language Acquisition*. New York: Routledge.

Loewen, S. and Sato, M. (2017) *The Routledge Handbook of Instructed Second Language*

Acquisition. New York: Routledge. https://doi.org/10.4324/9781315676968

Loschky, L. and Bley-Vroman, R. (1993) Grammar and task-based methodology. In G. Crookes and S. Gass (eds) *Tasks and Language Learning: Integrating Theory and Practice* 123–67. Bristol: Multilingual Matters.

Lu, X. (2010) Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4): 474–96. https://doi.org/10.1075/ijcl.15.4.02lu

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

Manchón, R. M. (2011) *Learning-to-Write and Writing-to-Learn in an Additional Language*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.31

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: 381–92. https://doi.org/10.3758/BRM.42.2.381

Meurers, D. (2015) Learner corpora and natural language processing. In S. Granger, G. Gilquin and F. Meunier (eds) *The Cambridge Handbook of Learner Corpus Research* 537–66. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.024

Meurers, D. and Dickinson, M. (2017) Evidence and interpretation in language learning research: opportunities for collaboration with computational linguistics. *Language Learning* 67(2): 67–96. https://doi.org/10.1111/lang.12233

Michel, M. (2017) Complexity, accuracy and fluency in L2 production. In S. Loewen and M. Sato (eds) *Routledge Handbook of Instructed Second Language Acquisition* 50–68. New York: Routledge. https://doi.org/10.4324/9781315676968-4

Murakami, A. (2016) Modeling systematicity and individuality in nonlinear second language development: the case of English grammatical morphemes. *Language Learning* 66(4): 834–71. https://doi.org/10.1111/lang.12166

Murakami, A. and Alexopoulou, T. (2016) Longitudinal L2 development of the English article in individual learners. In A. Papafragou, D. Grodner, D. Mirman and J. Trueswell (eds) *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* 1050–5. Austin, TX: Cognitive Science Society.

Nassaji, H. (2016) Interactional feedback in second language teaching and learning: a synthesis and analysis of current research. *Language Teaching Research* 20(4): 535–62. https://doi.org/10.1177/1362168816644940

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): 555–78. https://doi.org/10.1093/applin/amp044

Ott, N., Ziai, R. and Meurers, D. (2012) Creation and analysis of a reading comprehension exercise corpus: towards evaluating meaning in context. In T. Schmidt and K. Wörner (eds) *Multilingual Corpora and Multilingual Corpus Analysis* 47–69. Hamburg Studies in Multilingualism (HSM). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/hsm.14.05ott

Pallotti, G. (2015) A simple view of linguistic complexity. *Second Language Research* 31(1): 117–34. https://doi.org/10.1177/0267658314536435

Pander Maat, H., Kraf, R., van den Bosch, A. P. J., Dekker, N., Gompel, M. V., Kleijn, S. D., ... and Sloot, K. (2014) T-Scan: a new tool for analyzing Dutch text. *Computational Lin-*

guistics in the Netherlands (4):53–74. Retrieved on 19 September 2019 from http://hdl.handle.net/2066/134833

Polio, C. and Park, J.-H. (2016) Language development in second language writing. In R. M. Manchón and P. Matsuda (eds) *Handbook of Second and Foreign Language Writing* 287–306. New York: Routledge. https://doi.org/10.1515/9781614511335-016

Polio, C. and Yoon, H.-J. (2018) The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics* 28(1): 165–88. https://doi.org/10.1111/ijal.12200

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): 19–48. https://doi.org/10.1558/cj.v33i1.26543

Rebuschat, P., Meurers, D. and McEnery, T. (2017) Language learning research at the intersection of experimental, computational, and corpus-based approaches. *Language Learning,* 67(s1): 6–13. https://doi.org/10.1111/lang.12243

Robinson, P. (2001) Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics* 22(1): 27–57. https://doi.org/10.1093/applin/22.1.27

Sato M. and Loewen, S. (2019) *Evidence-based Second Language Pedagogy: A Collection of Instructed Second Language Acquisition Studies.* New York: Routledge. https://doi.org/10.4324/9781351190558

Skehan, P. (2009) Modelling second language performance: integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics* 30(4): 510–32. https://doi.org/10.1093/applin/amp047

Swain, M. and Lapkin, S. (2002) Talking it through: two French immersion learners' response to reformulation. *International Journal of Educational Research* 37(3–4): 285–304. https://doi.org/10.1016/S0883-0355(03)00006-5

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. https://doi.org/10.1075/ijlcr.1.1.03tra

Tracy-Ventura, N. and Paquot, M. (eds) (in preparation) *The Routledge Handbook of Second Language Acquisition and Corpora.* New York: Routledge.

Vajjala, S. and Meurers, D. (2012) On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)* 163–73. Montréal, Canada: ACL. Retrieved on 19 September 2019 from http://aclweb.org/anthology/W12-2019.pdf

Vajjala, S. and Meurers, D. (2014) Readability assessment for text simplification: from analysing documents to identifying sentential simplifications. *ITL – International Journal of Applied Linguistics* 165(2): 194–222. https://doi.org/10.1075/itl.165.2.04vaj

Vyatkina, N. (2012) The development of second language writing complexity in groups and individuals: a longitudinal learner corpus study. *Modern Language Journal* 96: 576–98. https://doi.org/10.1111/j.1540-4781.2012.01401.x

Vyatkina, N., Hirschmann, H. and Golcher, F. (2015) Syntactic modification at early stages of L2 German writing development: a longitudinal learner corpus study. *Journal of Second Language Writing* 29: 28–50. https://doi.org/10.1016/j.jslw.2015.06.006

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* 84(2): 171–84. https://doi.org/10.1111/0026-7902.00060

Weigle, S. C. (2002) *Assessing Writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wieling, M. (2018) Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70: 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wieling, M., Montemagni, S., Nerbonne, J. and Baayen, R. H. (2014) Lexical differences between Tuscan dialects and standard Italian: accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language* 90: 669–92. https://doi.org/10.1353/lan.2014.0064

Williams, J. (2012) The potential role(s) of writing in second language development. *Journal of Second Language Writing* 21(4): 321–31. https://doi.org/10.1016/j.jslw.2012.09.007

Wolfe-Quintero, K., Inagaki, S. and Kim, H.-Y. (1998) *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Manoa, Hawaii: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Wood, S. N. (2017) *Generalized Additive Models: An Introduction with* R (2nd edn). Boca Raton, FL: Chapman and Hall/CRC.

Yang, W., Lu, X. and Weigle, S. C. (2015) Different topics, different discourse: relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing* 28: 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

Yoon, H.-J. and Polio, C. (2017) The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly* 51(2): 275–301. https://doi.org/10.1002/tesq.296

Appendix 1: Illustrations of qualitative differences

A. Dependent clauses

At level 4, tasks 5 and 7 represent extreme points of dependent clause use. Both are descriptive tasks, so a low level of subordination would be expected. Task 4.5 requires learners to describe a photo of their family. Consequently, learners produce short descriptive sentences with the copula ‘be’ and possessive ‘have’ as the most common verbs, which do not take dependent clauses. Furthermore, there is no complex event structure to require adverbial clauses (temporal, causation, etc.) or complex descriptions of referents that would lead to more relative clauses.

In contrast, task 4.7 asks writers to complain to their flatmate about a list of chores she failed to do. As a whole, the texts are descriptive. However, the task prompt instructs writers to start their complaint with the phrases ‘I’m very angry because I did most of the chores this week. Let me tell you what I did.’ Most writers indeed used these exact nineteen words as a beginning. The rest of the 50–70-word text (word limit provided in instructional prompt) is covered by one or two sentences with a list of chores separated by commas. Given that out of three or four sentences, the two initial ones contain a subordinate clause each, a high score for subordination emerges.

Example 2: Scripts of atypical descriptive tasks for number of dependent clauses (at A2 level).

Low number of dependent clauses:
Level 4 Unit 5; Writing ID = 319869;
Nationality = German
This is my family. My husband is called Peter. He has short grey hair and blue eyes. He is wearing a blue shirt and black pants. Our son is called Ole Jens. He is very tall and thin.
Writing ID = 712568;
Nationality = Chinese
My mother has long straight black hair and big eyes. In her part time, she often dances in a beautiful dress. I have long curly brown hair. I look like my mother …

High number of dependent clauses:
Level 4 Unit 7; Writing ID = 580314;
Nationality = Mexican
I’m very angry because I did most of the chores this week. Let me tell you what I did. On Monday I washed the dishes and did the ironing, on Tuesday I washed the dishes and made the beds, on Wednesday I washed the dishes again, the next day I washed the dishes, made dinner, made the beds and paid the bills. On Friday I made the beds, on Saturday I swept and mopped the floor, washed the dishes, made the bed and did the shopping …

B. Verb tense

At the much higher level 15, tasks 3 and 7 are both descriptive tasks, which are characterised in EFCAMDAT by average scores on past tense and non-third person present tense verb forms. Interestingly, task 15.3 scores high on past tense verbs while 15.7 scores very low. The opposite pattern holds for present tense verbs (non-third person). Taking a closer look at the task instruction/prompt and writing samples explains these extreme behaviours: task 15.3 is asking learners to put themselves in the shoes of a counsellor, listen to a client and then describe her situation before and after treatment. The use of the past is important in order to accurately reflect the sequence of events, while the formal register of the task excludes use of the 'historic/narrative present' and necessitates the systematic past. Perhaps unlike other tasks set in the past, all events here are in sequence, rather than overlapping or simultaneous, so imperfective forms would be less appropriate. Similarly, the reason why task 15.7 is different is due to the task prompt asking learners to write about the ways in which life will be different in twenty years' time, which, unsurprisingly, results in high numbers of future verb forms.

Example 3: Scripts of atypical descriptive tasks for number of past tense verbs (at C1 level).

High use of past tense verbs:
Level 15 Unit 3; Writing ID = 505242;
Nationality = Korean
The client worked, and she had a toddler child. She was often busy, and she got annoyed at her child easily when she interrupted her because of the lot of work she had. She ended up staying up late, which made her snappy at her daughter, and she felt guilty …

Low use of past tense verbs:
Level 15 Unit 7; Writing ID = 410666;
Nationality = Italian
I think that the people's lives will improve in 20 years from now. I trust that strong and significant improvements will concern the transport system and the environment. We will have better public transportation and we will use new and greener forms of power …

Example 4: Scripts of atypical descriptive tasks for non-third person present tense verbs (at C1 level).

Low use of non-3rd person present:
Level 13 Unit 6; Writing ID = 81636;
Nationality = Brazilian
European Collection brochures: 1. The Impressionist Wing: Impressionism began in Paris as a reaction to traditional and rigid style of painting. Painters preferred to painting outside where they could see the impact and effect of the natural light on objects. The painting The Road Bridge at Argenteuil by Claude Monet was painted …

High use of non-3rd person present:
Level 13 Unit 9; Writing ID = 184175;
Nationality = Russian
My needs and goals according to Maslow's hierarchy. Physiological needs: I have enough fresh air to breathe, I do not go hungry or thirsty but I would prefer to diversify my food. I really suffer from cold in winter and heat in summer, so I would prefer another place to live in …

Tasks 6 and 9 stand out at level 13. Even though they are descriptive tasks, their use of non-third person present tense verb forms is low and high, respectively.

Taking a closer look reveals that task 13.6 asks writers to formulate a brochure text for a museum. Although the task is in the present, there is natural reference to the past when information is given about the history of artefacts and exhibits, as shown in the sample script. Task 13.9 asks learners to evaluate their needs and goals, which are typically expressed with habitual present tense verb forms in the first person.