1

The Effect of Task and Topic on Opportunity of Use in Learner Corpora

Andrew Caines and Paula Buttery

1 Introduction

Only a little attention has been paid in the field of LCR to the effect of situational variables such as document length, task and topic, and yet their true effect needs to be fully understood before strong conclusions can be made about, for example, proficiency-level profiling, learner progress and so on. As Tummers and colleagues (Tummers, Speelman and Geeraerts 2014: 482) state, 'The phenomenon of confounding variables … is hardly ever explicitly raised in (corpus) linguistics', even though their importance is acknowledged in other disciplines. However, there is a growing interest in this area (e.g. Ädel 2015; Biber, Gray and Staples 2016; Gries 2003; Hinkel 2009; Khabbazbashi 2017; Kobayashi and Abe 2014), one we pursue here by investigating how various linguistic features are affected by task and by topic – the subject matter and structure of a text, from the high level such as 'business' or 'society', to the fine-grained (e.g. 'write marketing strategy', 'what public transport is like in my hometown') – concluding that this is a factor that needs to be fully understood and controlled for in LCR and, by extension, language teaching and assessment. Finding that task and topic are inextricably linked, we refer to both variables in a hybrid way from now on, as the 'task-topic' of a given document.

In LCR the task and topic of any given text are usually dictated by a 'prompt' – a question or statement that prompts a response from the learner. This is typical of learner essay collections, such as the International Corpus of Learner English (ICLE) (Granger et al. 2009), the Longman Learners' Corpus[1] and the Cambridge Learner Corpus (CLC; Nicholls 2003). As a consequence of how corpora such as these tend to be collected in an ongoing process of accumulation, they are not typically balanced in terms of tasks and topics.

For instance, the CLC has been built up over many years from the whole suite of exams set by Cambridge English Language Assessment.[2] Due to the exam pathways designed by Cambridge English, their business English certificates (BEC) do not begin until CEFR[3] level B1 and go up to level C1, impacting the CLC in the sense that levels A1, A2 and C2 are devoid of business English content and are instead made up of general English essays. An initial topic distinction can be made therefore between 'business English' and 'non-business English'. If topic is not controlled for even at this basic level, apparent linguistic progression from level to level may in fact be confounded by the different constitutions of CEFR level sub-corpora.

It is understandable that the CLC has been collected in this way: its rapid and ongoing accumulation has been an invaluable resource for LCR. Nor is it the only such heterogeneous learner corpus. It is the purpose of this chapter to demonstrate that task and topic do affect language use, and that they are variables which need to be controlled for in LCR, along with other external variables such as document length. We refer to a concept we introduced in previous work, 'opportunity of use' – the opportunity the learner is afforded to use a linguistic feature, whether a lexical item, particular construction or discourse structure (Buttery and Caines 2012b). For instance, we showed that adverb use does not have a linear relationship with document length. That is, native speakers use disproportionately few adverbs in shorter documents compared to longer documents, and since lower-proficiency learners tend to write shorter documents than higher proficiency learners, they do not have the same opportunity to demonstrate use of adverbs. This needs to be taken into consideration when comparing learners for research or assessment purposes.

Hence opportunity of use should be controlled for a fair comparison between proficiency levels, between native speakers and learners, or among individual learners. In this chapter, we discuss the following four areas: (i) a task-topic taxonomy for learner essays, (ii) lexico-syntactic usage differences across task-topic types, (iii) unsupervised identification of task-topic-type clusters in the scenario where the prompts are unknown to the researcher and (iv) implications of task-topic effect for researchers, assessors and teachers.

2 Previous work

Previous relevant work in LCR has mainly focused on the effect of task on language features, with only incidental mention of topic. A representative example of this would be Newton and Kennedy (1996) who report on a 29,000-

word corpus constructed from four adult learners of English, at what is described as 'pre-university' proficiency. The learners were presented with two tasks. The first of these was a spatial shared information map-task regarding the optimal layout of a zoo: each learner was provided with a partial layout and then asked to complete a map of the entire zoo via spoken communication only. The second task involved sharing non-spatial domain-specific information about a medical dilemma. Here learners were required to reach consensus regarding the order in which patients should receive treatment: each learner was again provided with only partial information and via conversation expected to construct a priority list for patients awaiting surgery.

The study focused on the effects of the different task types (spatial vs. non-spatial) on the language that was used. For instance, more subordinating conjunctions were used in the non-spatial information task because of the need to argue a case with relationships of cause and effect, condition, result, purpose. The spatial zoo task elicited a greater number of prepositional phrases than the non-spatial medical task because of the need to verify location to create the map.

Another area of relevant work is probabilistic topic modelling, which is used for discovering topics in a collection of texts based on lexical features. This tends to proceed via modelling techniques such as 'latent Dirichlet allocation' in which a set of documents are considered to be distributed over a fixed vocabulary (Blei 2012). In this approach, identifiable topics emerge through interpretation of the 'topic' wordlists generated by the model. It is up to the researcher to generalize over the wordlists and assign each one a label.

For example, the most frequently occurring topic in a corpus of 17,000 *Science* articles features the terms such as 'human, genome, dna, genetic, genes' – and we might therefore label this topic 'genetics'; the second most frequent topic contains words such as 'evolutionary, species, organisms, life, origin' – we can label this topic 'evolution'; and so on. Topic wordlists are not always so easy to label, but the fact that many are as coherent as shown in these examples is a natural consequence of 'the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of [latent Dirichlet allocation]' (Blei 2012: 84). But note that in itself, this observation reflects the fact that lexical choice, at least, is statistically skewed by topic, else the linear discriminant analysis (LDA) technique would not work.

Initially we present three case studies (sections 4–6) in which we opt not to employ a bottom-up topic modelling approach in this work. Our interest lies not so much in the discovery of topics and their frequency, but rather in the effect of predefined topic classes on linguistic features, thereby adopting a top-down view of topic and preparing a corpus of labelled texts of even frequency. These studies are based on the assumption that the prompts linked to the essays in a corpus are known or can be obtained. However, there is an alternative scenario in which the prompts are unknown, and so we do discuss how clustering might be used to group prompts in a corpus in a machine learning procedure, demonstrating that essays are readily clustered into something like the task-topic groupings we define.

3 Cambridge Learner Corpus

We use the CLC for this study, a collection of essays[4] written by students from around the world sitting Cambridge English exams. We were provided with the prompts for a subset of exams taken in 2009.[5] Using these, we could identify the expected task-topic for a selection of essays written that year. Having inspected the prompts for the essays in the corpus we settled upon a set of six labels to evenly represent the range of tasks and topics presented by the prompts that year: 'administrative', 'autobiographical', 'narrative', 'professional', 'society', 'transactional'. We experimented with higher and lower level label sets, but settled upon this set on the grounds that they formed quite large groups while being at a sufficiently granular level in descriptive terms. For example, we labelled the following prompts 'autobiographical' (i), 'narrative' (ii) and 'society' (iii):

(1) This is part of a letter you receive from your new penfriend, Jenna.

I've got one close friend who I spend a lot of time with. What about you? Tell me about your friends. How important are they to you?

- Now write a letter to Jenna about your friends.
- Write your letter in about 100 words on your answer sheet.

(2) Your English teacher has asked you to write a story.

- Your story must begin with this sentence:

It was getting dark and I was completely lost.

- Write your story in about 100 words on your answer sheet.

(3) Your class is doing a project about education. Your teacher asks you to write about education in a country you know. Write about:

- what you think is good and bad about education in that country
- how you think education in that country will change in the future.
- Write about 100 words.

From our sample of 84 prompts from the year 2009, we were able to match and assign task-topic labels to 6,953 essays written by 4,784 individual students, giving us a sub-corpus we hereafter refer to as 'CLC_2009'. B1 is the largest proficiency subset of CLC_2009 and so we limit ourselves to essays at this level only, to control the effect of proficiency as far as practical (a CEFR level still represents a wide range of proficiency), and moreover to constrain document length – a variable we have previously recommended should be controlled (Buttery and Caines 2012b).

Since in this project the document is the target entity in question, we needed a corpus with a balanced number of essays for each of our task-topic labels. We considered balancing task-topic sub-corpora by number of tokens, but decided this would have the unsatisfactory consequence of (i) introducing incomplete documents where the requisite token count had been reached before a document break, and (ii) introducing an unwanted label bias in the training data, whereby there would be more occurrences of some labels than others (i.e. the task-topic types with shorter documents would contain more instances than the types with

Table 1.1 Word counts in CLC_2009_B1_balanced for each of six task-topic labels

Task-topic	**Examples**	**Essays**	**Words**	**Mean document length (SD)**
administrative	•write to all staff about office equipment	68	2986	43.9 (9.7)
autobiographical	•write about your friends •write about last weekend	68	7450	110.0 (22.7)
narrative	•write a story which begins with the following sentence	68	8591	126.3 (33.4)
professional	•write a reply to a conference invitation •write a covering letter for a job application	68	5827	85.7 (21.4)
society	•write about the education system in your country	68	9720	142.9 (42.0)
transactional	•invite a friend to a picnic •write to your teacher to explain your absence from class	68	3651	53.7 (17.9)
Total		*408*	*38,225*	93.7 9 (45.0)

longer documents). Thus, we opted for a per-document or per-essay perspective in our analysis tasks.

We back off to the smallest class in our task-topic taxonomy – namely, 'society', for which we have 68 essays. We thus gather 68 essays from each of our six labels to produce a 408-essay (38,225 tokens) section of CLC_2009_B1 (henceforth 'CLC_2009_B1_balanced') for all experiments described below. In the case of 'society', our 68-essay sub-corpus will include all available essays, whereas in the case of the 'administrative', 'narrative' and 'professional' classes it will include about half of the available essays for each, and in the case of 'autobiographical' and 'transactional' we sample approximately 1-in-20 essays. The structure of CLC_2009_B1_balanced is given in Table 1.1.

4 Case study 1: Classification based on lexical features

We first ask whether task-topic has a noticeable effect on 'lexical features' – that is, vocabulary. To address this question, we turned to a machine learning procedure to test whether essays could be correctly assigned their label based only on their lexical contents. We prepared a training set of 90 per cent of the essays in CLC_2009_B1_balanced, each one as an unordered 'bag of words' along with its task-topic label, in order to train a naive Bayes classifier.[6] The idea is that each word in the bag becomes associated with that label, and the classifier, if presented with a sufficient quantity and distinctiveness of training data, can then use the presence of those features (words) in unlabelled essays to hypothesize which of the set of labels to apply to them. The remaining 10 per cent of the essays from CLC_2009_B1_balanced were presented as unlabelled bags of words to the classifier, and its accuracy in identifying the labels in this test-set is then calculated by comparison with the true labels which had been held back. This procedure was repeated a further nine times, shifting the windows of the 90 per cent training set and the 10 per cent test-set across CLC_2009_B1_balanced, and taking average accuracy values from all ten iterations – a method known as 'tenfold cross-validation'.

In preparing the essays for classification, not only did we follow the standard practice of removing common function words (e.g. articles, prepositions, pronouns, *wh*-words, auxiliary verbs), or 'stopwords', from the essays, but we also omitted any lexical items given in the prompt itself. Thus, the classifier is trained purely on spontaneously composed lexical features rather than a restricted set of imitated lexical items copied from the given exam materials, and we were not simply identifying the labels of essays according to the inevitable repetition of certain keywords from the prompts.

The classifier was found to have a mean accuracy of 88.8 per cent (versus a chance baseline of 1 in 6, or 16.7%) averaged across our ten iterations.[7] Mean precision and recall are given in Table 1.2, with 'precision' for each label *l* being the proportion of correctly identified documents out of the total number of documents the classifier hypothesizes to have label *l*, and 'recall' being the number of correctly identified documents with label *l* out of the actual number of documents with label *l*. As shown in Table 1.2, the balance of precision and recall scores vary by task-topic label. For instance, precision is just 76.8 per cent for 'autobiographical' whereas recall is 94.6 per cent. The reverse is true of 'transactional', with 96.4 per cent precision and 55.7 per cent recall. The '*F*-measure', the harmonic mean of precision and recall, is found to be high (greater than 9 in 10) for all labels except 'autobiographical' and 'transactional'.

A further aspect of this exercise is that we can query the trained classifiers for the features deemed most informative – that is, words which distinguish strongly between task-topic labels. We obtained the twenty most informative features from each training iteration and list those that occur in at least two of these lists in Table 1.3. Each set of word tokens makes sense when compared with the kinds of prompts associated with each label (Table 1.1). Notably, the number of highly informative words for 'transactional' is small – only 'see', in fact – symptomatic of the poor *F*-measure returned for this label in the classification exercise, though its high precision hints at a latent set of mildly informative features (Table 1.2).

Evidently, the classifier is over-hypothesizing essays to be 'autobiographical', presumably because its most discriminating lexical features occur frequently in other task-topics too. The 'transactional' label is underused, however, indicating that its lexical features are relatively indistinguishable from those of other labels. We infer that the 'transactional' task-topic is relatively bland in terms of encouraging distinctive lexical choice, a hypothesis confirmed

Table 1.2 Mean precision, recall, and F-measure for a naive Bayes classifier trained on CLC_2009_B1_balanced for each of six task-topic labels

Task-topic	Precision	Recall	F-measure
administrative	0.977	0.846	0.912
autobiographical	0.768	0.946	0.845
narrative	0.919	0.962	0.944
professional	0.905	0.992	0.948
society	0.869	0.989	0.928
transactional	0.964	0.557	0.694

Table 1.3 Word tokens which are highly informative in at least two iterations of the tenfold topic classification exercise

Task-topic label	Highly discriminative word tokens
Administrative	Budget, department, computers, soon, old, please
Autobiographical	Cinema, play, time
Narrative	Happy, house, love, person
Professional	Opportunity, talk, company, work, staff
Society	School, learn, teachers, years
Transactional	See

by training a binary classifier on just the 'administrative' and 'transactional' essays in CLC_2009_B1_balanced and finding that the most informative features for the latter are common verbs such as 'come', 'go' and 'think' – so it is little wonder that these features would be relatively indeterminate against five competitor labels.

The other four labels are identified more accurately, especially 'narrative' and 'professional', in both cases indicative of more restricted lexical domains – because in the former the story is defined by an opening line given in the prompt (even though keywords from the line itself are excluded as stopwords), and in the latter there is a need to discuss work matters. This gives rise to informative features such as 'opportunity' and 'staff' in the case of the 'professional' type, while the opening line for most of the 'narrative' essays in CLC_2009_B1_balanced initiated a story about 'the best day of Lisa's life', hence 'happy' and 'love' being discriminative features for this group.

In summary, this case study has underlined that not all task-topic types afford equal opportunity of use for learners to demonstrate lexical knowledge in English. That is, the selection of prompt topic is found to have a somewhat deterministic effect on learners' vocabulary use in various ways, in such a way that we can train a classifier to correctly identify the topic of nine in ten essays based on word features alone. The implication is that the choice of prompt in an exam affords different opportunity to demonstrate knowledge of different vocabulary sets related to the topic. Thus, if the prompt topic is more abstract than concrete, for example, then the learner can be expected to employ more abstract than concrete vocabulary. It remains a matter for future work whether such differences affect assessment of learners – whether a highly abstract vocabulary set correlates with higher grades than a more concrete set.

5 Case study 2: Word-class frequency

At a level away from direct lexical features, one can generalize by examining linguistic use in terms of part-of-speech tags. Such an approach is commonly used in LCR, often as a foundation for analyses of larger syntactic patterns (e.g. Hawkins and Buttery 2010; Hawkins and Filipović 2012). We investigate whether task-topic has an effect on part-of-speech (PoS) frequencies: does task-topic entail similar distributions of PoS tags? Since they are foundational for much work in LCR, this is an important question to answer, especially having found that document length has a nonlinear effect on adverb use, for example (Buttery and Caines 2012b).

We used the RASP System ('Robust Accurate Statistical Parsing'; Briscoe, Carroll and Watson 2007) to process CLC_2009_B1_balanced, our dataset containing sixty-eight essays for each of our six labels. We gathered PoS tag frequencies for each of four major classes – nouns, adjectives, verbs, adverbs – from each essay. The distribution of these frequencies is presented as density plots in Figure 1.1.[8] Density plots serve a similar purpose to histograms, in that they portray distributions, but 'smooth' the distribution rather than 'bin' it (as in histograms) and therefore offer a way to visualize the underlying distribution. We use the normal (Gaussian) distribution as the 'kernel' – the weighting function – for our density plots, and the bandwidth smooths the values by the standard deviation of the kernel (Silverman 1998). Note that the area under each curve sums to one, and thus gives a proportional representation of the underlying PoS distributions.

It is apparent from Figure 1.1 that the density distributions of noun, adjective and verb frequencies are broadly similar across the six task-topic labels. There are some differences in height (e.g. nouns, adjectives) and position (e.g. verbs) of the 'peaks'. Adverb distributions are noticeably more varied, though it should be observed that these are the least frequent word-class represented in Figure 1.1 (see the x-axis) and we might speculate that the differences would in fact smooth out with more data.

In any case, we can measure the apparent differences statistically, using the two-sample Kolmogorov-Smirnov test[9] (K-S) to determine how the distributions compare in a pairwise fashion: that is, taking the PoS frequencies from each set of essays. The results of these K-S tests, known as the '*D* statistic', indicate how strongly the two samples differ and are presented in Table 1.6 (nouns and adjectives) and Table 1.7 (verbs and adverbs), with *p* values smaller than 0.001 being marked with an asterisk.

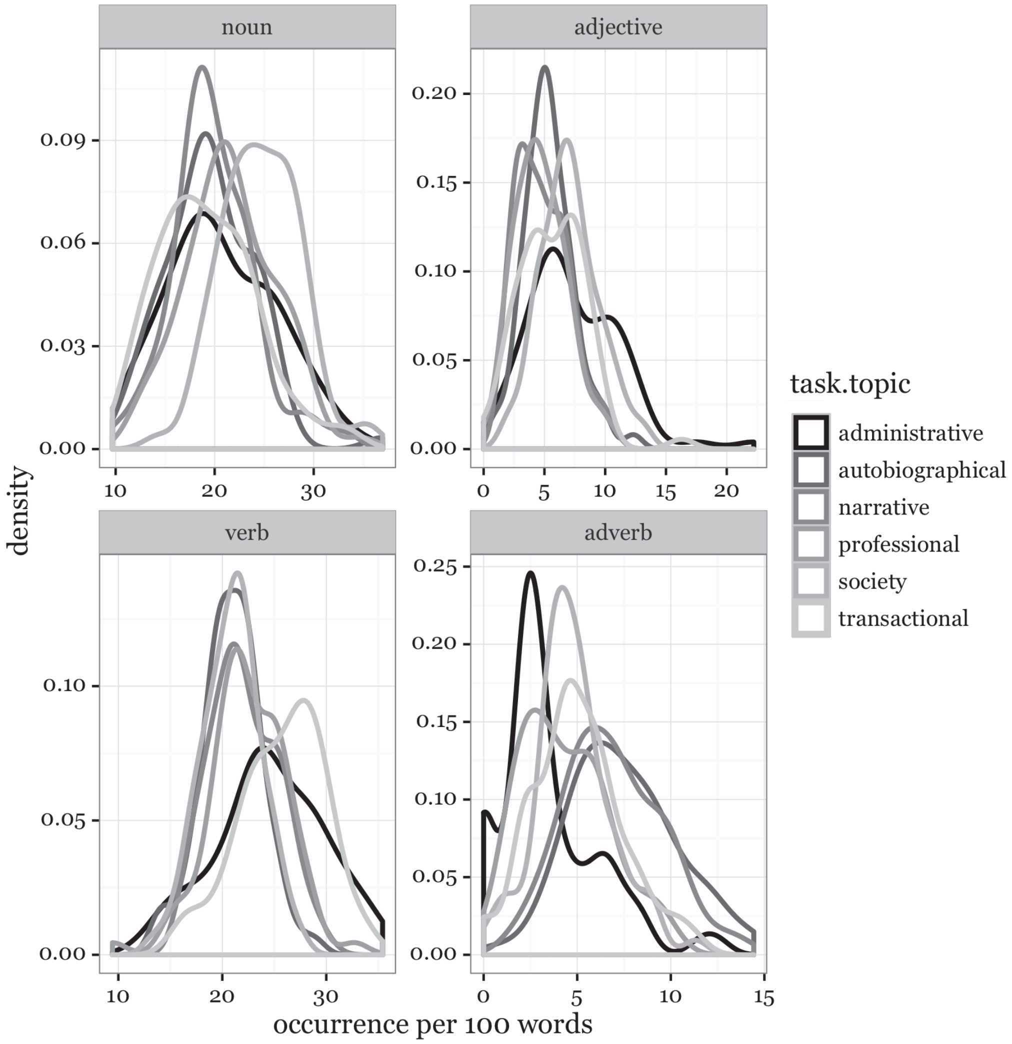

Figure 1.1 Noun, adjective, verb and adverb frequencies by task-topic in CLC_2009_ B1_balanced.

As shown in Table 1.4, the majority of noun and adjective comparisons are not significantly different and therefore we assume the frequencies are drawn from the same distribution in these cases. For those few tests which do produce significant differences, we note that most involve 'society' documents.

Table 1.5, however, confirms the visual impression from Figure 1.1, namely, that the verb and adverb distributions are more heterogeneous than the noun and adjective ones. More than half the tests are significantly different, with the 'transactional' group the main source of disparity for verbs, and 'administrative', 'autobiographical' plus 'narrative' topics showing significant differences for four of five comparisons each.

Table 1.4 Kolmogorov-Smirnov pairwise task-topic label comparisons of noun (upper left, white background) and adjective (lower right, grey background) frequency distributions in CLC_2009_B1_balanced (D = x, with * indicating p < 0.001)

	Transactional	**Society**	**Professional**	**Narrative**	**Autobiographical**	**Administrative**
Administrative	0.221	0.368*	0.103	0.235	0.206	–
Autobiographical	0.118	0.544*	0.265	0.118	–	0.324
Narrative	0.176	0.515*	0.235	–	0.206	0.338*
Professional	0.279	0.353*	–	0.088	0.206	0.353*
Society	0.515*	–	0.426*	0.397*	0.382*	0.176
Transactional	–	0.279	0.250	0.221	0.176	0.294

Table 1.5 Kolmogorov-Smirnov pairwise task-topic label comparisons of verb (upper left, white background) and adverb (lower right, grey background) frequency distributions in CLC_2009_B1_balanced (D = x, with * indicating p < 0.001)

	Transactional	**Society**	**Professional**	**Narrative**	**Autobiographical**	**Administrative**
Administrative	0.162	0.485*	0.265	0.324	0.485*	–
Autobiographical	0.574*	0.074	0.250	0.206	–	0.647*
Narrative	0.441*	0.176	0.118	–	0.103	0.603*
Professional	0.426*	0.235	–	0.412*	0.471*	0.279
Society	0.574*	–	0.250	0.397*	0.456*	0.515*
Transactional	–	0.132	0.221	0.338*	0.412*	0.441*

Considered together with Figure 1.1, these results indicate that 'society' essays involve greater use of nouns and adjectives, 'transactional' essays involve heavy use of verbs, while adverbs are used with high frequency in 'autobiographical' and 'narrative' essays and with low frequency in 'administrative' essays. The main outcome of this case study is that the task-topic of an essay, set by the prompt, entails different frequencies of word use. The implication is that for much work in LCR, which as a field tends to treat proficiency sub-corpora holistically, there is a risk of confounding any findings with the effects of task-topic. An increase in *x*, where *x* could be the frequency of a noun, for example, could in fact be a consequence of differing proportions of discursive exercises at proficiency level Y compared to proficiency level Z.

For example, Cambridge English business certificates are only intended to cover CEFR levels B1, B2 and C1, and thus the CLC A1, A2 and C2 subsections do not include business topic texts. Indeed, we find that between 10 and 20 per cent of the essays in CLC B1, B2 and C1 come from business exams: in CLC_2009_B1 at least, all such essays are exclusively 'administrative' or 'professional' essays, and if that is true of B2 and C1 business texts also then these three levels each receive a distinctive injection of rather homogeneous data. Such differences in the constitution of proficiency sub-corpora might lead researchers to mis-identify linguistic feature correlates, such as the ones described here, as entirely proficiency-driven, rather than partly or wholly based on task-topic.

6 Case study 3: Subcategorization frames

Our third and final investigation of task-topic effect involves 'subcategorization frames' (SCFs), a set of 163 frames which describe verbs and their arguments. These range from the single-argument intransitive frame (4), to the two-argument transitive (5), the ditransitive with three arguments (6) and then more complex constructions involving extraposition, clausal complements and so on[10] (Briscoe and Carroll 1997; Buttery and Caines 2012a).

(4) Stephen surfs. SCF 22: INTRANS
(5) Vic bought a juicer. SCF 24: NP
(6) Lindsay put Harvey on the floor. SCF 49: NP-PP

Note that verbs may be associated with more than one SCF depending on argumentation. For instance, ‘surf’ can also be transitive (7), ‘buy’ can be used as a ditransitive (8) and ‘put’ may take a phrasal particle (9):

(7) Stephen surfs the internet. SCF 24: NP
(8) Vic bought me a juicer. SCF 37: NP-NP
(9) Lindsay put up with his foibles. SCF 76: PART-NP / NP-PART

We extracted SCFs from each of the 408 essays in CLC_2009_B1_balanced, automatically identifying argumentation patterns on the basis of RASP System output. Where there is syntactic ambiguity as to which frame is in use, the possible candidates are concatenated with commas. For example, ‘49,56’ would indicate a syntactic ambiguity between frame 49 (10) and frame 56 (11).

(10) He posted a sign to the wall. SCF 49: NP-PP
(11) He posted a letter to her. SCF 56: NP-TO-NP

As can be seen in (10) and (11), semantic information actually disambiguates between the two possible syntactic analyses, but since this kind of information is not available to the parser, a frame of concatenated SCF options is posited instead.

Through this analysis of SCFs we aim to gain a fuller insight into the variance in constructional use across task-topics. The issue of task-topic has mainly focused on lexical features thus far, and so by moving to the constructional level we test whether this variable affects syntactic use as well as lexical selection.

We identified 66 unique SCFs in CLC_2009_B1_balanced, of which 18 are concatenated multiple frames of the type ‘SCF, SCF, (SCF),…’. We obtain frequency counts for these 66 SCFs for each of the 408 essays in CLC_2009_B1_balanced. This 26,928-cell matrix (408 row, 66 column) may be reduced to a low-dimensional space through LDA using R and the *MASS* package (R Core Team 2015; Venables and Ripley 2002). In LDA, the vertical dimensions of our data table, the 66 SCFs, are transformed into new axes which combine these variables so that between-group differences are maximized while within-group differences are minimized. Each new dimension ‘explains’ a certain amount of variance found in the data, a process which conventionally allows reduction of the original multiple dimensions to just a few which best account for the dataset (Kuhn and Johnson 2013).

In our case, we first establish through ‘principal components analysis’, an unsupervised clustering method, that five dimensions account for approximately

75 per cent of the variance in the data. Our LDA model, also reduced to five dimensions, correctly classifies the task-topic label for 65 per cent of our 408 essays, a highly significant result according to a binomial test ($p < 0.001$). Figure 1.2 shows a density plot of the 408 essays from CLC_2009_B1_balanced in dimensions 1 and 2 of the LDA, grouped by topic. What is apparent is the relatively high variance in terms of SCF use within the 'professional' set of documents, the relatively low variance within 'administrative' essays, and the overlap of 'autobiographical' and 'society' essays on the one hand, versus the four remaining, more distributed labels.

The accuracy of the LDA and the clusters apparent in Figure 1.2 indicate that constructional use is somewhat affected by task-topic. To further establish the nature of this effect, we plot a heatmap of frequencies for the twenty most frequent SCFs in CLC_2009_B1_balanced (Figure 1.3).

In Figure 1.3 we can see that the 'autobiographical', 'narrative' and 'society' essays contain a greater range of the most frequent SCF types, whereas 'administrative', 'professional' and 'transactional' are more limited in this regard, including some zero (white) values. Moreover, the latter set are relatively low

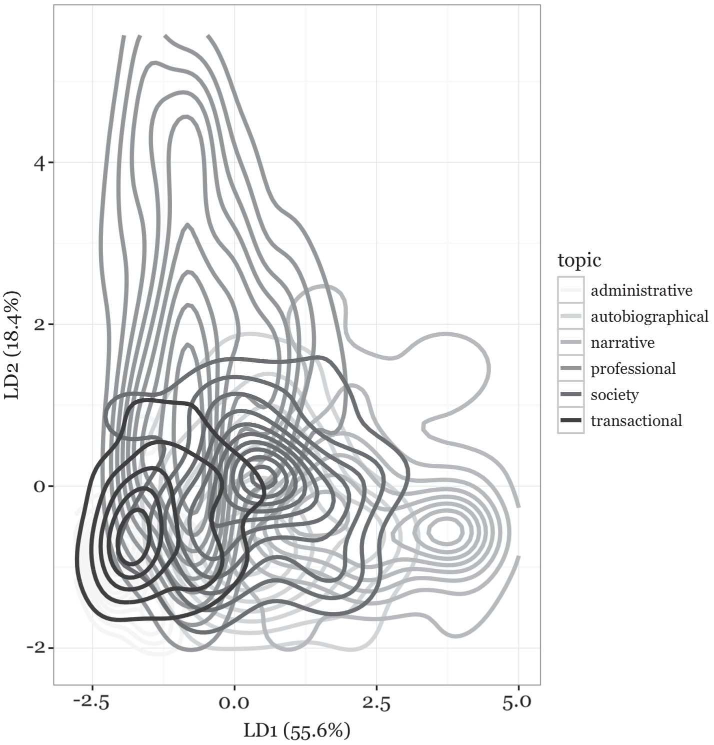

Figure 1.2 Linear discriminant analysis of subcategorization frames by task-topic in CLC_2009_B1_balanced.

frequency for SCF 87,96 – prepositional arguments of some kind – while maintaining a high frequency for SCF 18,142 (bare infinitive arguments) and SCF 112,111,110 (*to*-infinitives), 'professional' in particular.

In Table 1.6 we present the ten most frequent SCFs by task-topic. Again, we see some notable differences in ranking, with the prepositional SCF 87,96 of lower frequency in 'professional' and 'transactional', and not even among the ten most frequent SCFs for 'administrative'. Meanwhile, subordinate *that*-clauses (SCF 104,109) are ranked more highly for 'narrative' and 'administrative' than other topics, while 'society' uniquely has adjectival arguments (SCF 1,2) in its top ten.

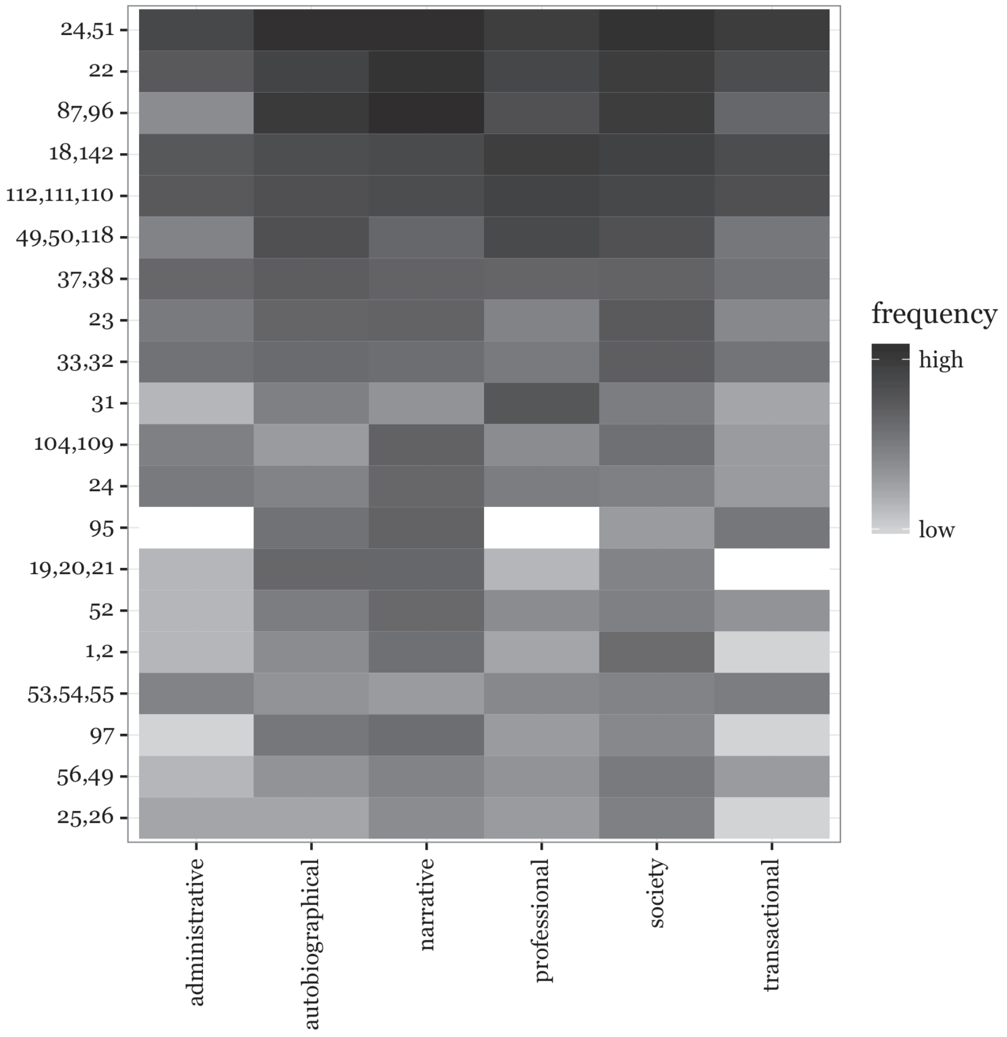

Figure 1.3 Heatmap of frequencies by task-topic for the twenty most frequent subcategorization frames in CLC_2009_B1_balanced, with darker shades indicating higher frequency.

Table 1.6 The ten most frequent subcategorization frames for each task-topic sub-corpus in CLC_2009_B1_balanced

Rank	Administrative	Autobiographical	Narrative	Professional	Society	Transactional
1	24,51: NP(+RS)	24,51: NP(+RS)	87,96: PP	18,142: INF/SC	24,51: NP(+RS)	24,51: NP(+RS)
2	18,142: INF/SC	87,96: PP	24,51: NP(+RS)	24,51: NP(+RS)	22: INTRANS	18,142: INF/SC
3	110,111,112: TO-INF	22: INTRANS	22: INTRANS	110,111,112:TO-INF	87,96: PP	22: INTRANS
4	22: INTRANS	18,142: INF/SC	18,142: INF/SC	22: INTRANS	18,142: INF/SC	110,111,112: TO-INF
5	37,38: NP-NP	49,50,118: NP-PP	110,111,112:TO-INF	49,50,118: NP-PP	110,111,112: TO-INF	87,96: PP
6	32,33: NP-INF	110,111,112: TO-INF	104,109: THAT-S	87,96: PP	49,50,118: NP-PP	37,38: NP-NP
7	23: INTRANS-RECIP	37,38: NP-NP	23: INTRANS-RECIP	31: NP-FOR-NP	23: INTRANS-RECIP	32,33: NP-INF
8	24: NP	23: INTRANS-RECIP	95: PP-PP	37,38: NP-NP	32,33: NP-INF	49,50,118: NP-PP
9	104,109: THAT-S	19,20,21: ING	37,38: NP-NP	32,33: NP-INF	37,38: NP-NP	95: PP-PP
10	49,50,118: NP-PP	32,33: NP-INF	19,20,21: ING	24: NP	1,2: ADJP	53,54,55: NP-TO-INF

This sketch of SCF frequency differences across task-topic types is presented as evidence for this variable's constructional effects, which needs to be kept in mind when comparing learners within and between proficiency levels.

7 Differentiation at the prompt level

We recognize that the lexical content of the prompts associated with the essays in large learner corpora may not always be available. Thus we attempt to model linguistic use at a per-prompt level, assuming that researchers can at least group the essays by a prompt identifier. As with CLC_2009_B1_balanced, we include only those prompts answered by at least 68 essays. This gives us a new corpus of 612 essays responding to nine prompts, which we again process with the RASP System, and extract PoS frequency counts for nouns, adjectives, verbs and

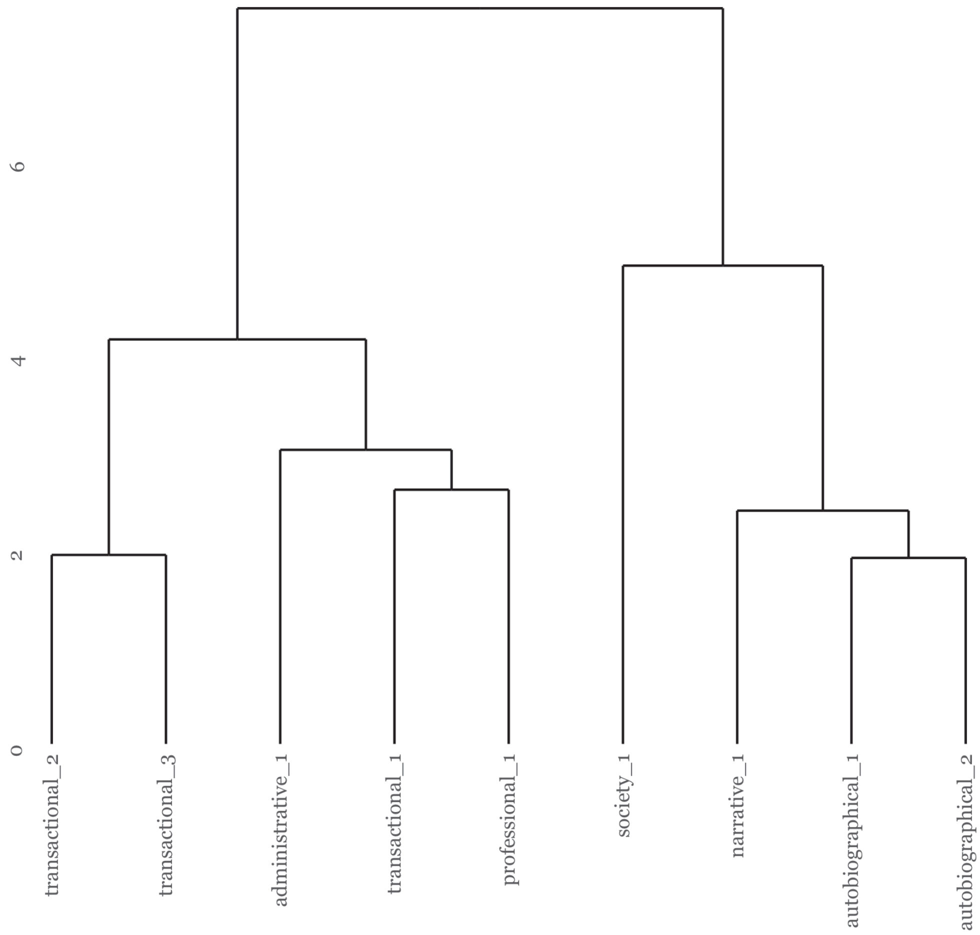

Figure 1.4 Dendrogram of prompt sub-corpora from CLC_2009_B1_balanced.

adverbs. We then normalize the relative frequencies and type:token ratios and apply a hierarchical clustering algorithm using the core *stats* package in R (R Core Team 2015). The results of the clustering are presented as a dendrogram in Figure 1.4 (de Vries and Ripley 2015).

First, the bottom leaves of Figure 1.4 rather support our topic label taxonomy, in that the 'autobiographical' prompt subsets are grouped together, two of the three 'transactional' subsets are grouped together, while the singleton 'administrative', 'narrative' and 'society' prompts branch off individually.

In the common scenario where the prompts themselves are not available to researchers, but the essays in a learner corpus may still be grouped by a prompt identifier, one could use this method to group the prompts into pseudo-topics whose identity is initially unknown but could be discovered via the classification or LDA techniques employed above. For example, the second branching would give four groups, which we have here known as 'transactional', 'administrative' + 'professional' + 'transactional', 'society', 'narrative' + 'autobiographical', but having been rearranged in this bottom-up fashion could be relabelled appropriately.

8 Implications for research, assessment and pedagogy

We coined the term 'opportunity of use' and emphasized that variables such as document length need to be taken into account when comparing language use across proficiency levels. The point is that learners are not afforded the same opportunity to use adverbs across the learning and examination pathway – there is a bias in higher proficiency-level tasks that allows them more opportunity to demonstrate this know-how, and thus document length introduces a confound to any LCR that treats the corpus as a homogeneous set of documents. Here we consider document task-topic as another variable that needs to be controlled for, in continuing work on opportunity of use.

In terms of how this work relates to LCR more generally, we advise researchers, if they do not already, to control for variables such as document length, task, topic and first language as far as possible. We recognize this is not always feasible, given limitations on obtainable data and metadata, and therefore these factors need to be acknowledged and understood where they cannot be fully controlled.

With regard to document task-topic we advise the following precautionary steps:

1: Are the prompts available for the essays in your corpus?
Yes. Then review the prompts and design a set of task-topic labels to reflect their diversity; if the distinctiveness of the label set can be confirmed through language modelling and/or statistical tests, all the better.
No. Adopt an unsupervised machine learning approach to group the essays into sub-corpora, since document 'labels' (i.e. prompts) are unknown. For example, one might employ hierarchical clustering where at least question numbers for the essays are known (as in our corpus), or *k*-means clustering if no such information is available – where *k* is set *a priori* as the number of desired document groupings.
2: Make comparisons across proficiency-level sub-corpora restricted to the same or similar task-topics, using methods such as those described here.

We acknowledge that in reality there are often limitations to what information is readily available, and that despite best efforts the effects of other variables persist – such as the document length confound in our restricted dataset CLC_2009_B1_balanced (Table 1.1). However, by attempting to control such factors, or at least being aware of them, researchers can avoid making inappropriate inferences over highly heterogeneous data.

Examiners and assessors should also be aware that 'opportunity of use' is not necessarily equal for certain linguistic features across different task-topic types. In its document on CEFR, the Council of Europe sets out the lexical knowledge expected at each level, as shown in Table 1.7.

Milton (2010: 214) raises some valid questions in response to the vocabulary range descriptors set out in Table 1.7, relating to how these broad characterizations of knowledge are to be measured in practice, and furthermore, 'as to how learners are to demonstrate this knowledge when the tasks presented to them … only allow them to produce a few hundred words, and most of these will be highly frequent and common to most learners'. On the assessment side, Daller and Phelan (2007) found that raters can be inconsistent in applying these vocabulary criteria, while we demonstrate that not all prompts induce the same range of lexical items – indeed we might infer that not all essays encourage or necessitate the same range of lexical items. However, if the topics and tasks are deemed to be the most suitable for the relevant exams, then what's needed here is awareness of the linguistic consequences on the part of researchers and assessors.

From a pedagogical perspective, the effects of prompt topic are to be viewed as a by-product of task and topic variants rather than something to be altered. We suggest that teachers may use the information presented here to their advantage, in order to, for example, focus on a particular lexical set, or a

Table 1.7 CEFR level vocabulary range descriptors (Council of Europe 2001)

CEFR level	Vocabulary range
C2	Has a very good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms, shows awareness of connotative levels of meaning.
C1	Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.
B2	Has a good range of vocabulary for matters connected to his or her field and most general topics. Can vary formulation to avoid repetition, but lexical gaps can still cause hesitation and circumlocution.
B1	Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel and current events.
A2	Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics. Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs.
A1	Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.

particular construction type, per the observation that different tasks and topics give learners practice in different aspects of complexity, accuracy and fluency (Yuan and Ellis 2003). As has been shown here, there are certain task-topic types that encourage the use of certain lexico-syntactic constructs. For example, 'narrative' essays encourage the use of verb construction with prepositional arguments, 'autobiographical' task-topic types encourage use of adverbs, and 'professional' prompts lead to a greater use of business vocabulary. It is analyses such as these that may be harnessed in pedagogy to further broaden learners' linguistic repertoire, as diversity in this respect is known to associate strongly with increasing proficiency (Vercellotti, 2017).

Acknowledgements

This chapter reports on research supported by Cambridge English, University of Cambridge. We gratefully acknowledge the help of Dr Nick Saville and Dr Fiona

Barker of Cambridge English, Professor Ted Briscoe and Dimitrios Alikanoitis of the University of Cambridge and Professor Michael McCarthy of the University of Nottingham. We also thank two anonymous reviewers and the two editors for their detailed and helpful feedback which greatly improved this chapter.

Notes

[1] See www.pearsonlongman.com/dictionaries/corpus/learners.html.

[2] See www.cambridgeenglish.org.

[3] The 'Common European Framework of Reference for Languages': see www.coe.int/lang-CEFR.

[4] We use 'essay' here in its general sense: 'a composition of moderate length on any particular subject' (*The Oxford English Dictionary*).

[5] Our thanks to Dr Fiona Barker of Cambridge English Language Assessment for her help in this. All prompts are © UCLES 2009 (1), and 2014 (2), (3).

[6] This type of classifier is Bayesian as it implements Bayes's theorem, and it is 'naive' in the sense that it assumes independence between features; see Bird, Klein and Loper (2009) for further background information.

[7] Classifier accuracy without excluding keywords repeated from the prompts was 96.3 per cent, a large improvement on the prompt stopword classifier presented here, indicating that the presence of prompt keywords makes the task of assigning labels much easier, and hinting at a method for unsupervised grouping of unlabelled essays through clustering and the use of prompt keywords as labels.

[8] This and all further plots produced using *ggplot2* for R unless otherwise stated (Wickham 2009; 25R Core Team 2015).

[9] For background information about this test, see Corder and Foreman (2014).

[10] For a full list of all 163 frames, download the VALEX package from http://ilexir.co.uk/applications/valex (accessed on 20 June 2016).

References

Ädel, A. (2015), 'Variability in learner corpora', in S. Granger, G. Gilquin and F. Meunier (eds), *The Cambridge Handbook of Learner Corpus Research*, 379–400, Cambridge: Cambridge University Press.

Biber, D., Gray, B. and Staples, S. (2016), 'Predicting patterns of grammatical complexity across language exam task types and proficiency levels', *Applied Linguistics,* 37(5): 639–68.

Bird, S., Klein, E. and Loper, E. (2009), *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, Sebastopol, CA: O'Reilly Media.

Blei, D. M. (2012), 'Probabilistic topic models', *Communications of the ACM*, 55(4): 77–84.

Briscoe, T. and Carroll, J. (1997), 'Automatic extraction of subcategorization from corpora', *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, Washington, DC, USA, 31 March to 3 April. Association for Computational Linguistics. Available at: https://arxiv.org/pdf/cmp-lg/9702002.pdf (accessed 9 January 2017).

Briscoe, T., Carroll, J. and Watson, R. (2007), 'The second release of the RASP system', *Proceedings of the COLING/ACL Interactive Presentation Sessions,* Sydney, Australia, 16–17 July. Association for Computational Linguistics. Available at:https://pdfs.semanticscholar.org/7410/c010a38e7e23f38c3c6e898d5695a4874c61.pdf (accessed on 9 January 2017).

Buttery, P. and Caines, A. (2012a), 'Reclassifying subcategorization frames for experimental analysis and stimulus generation', *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), Istanbul, Turkey, 23–25 May. European Language Resources Association. Available at; http://www.lrec-conf.org/proceedings/lrec2012/pdf/1063_Paper.pdf (accessed 9 January 2017).

Buttery, P. and Caines, A. (2012b),'Normalising frequency counts to account for "opportunity of use" in learner corpora', in Y. Tono, Y. Kawaguchi and M. Minegishi (eds), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*, 187–204, Amsterdam: John Benjamins.

Corder, G. W. and Foreman, D. I. (2014), *Nonparametric Statistics: A Step-by-step Approach,* 2nd edn, Hoboken, NJ: John Wiley & Sons.

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

Daller, H. and Phelan, D. (2007), 'What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness', in H. Daller, J. Milton and J. Treffers-Daller (eds), *Testing and Modelling Lexical Knowledge*, 234–45, Cambridge: Cambridge University Press.

de Vries, A. and Ripley, B. D. (2015), ggdendro: Create Dendrograms and Tree Diagrams using 'ggplot2', R package version 0,1–17, URL http://CRAN,R-project,org/package=ggdendro

Granger, S., Dagneaux, E., Meunier, F. and Paquot, M. (2009), *International Corpus of Learner English v2*, Louvain-la-Neuve: Presses universitaires de Louvain.

Gries, S. (2003), *Multifactorial Analysis in Corpus Linguistics: A study of particle placement*, London: Continuum Press.

Hawkins, J. and Buttery, P. (2010), 'Criterial features in learner corpora: theory and illustrations', *English Profile Journal*, 1(1): e5.

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

Hinkel, E. (2009), 'The effects of essay topic on modal verb uses in L1 and L2 academic writing', *Journal of Pragmatics,* 41: 667–83.

Khabbazbashi, N. (2017), 'Topic and background knowledge effects on performance in speaking assessment', *Language Testing,* 34(1): 23–48.

Kobayashi, Y. and Abe, M. (2014), 'A machine learning approach to the effects of writing task prompts', *Learner Corpus Studies in Asia and the World,* 2: 163–75.

Kuhn, M. and Johnson, K. (2013), *Applied Predictive Modeling*, Berlin: Springer.

Milton, J. (2010), 'The development of vocabulary breadth across the CEFR levels', in I. Bartning, M. Martin and I. Vedder (eds), *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research,* 211–32, EUROSLA Monograph Series, 1.

Newton, J. and Kennedy, G. (1996), 'Effects of communication tasks on the grammatical relations marked by second language learners', *System,* 24: 309–22.

Nicholls, Diane (2003), 'The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT', *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster University, UK, 28–31 March. Available at: http://ucrel.lancs.ac.uk/publications/cl2003/papers/nicholls.pdf (accessed on 9 January 2017).

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing, URL http://www,R-project,org (accessed 20 June 2016).

Silverman, B. (1998), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.

Tummers, J., Speelman, D. and Geeraerts, D. (2014), 'Spurious effects in variational corpus linguistics', *International Journal of Corpus Linguistics*, 19: 478–504.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, 4th edn, Berlin: Springer.

Vercellotti, M. L. (2017), 'The development of complexity, accuracy, and fluency in second language performance: a longitudinal study', *Applied Linguistics*, 38 (1): 90–111.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Berlin: Springer,

Yuan, F. and Ellis, R. (2003), 'The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production', *Applied Linguistics*, 24: 1–27.