Syntactic complexity across registers: Investigating (in) formality in second-language writing

Tove Larsson a, b, *, Henrik Kaatari c

a Department of English, Uppsala University, Sweden
b Department of English, Northern Arizona University, USA
c Department of Humanities, University of Gävle, Gävle, Sweden

ARTICLE INFO

Article history:
Received 31 August 2019
Received in revised form 16 December 2019
Accepted 3 February 2020
Available online 4 February 2020

Keywords:
Syntactic complexity
Register
Formality
L2 writing
Noun-phrase complexity

ABSTRACT

Syntactic complexity (i.e. the grammatical sophistication exhibited in language production) has been found to be positively correlated with formality. In viewing formality as a cline rather than as a dichotomy, the present study revisits previous claims about (in) formality in learner writing in relation to syntactic complexity. The use of commonly utilized measures of syntactic complexity is explored in learner writing and across four registers from the British National Corpus (academic prose, popular science, news and fiction). The results show that while the learners generally exhibited appropriate register awareness, there were some differences noted between their writing and that of the published writers, in particular with regard to the measure of complex nominals. A detailed analysis of this measure in the academic register showed that the learners make less frequent use of adjectival and prepositional modifiers than the expert writers. Our results thus confirm previous claims about the importance of phrase-level complexity measures as a predictor of formality. It would seem that learners would benefit from some targeted instruction of such structures for increased register awareness.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Syntactic complexity has received much attention in many recent studies in subfields such as Second Language Acquisition (SLA) and Learner Corpus Research (LCR) where it is often used as a measure of linguistic development in second-language (L2) writing (e.g. Ansarifar, Shahriari, & Pishghadam, 2018; Housen & Simoens, 2016; Larsen-Freeman, 1978; Lu, 2011). Following Ortega (2015, p. 82), syntactic complexity is here defined as “the range and the sophistication of grammatical resources exhibited in language production”. As such, syntactic complexity thus covers related concepts such as diversity, variety and degree of linguistic elaborateness. It is therefore not surprising that syntactic complexity has been found to be positively correlated with formality, as formal, academic texts tend to be characterized by elaborate and diverse language (e.g. Biber, 2016). In the present study, formality is defined as a situational construct and operationalized through the use of registers, as explained in further detail below.

Achieving an appropriately high level of formality in one’s academic texts requires practice, and is something that novice writers, in particular learners, have been found to struggle with. In fact, learners are often described as being overly informal in
their writing (c.f., e.g., Altenberg & Tapper, 1998). However, many previous claims of the informal nature of learner writing are based on linguistic (rather than situational) definitions of formality and on reports of over/underuse of a limited number of features, leading to a dichotomous view where student production is either viewed as formal or informal; few studies have sought to investigate formality in a more systematic manner.

In order to provide a more nuanced picture of formality, the present exploratory study makes use of a method developed in Larsson and Kaatari (2019) to operationalize (in)formality and to test previous claims about (in)formality in relation to syntactic complexity in learner writing and across four registers from the BNC: academic prose, popular science, news and fiction (see also Larsson, 2019). This method makes two assumptions: (i) that formality can be viewed as a continuum rather than as a dichotomy (cf., e.g., Adel, 2008; Smith, 1986), and (ii) that registers can be placed along this continuum based on their situational characteristics (cf. Biber, Johansson, Leech, Conrad, & Finegan, 1999, p. 16).

In more detail, the registers are placed on the informal-to-formal continuum based on their a priori situational characteristics, and thus function as reference points on this continuum. This enables us to explore how complexity measures pattern across the registers and, by extension, whether they are correlated with formality (i.e. whether we get increasingly higher/lower scores as we move from one end of the continuum to the other). We then look at the same measures in the learner data to see where on this continuum we can place them. This method thus allows for a more fine-grained discussion of formality in relation to learner writing (Larsson, 2019; Larsson & Kaatari, 2019). Following Biber et al. (1999, p. 16), an ordering of the registers included in the present study from more formal to less formal looks as follows: academic prose, popular science, news and fiction, as shown in Fig. 1 (adapted from Biber et al., 1999, p. 43; cf. also Larsson, 2019).

In the first part of the study, we start out from Lu’s (2017) 14 complexity measures (e.g. phrasal sophistication, subordination and sentence complexity) and map out their distribution across the registers and learner data. However, as these complexity measures confound grammatical units (see Biber, Gray, Staples, and Egbert (2020), and Section 2.1 for a more detailed discussion), a complementary, more detailed analysis was subsequently carried out in the second part of the study focusing on the expert and learner academic prose. The following research questions are used to guide the analysis:

- What is the relative importance of the complexity measures investigated vis-à-vis formality, and to what extent are the measures correlated?
- Which of the BNC registers is the learners’ use closest to, and what can this tell us about (in)formality in learner writing?
- What differences and similarities can be found between the experts’ academic writing and that of the learners?

2. Measures of syntactic complexity

2.1. Operationalizing – and problematizing – syntactic complexity

Syntactic complexity has been reported to be an important measure for assessing second-language (L2) writing proficiency (e.g. Larsen-Freeman, 1978; Lu, 2011). While most previous studies largely agree on the general definition of syntactic complexity, there has been some disagreement with regard to how to best operationalize and measure it. Many studies make use of global measures such as mean length of T-unit and somewhat more fine-grained measures of clausal constructs (e.g. subordinated clauses per T-unit) (cf., e.g., Lu, 2010, 2017; Bulte & Housen, 2018); however, these measures have been criticized, in particular in relation to their usefulness for descriptive accounts of language.

In their extensive survey of linguistic complexity (drawing on Norris & Ortega, 2009), Biber et al. (2020) discuss two major problems of global measures of complexity: the distinctness problem and the redundancy problem. The distinctness problem refers to the fact that global measures confound different linguistic categories. For example, global measures fail to capture differences between clausal and phrasal complexity. Biber, Gray, and Poopon (2011) and Biber et al. (2020) show that the measure mean length of T-unit fails to distinguish between a sentence with extensive use of embedded subordinate clauses and a sentence with several embedded prepositional phrases, but no subordinate clauses. The redundancy problem refers to

<table>
<thead>
<tr>
<th>Formality</th>
<th>Register</th>
<th>Situational characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher</td>
<td>Academic prose</td>
<td>Informational/argumentative text written for a specialist audience</td>
</tr>
<tr>
<td></td>
<td>Popular science</td>
<td>Informational/argumentative text written for a non-specialist audience</td>
</tr>
<tr>
<td></td>
<td>News</td>
<td>Informational/evaluative text written for a regional, wide-public audience</td>
</tr>
<tr>
<td>Lower</td>
<td>Fiction</td>
<td>Written text (typically including written dialogue) for a wide-public audience</td>
</tr>
</tbody>
</table>

Fig. 1. The registers mapped onto the informal-formal cline.
the fact that many global measures of complexity measure similar things. For example, global measures that are based on word length will give more weight to dependent clauses than to dependent phrases since clauses are longer than phrases. It is also noted that most ratio based measures suffer from this problem, as they typically include the frequency of subordinate clauses in the numerator. The authors therefore argue in favor of measures that are based on actual (and more fine-grained) linguistic categories, such as finite dependent clauses along with its subcategories (e.g. conditional clauses, relative that-clauses) (e.g. Biber & Gray, 2016; Biber et al., 2011; Biber, Gray, Staples, & Egbert, 2020).

Using these measures, Biber et al. (2011) noted that clausal complexity is in fact a more prominent feature of conversation than of academic prose, thus calling into question the reliability of clausal measures for assessing proficiency in novice academic writing, and concluding that phrasal measures are preferable. This finding was confirmed in Kyle and Crossley’s (2018) study where it was shown that phrasal complexity indices are better predictors of writing quality than both global and clausal indices. Similarly, other studies have noted that noun pre-modifiers, attributive adjectives and prepositional post-modifiers are particularly important for distinguishing between novice and expert writers. In Ansarifar et al. (2018), where abstracts written by L1 Persian MA students were compared to those written by L1 Persian PhD students and expert writers, it was found that the MA students tended to use noun, participle, adjective and prepositional noun modifiers compared to the PhD students and expert writers. Additional factors such as L1 and text type/genre have also been found to differ with regard to usage of attributive adjectives and pre-modifying nouns (Staples & Reppen, 2016).

It is thus clear that linguistic complexity can be studied at different levels of abstraction; in fact, studies on the topic increasingly operationalize syntactic complexity as a “multidimensional construct” (Norris & Ortega, 2009), encompassing both global and more fine-grained measures (Casal & Lee, 2019). The points raised by Biber et al. (2020) highlight the importance of not relying solely on global measures of complexity for linguistic description. In the present study, the more general measures of complexity are thus merely used as stepping stones for further analysis for us to see where to direct our attention. In the subsequent analysis, we use more fine-grained measures of phrasal complexity (see Section 3.1) that do not suffer from the problems discussed.

2.2. Syntactic complexity and formality

Using a linguistic (rather than situational) definition of formality, several studies on novice writing, in particular learner writing, have found a tendency for these writers to be informal. For example, a higher degree of interpersonal involvement in the learner data (L1 French, Swedish, Finnish and Dutch) than in the NS student data was reported in Petch-Tyson (1998, p. 116), as almost all features of reader/writer visibility (e.g. stance marking, first-person pronouns) were overused by the learners. L1 Swedish users of English have been reported to be particularly informal in their writing, with studies reporting results that led the authors to conclude that L1 Swedish students’ argumentative writing “is similar to NS’ spoken language” (Herriman & Boström Aronsson, 2009, p. 118). This is also echoed in Tapper (2005, p. 124) who notes that “several other studies have also shown that many of the learner sub-corpora in the ICLE corpus — and in some cases the Swedish sub-corpus especially — contain writing that is more informal in style than the native English-speaking student writing”. Although it has since been shown that this alleged lack of formality, at least to some degree, might be explained by the text type that the students have been asked to produce (argumentative writing vs. academic prose, etc.) (Larsson & Kaatari, 2019), the intricate network of syntactic complexity measures can be used to gain further insights about students’ register awareness and, by extension, their adherence to academic norms and standards in terms of formality.

In studies looking specifically at syntactic complexity and register, it has been shown that academic prose differs significantly from face-to-face conversation in that clausal subordination measures are more common in conversation that in academic writing, whereas complex noun phrase constituents and complex phrases are more frequent in academic writing (Biber et al., 2011; Biber & Gray, 2010). Similarly, Biber and Gray (2016) note that academic writing tends to be structurally compressed. With regard to second-language (L2) production, Lintunen and Mäkilä (2014) also found considerable differences in syntactic complexity between the learners’ written and spoken production.

However, while clear differences have been noted between speech and writing in relation to syntactic complexity, much less attention has been paid to more fine-grained register differences. The approach used in the present paper allows us to get a more detailed view of register, which allows us to nuance the perceived formal-informal dichotomy.

3. Method and corpora

3.1. Data processing and automatic analysis of syntactic complexity

In order to investigate syntactic complexity, two different automated tools are used in the present study: the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010; 2017) and the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle, 2016). The latter is used to attain a more detailed picture of noun phrase (NP) complexity.

L2SCA (Lu, 2010; 2017), which is commonly used in studies of complexity (e.g. Casal & Lee, 2019; Paquot, 2018), is a fully automated tool that takes raw text files as input. The texts are tokenized, part-of-speech (POS) tagged and parsed, using the Stanford parser (see Lu, 2010, p. 479). Scores for each of the fourteen complexity measures included are then computed drawing on information from the parse trees. An overview of the complexity measures included in L2SCA is given in Table 1 (see Kyle & Crossley, 2018, p. 338, for a detailed description of the syntactic structures counted by L2SCA). Following Hunt...
A T-unit is here defined as “one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it”.

The measures included in L2SCA were selected by Lu (2010) as they have been found to be correlated with proficiency in previous studies (Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). Lu (2010, p. 488) reports high system-annotator agreement scores for L2SCA. F-scores, which take both precision and recall into consideration, vary between 100% (for sentence identification) and 86.6% (for complex nominals), suggesting that L2SCA achieves a relatively high degree of reliability in terms of identifying the structures necessary for the computation of the complexity measures included.

For the subsequent more detailed analysis, we made use of TAASSC (Kyle, 2016). TAASSC includes a broad range of fine-grained linguistic measures. Like L2SCA, it is fully automated, taking raw text-files as input. Since our main focus for this part of the analysis was to zoom in on NPs to refine our analysis of complex nominals (see Section 4.1), we selected only measures relevant for this phrasal category. The measures considered in this study are listed and exemplified in Table 2. The terminology used in Table 2 comes from Kyle and Crossley (2018, p. 341).1

The study used the statistical environment R (R Development Core Team, 2019) to manage the data and carry out statistical analyses such as conditional inferences trees (CIT) and random forests, as detailed in the corresponding subsection. The (party) package and the functions ‘ctree’ (Hothorn, Hornik, & Zeileis, 2006a) and ‘cforest’ (Hothorn, Buehlmann, Dudoit, Molinari, & Van Der Laan, 2006b; Strobl et al., 2007, 2008) were used for this part of the analysis.

### 3.2. Corpus data

In order to investigate the interplay between syntactic complexity and register, data from two learner corpora and one reference corpus were used: the Advanced Learner English Corpus (ALEC), the Varieties of English for Specific Purposes dAtabase (VESPA; Paquot, Hasselgård, & Oksefjell Ebeling, 2013) and BNC-15.

ALEC comprises 1.3 million words of learner academic writing in English by Swedish university students in English linguistics and English literature. The full VESPA corpus is a multi-million-word corpus of learner academic writing comprising data from learner writers of different mother tongue backgrounds (Paquot et al., 2013); only the Swedish component (VESPA-SE) was used in the present study to enable comparability with ALEC. For this study, we limited the analysis to texts from ALEC (109 texts) and VESPA-SE (22 texts) that were written by students whose self-reported first language (L1) is Swedish and who are in their third year of university studies on average. ALEC and VESPA-SE both comprise (untimed) student theses from the same two universities in Sweden and are thus grouped together in this study (see Larsson & Kaatari, 2019). The reference corpus, BNC-15, is a methodically sampled 3-million-word subset of the British National Corpus (BNC; Burnard, 2007). BNC-15 is designed to enable register comparisons with comparable subsets (for more detailed information about BNC-15, see Kaatari, 2017). Five registers are included in the full BNC-15: academic prose, popular science, news, fiction and conversation. However, for the purpose of the present study, conversation had to be excluded.2 BNC-15, unlike the learner corpora, includes samples of 10,000 words rather than full texts but the average text length is similar to that of VESPA-SE and ALEC.3 An overview of the subsets included in this study is provided in Table 3.
In the subsequent section, the relative importance of the measures introduced in Section 3.1 will be explored across register. We will then add the learner data to see which of the registers their use is closest to, which will allow us to discuss these findings in relation to (in)formality.

4. Results and discussion

Section 4.1 provides a quantitative overview of the L2SCA measures and how they pattern across registers and in the learner writing, which allows us to further investigate claims about level of formality made in previous studies. Section 4.2 presents the results of a more detailed comparison between the expert academic writers and the learners using TAASSC and manual linguistic investigation.

4.1. Predicting and assessing formality

To help us make sense of the measures and their relative importance, we fitted a Random Forest (RF) on the data. RFs are built from averaged predictions of a large set of Conditional Inference Trees (CITs). CITs belong to the family of recursive partitioning methods, where a series of splits are made to achieve minimal impurity in the terminal nodes (see Levshina, forthcoming, for a more detailed description). The results of the RF (mtry: 3; ntree: 4000) display the relative importance of each of the syntactic measures for predicting register (and, by extension, formality), as can be seen in Fig. 2.

The factors are plotted in descending order of importance; the dotted line represents the cut-off point for statistical significance at the 0.05 level. The x-axis displays the conditional variable importance. The most important predictors of register in descending order are complex nominals per clause (CN.C), complex nominals per T-unit (CN.T), mean length of T-unit (MLT), mean length of sentence (MLS), number of coordinated phrases per T-unit (CP.T), mean length of clause (MLC), number of T-units per sentence (T.S) and number of coordinate clauses per clause (CP.C). To assess how well the model fitted the data, the prediction accuracy was computed using a prediction matrix (cf. Levshina, 2015, p. 299), which showed that the model had a high prediction accuracy: 0.87.

The eight most important measures fall into three main categories in Lu’s (2017) classification: degree of phrasal sophistication (CN.T and CN.C), length of production unit (MLT, MLC, MLS) and amount of coordination (CP.T, CP.C and T.S).
remaining two categories, *amount of subordination* and *overall sentence complexity*, thus proved to be of less importance for predicting register.

Not surprisingly, a closer look at these eight measures using pairwise comparisons for each category showed that measures based on T-units vs. clauses were strongly correlated, as shown in Figs. 3 and 4 below displaying CN, T vs. CN, C and CP, T and CP, C, respectively. The three measures of length of production unit (MLT, MLC and MLS) were also highly correlated (correlation coefficients: 0.80–0.97; *p* < 0.001).4

Due to the high degree of correlation, we thought it preferable not to include all eight measures in the subsequent analysis, but instead use one measure to represent each category. As the T-unit-based measures performed slightly better overall, the following measures were chosen for further investigation: *complex nominals per T-unit*, *mean length of T-unit* and *number of coordinate phrases per T-unit*.

The complex nominals measure includes nouns plus adjectives, possessives, prepositional phrases, relative clauses, participles, or appositives; it also includes nominal clauses and gerunds and infinitives when found in subject position (Lu, 2010, p. 483). Mean length of T-unit, is, as the name suggests, a measure of the length (in number of words) of a main clause and any subordinate or non-clausal structures attached to it (Lu, 2010, p. 482). Number of coordinated phrases per T-unit measures the number of coordinated adjectives, adverbs, nouns and verb phrases (Lu, 2010, p. 483). See Lu (2010, 2017) for a more technical description of how the measures are calculated. Three example sentences can be found in (1)–(3) below, along with the corresponding measure.

(1) At the outer rim of awareness, the known and declared nationalists fade into a host that have still to declare themselves. (Acad_humanities.ANT) [3 complex nominals; 1 T-unit = 3 complex nominals per T-unit]

---

4 In addition to the correlations noted between measures based on T-units vs. clauses and between measures of length of production unit, we found strong correlations between many of the other measures as well, thus supporting the points discussed in Section 2.1 relating to measure redundancy. This fact also meant that we opted for Random Forests (RF) rather than regression analysis, as inclusion of correlated variables does not violate any assumptions of this test (cf. Levshina, forthcoming).
(2) Colours are apprehended as spatially extended. (Acad_humanities.CK1) [6 words; 1 T-unit = T-unit length: 6 words]

(3) [W]e do know the main activities which generate the greatest growth and demand for space and we can identify current and potential conflicts with other land use activities. (Acad_social_sci.FR2) [2 coordinated phrases; 2 T-units = 1 coordinate phrase per T-unit]

Clear register-specific differences emerged when these three measures were mapped onto the registers. For all three measures, the medians were significantly higher in the a priori most formal register, academic prose, than in the least formal registers, news and fiction (the conditional inference trees testing the differences for statistical significance can be found in Appendix A and the dispersion across individual texts can be found in Appendix B). That is, when we move along the scale from less formal to more formal, we generally get increasingly higher scores for the measures of complex nominals, T-units and coordinate phrases.

After having established that the measures investigated are correlated with formality, we added the learner writing to the model to see which of the BNC registers the learners’ texts are most similar to, and thereby where on the formal-informal cline they can be placed. Following Wulff, Gries, and Lester (2018), we make use of CITs to see how the subcorpora group for each of the measures. Figs. 5–7 below show the distribution per measure, as detailed below.

As can be seen in Fig. 5, the scores for complex nominals per T-unit in fiction are significantly lower than in the other registers (the first split). There is also a statistically significant difference between the expert academic writers and the learners on the one hand, and popular science and news on the other (the second split). In the final split, the learners branch off from academic prose. It thus seems that learners’ use is somewhere in-between academic prose, popular science and news in terms of frequency of complex nominals, suggesting somewhat lacking register awareness on the part of the learners. By contrast, with regard to coordinated phrases per T-units (Fig. 6), the learners’ usage does not differ significantly from that of the expert academic writers. For the final measure, mean length T-units (Fig. 7), the learners again group with the expert academic writers, suggesting relatively good register awareness.

We can thus conclude that while the learners’ frequencies are most similar to those of the expert academic writers for two of the three measures investigated here, some differences can still be noted with regard to the measure of complex nominals, which would merit further attention to this measure. However, as can be recalled from the beginning of this section, the complex nominal measure is very inclusive in that it covers several different kinds of pre-modification and post-modification in addition to gerunds and nominal clauses in subject position, thus suffering from several of the limitations discussed in Biber et al. (2011) and Biber et al. (2020). Nonetheless, as the present study is exploratory in nature, we found L2SCA to be a useful first step of the analysis to help us narrow down the field of possible linguistic structures to look more closely at, in an attempt to get an overview of the distributional tendencies across registers — albeit, admittedly, a coarse-grained one. The subsequent subsection provides a more detailed study of what kind(s) of NP modification is responsible for the differences between the learners’ and expert academic writing. First, however, we report on the findings of a Random Forest that is fitted on the learner and expert academic texts to ascertain that the complex nominals measure is, in fact, the most interesting measure to look more closely at even when we limit the analysis to academic prose.
Fig. 5. Complex nominals per T-unit.

Fig. 6. Coordinated phrases per T-unit.
4.2. Learner vs. expert academic writing

In order to investigate what lies behind the differences noted for the measure of complex nominals in the learner and expert academic writing, analyses at three different levels were applied: (i) a Random Forest was fitted exclusively on the texts written by the learners and the expert academic writers, thus allowing us to return to all 14 measures to assess the relative importance of the complex nominal measures for academic prose only; (ii) the differences between the groups was subsequently investigated more closely using TAASSC (Kyle, 2016); and (iii) a manual analysis of the texts was carried out to gain a more complete picture.

The Random Forest fitted with all 14 measures confirms that the most important measures for distinguishing between the learners’ and the experts’ academic writing are, indeed, complex nominals per clause (CN.C) and complex nominals per T-unit (CN.T), thereby once again highlighting the importance of complex nominals as a measure worthy of more detailed analysis in the learner and expert data. The output of the Random Forest (mtry: 3; ntree: 4000; prediction accuracy: 0.88) is shown in Fig. 8.

However, unlike what was the case in the Random Forest that was used to distinguish between the BNC-15 registers, there is a clear difference between the clause-based measure and the T-unit-based one for complex nominals, with the clause-based measure being the most important predictor. As can be seen from Fig. 9 below, the difference results from lower scores on average in the learner data. The figure displays the dispersion across all the texts, with each data point marked by a black dot. The mean is marked by a red square and the median is marked by the vertical line in the middle of the box. The box itself shows the inter-quartile range; x-es above the plot mark outliers.

Fig. 7. Mean length of T-units.

Fig. 8. Random Forest showing the relative importance of the 14 measures for distinguishing between the learner and expert academic writing.
The question as to why there is such a large difference between the clause-based measure of complex nominals and the T-unit based one may seem puzzling at first, especially as these measures were highly correlated in our investigations of the whole data set. However, it would appear as if the answer lies in one important difference between the measures that proved to be central when comparing the learners and the experts: for a T-unit-based measure to show a high average score for complex nominals, it is enough for just one clause (dependent or independent) per T-unit to have a high score for complex nominals. By contrast, for the average number of complex nominals per clause to be high, there would have to be a high score for complex nominals in each clause. It thus seems as if the complex nominals are more evenly spread out across the clauses in the expert data than in the learner data.

As the complex nominal measure includes many different types of modifiers, TAASSC (Kyle, 2016) was used to further investigate what differed between the learners’ and experts’ academic writing. A Random Forest (mtry: 3; ntree: 4000; prediction accuracy: 0.89) of these measures is displayed in Fig. 10.

The results show that the main differences between the learners and the experts are to do with their uses of adjectives and prepositions inside the NPs. As shown in Fig. 11 below, these differences resulted from an underuse in the learner data. These results are thus largely in line with those of Ansarifar et al. (2018) and Staples and Reppen (2016) where the importance of adjectival and prepositional modifiers was emphasized. The differences between the groups with regard to their use of possessives, determiners, nouns as nominal dependents, relative clauses (RC) and the conjunction ‘and’ all proved to be significant, unlike verbal and adverbial modifiers and the use of the conjunction ‘or’.

One could certainly argue that a less dense style of writing is oftentimes more accessible. However, adherence not only to stylistic requirements such as word limits, but also to general expectations of the research community is required for an article to be published (cf. Hyland, 2008). Since frequent use of prepositional phrases has been described as one of the defining characteristic of academic writing in previous studies (e.g. Biber, 1988; Biber et al., 2011), it would seem important for novice writers wishing to start an academic career to be made aware of these more or less explicit standards and expectations. Novice writers, in particular learners, might therefore benefit from some targeted instruction on how to use a denser style of writing, thereby allowing them to meet the expectations of the research community and have their disciplinary writing be recognized as such.

5 Barring that, fewer dependent clauses would also bring scores for clause-based measures closer to those of T-unit-based measures; however, no statistically significant differences related to this were found.
Indeed, when we took a closer look at text excerpts from texts that scored high vs. low on adjectival modification and prepositions per noun phrase, it became clear that the quantitative differences extended to more qualitative ones as well. Whereas the expert writers used a dense, fact-based style of writing, many of the learners tended to adopt a more narrative style, even for analyses of results, which is a subsection often associated with condensed linguistic structures. In more detail, the expert writers primarily used adjectival and prepositional modification to add more detail to technical descriptions of methods or results. An example can be found below, where the adjectival modifiers are marked in bold and prepositional phrases that modify a noun are underlined.

Hence at present three main kinds of problems are encountered in the use of existing systems: first, the generally inappropriate design of GUIs for GIS purposes; second, the specification of the language of interaction for spatial operations and queries; and, third, the limited content and rigid structure of standard help systems. It is therefore suggested that the design of a user interface must be rooted in the creation of a user environment which integrates the data-modelling process, the availability of spatial operations and queries, and the system of assistance and concept support within the interface.

While there certainly were texts in the learner data that scored relatively high on adjectival modification and prepositions per noun phrase too, a large proportion of the texts included simpler noun phrases without adjectival or prepositional modification. The writers of the remaining texts seem to have relied on other strategies to be able to add sufficiently detailed technical descriptions, such as adverbial clauses, which led to lower textual density. This is exemplified in the following excerpt from the analysis section of a linguistics thesis where the learner is applying incongruity theory to study the relation between conflict and humor.

When children argue, it seems typical for them to make threats that they will not follow through on, but here Agnes manages to make a threat and go through with it. She decides to hold her breath until she gets her way. This is her way of disagreeing with Gru, who says he cannot fix her toy. Even if it is not verbal, Margo steps in for her, explaining what she is doing and that Agnes will hold her breath until he compromises or submits to Agnes’ so called argument. This is a threat that would be difficult to follow through with, but since it is a movie, it is of course very possible; as well as exaggerated, which makes it humorous.

As can be seen, the lack of NP modification makes the text less dense than what is traditionally to be expected from expert academic writing. The style of writing arguably comes across as more narrative than technical, and more words are required to reach the intended rhetorical goals. It can also be noted in passing that whereas this learner hardly uses adjectival or prepositional modification, there are several instances of relative clauses in the text excerpt. Relative clauses are listed as towards the least compressed end of Biber and Gray’s (2016, p. 207) scale of the structural density of NPs, whereas prepositional and adjectival modification are found at the most compressed end of the scale, thus offering more evidence to support the claim that the learners could benefit from some explicit teaching of NP structure in academic prose.
5. Concluding discussion

The present study has explored the extent to which measures of syntactic complexity can be used to predict and describe (in)formality across four registers from the BNC-15 and in learner writing. As complex nominals proved to be a particularly important measure, we used a selection of measures from TAASSC (Kyle, 2016) along with manual analysis to look more closely at NP complexity in the learner and expert academic writing.

The results from the first part of the study showed that while the learner texts bear some resemblance to the expert academic writing (e.g. both groups make use of comparatively long T-units), they also exhibit features that are associated with the non-academic registers (e.g. comparatively lower scores for the automatic measure of ‘complex nominals’), which could offer some support for the claim in previous studies that learners tend to be somewhat informal in their writing. As much of the learners’ input in a foreign language comes from non-academic registers such as news, fiction and different online registers, it is perhaps not surprising that their output, while at a very high level, is not fully register-appropriate. Whereas it does not make sense, of course, to ask students to use longer T-units to make their writing sufficiently formal, studies like ours can hopefully be used to raise awareness of the more subtle differences across registers, thereby helping students improve their writing. In the second part, the more detailed analysis of NP complexity in the expert and learner academic writing showed that the learners tended to underuse adjectival and prepositional modification compared to the expert writers, and instead rely on clausal modification. This seems to point to a need for more detailed instruction of how to move from a more narrative style of writing to a more economic and dense style in order to meet the requirements and expectations of the research community.

Although it was not an aim of the study itself, using these automated tools for complexity detection enabled us to informally evaluate their usefulness for linguistic description. On the plus side, both tools, which have generously been made freely available by their respective creators, are easy to use and facilitate large-scale investigations of complexity of great use for the field. Tools such as these allow linguists like us (who lack sufficiently advanced knowledge of computer programming to write similar programs of our own) to investigate linguistic complexity in large corpora, thus adding to the diversity of studies published on the topic. Furthermore, computers can generally be trusted to have good intra-rater reliability, which is likely to lead to more trustworthy comparisons between the subcorpora investigated than any human analyst would achieve. However, it is important to keep in mind that there are some unavoidable downsides to using these kinds of tools. For example, using such tools ultimately entails employing a “black box” method, where the analyst is forced to trust the developer’s (grammatical) judgement, as no direct link between the input texts and the output figures is provided. Moreover, as became clear to us during the course of this project, the main downside to using L2SCA for the purpose of our study was the fact that the categories provided by the tool confounded a large number of grammatical structures (cf. Biber et al., 2020), thus requiring further analysis to learn exactly what was causing the differences found. Researchers wishing to use this tool for purposes similar to ours would therefore be advised to keep in mind that further, complementary analyses (and/or linguistically more informative measures) are needed in order to gain better understanding of the data. In our case, these steps involved incorporating more detailed complexity measures from TAASSC and complementary manual investigations of the files that ranked high and low respectively on the different measures to interpret the results linguistically.

Through this study, we hope to have contributed a piece of the puzzle with regard to how syntactic complexity relates to formality and register awareness in learner writing. However, as the picture is still far from complete, there are numerous avenues for future research. For example, as we have primarily taken a quantitative approach, more fine-grained, qualitative analyses, perhaps also taking spoken data into consideration, would serve to complement the present analysis. It would also be interesting to investigate register awareness in relation to proficiency, to see to what extent native-speaker status and/or level of proficiency might affect syntactic choices made. In addition, also including an investigation of lexical and lexico-grammatical complexity would serve to add to previous research in the field and enrich the analysis. Finally, due to the corpora selected for the present study, we have not been able to carry out investigations of possible disciplinary variation or differences across different first-language backgrounds, which is likely to prove fruitful.

All in all, we believe strongly in the importance of treating (in)formality as a cline rather than a dichotomy, and we hope to have shown how a register analysis can be used to concretize this cline. In doing so, we also hope to have contributed to a more nuanced view, one that will allow us to move away from the categorical perception of (in)formality in learner writing that sometimes permeates writing instruction, which will hopefully benefit both L2 instruction and theory.

CRediT authorship contribution statement

**Tove Larsson:** Project administration, Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing. **Henrik Kaatari:** Visualization, Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing.

Acknowledgements

We are grateful to the two anonymous reviewers for their valuable comments and suggestions. We would also like to thank Doug Biber for his very helpful comments on an earlier draft of this article.
Appendix A

Conditional inference trees showing the groupings of the BNC-15 registers for the three top predictors.

Fig. A1. Complex nominals per T-unit.

Fig. A2. Mean length T-unit (in number of words).

Fig. A3. Coordinated phrases per T-units.

Appendix B

The dispersion across the individual texts can be found in Figures B1–B3 below.
Fig. B1. Boxplots showing the dispersion, mean and median of each register for the three top predictors.
Tove Larsson received her PhD from Uppsala University in Sweden. Her research interests include analysis and applications of learner corpora, register variation, lexico-grammar and English for Academic Purposes; she also has a keen interest in research methods. She is currently a visiting researcher at Northern Arizona University; she is also affiliated with Uppsala University and UCLouvain.

Henrik Kaatari received his PhD from Uppsala University in Sweden. His research interests include syntactic/grammatical variation, complexity and learner corpus research. He is currently senior lecturer in English linguistics at the University of Gävle in Sweden.