DE GRUYTER
OPEN

ICAME Journal, Volume 42, 2018, DOI: 10.1515/icame-2018-0003

Is there a correlation between form and function? A syntactic and functional investigation of the introductory *it* pattern in student writing[1]

Tove Larsson, Uppsala University and Université catholique de Louvain[2]

Abstract

The introductory it *pattern, as in 'It is important to note that information was added', is a tool used by academic writers for a range of different rhetorical and information-structural purposes. It is thus an important pattern for students to learn. Since previous research on student writing has indicated that there seems to be a correlation between form and function of the pattern, the present study sets out to investigate this more systematically in non-native-speaker and native-speaker student writing in two disciplines (linguistics and literature). In doing so, the study adds to and extends previous research looking into factors such as NS status and discipline. It uses data from three corpora: ALEC, BAWE and MICUSP. The results show that there is indeed a correlation between form and function, as the most common syntactic types of the pattern each display a preferred function and vice versa. While very few differences across NS status were found, there were certain discipline-specific disparities. The findings, which could be useful for teaching students about the use of the introductory* it *pattern, also have implications for the automatized functional tagging of parsed corpora.*

1 Introduction

It is often said that scholars construct an academic identity through their writing, using both conventional and non-conventional forms of expression to project different voices (Thompson 2009: 53). While individuality is of importance, academic writers must largely adhere to the discourse conventions of their fields and disciplines in order for their findings to be recognized (Hyland 2008a: 3). The use of formulaic language and linguistic patterns offers writers an opportunity to structure their arguments or position themselves in relation to a claim in an appropriate manner that is familiar to their discourse community. One such linguistic pattern that is commonly used for these purposes is the introductory *it* pattern, as exemplified in (1)–(3) below. Investigation of the use of this pattern in academic discourse is the focus of the present study.

(1) […] *it* is interesting *to consider the three metaphors as a rhetorical sequence* […]. (ALEC_LING.075)

(2) […] *it* could be the case *that the issues addressed are so oblique, so illusive that you cannot approach them gradually* […]. (ALEC_LING.011)

(3) […] *it* appears *that DeLillo's perspective on subjectivity has shifted traumatically* […]. (ALEC_LIT.037)

The introductory *it* pattern (also referred to as *subject extraposition*) is here defined as a pattern that has two subjects: introductory *it*, which does not have anaphoric reference, and a clausal subject. The pattern is described in more detail in Section 2.2.

The introductory *it* pattern is commonly used in academic discourse (Zhang 2015; see also Biber et al. 1999: 722), and its functional diversity has been emphasized in many studies (e.g. Kaltenböck 2005). It thus stands out as an important pattern for apprentice academic writers to master. However, previous research has found that apprentice learners in particular tend to struggle with how to use the pattern (e.g. Hewings and Hewings 2002; Römer 2009). For example, with regard to its functional distribution, learners have been reported to have a tendency to overuse the pattern to emphasize the validity of claims (e.g. *it is clear that…*) and underuse the pattern to hedge claims (e.g. *it might be that…*) compared to NS-student and expert writers (Hewings and Hewings 2002; Larsson 2017). Learners have also been found to struggle with making appropriate use of high-frequency syntactic realizations of the pattern, such as subject-verb-complement (SVC: *it is important to…*) and subject-verb (SV: *it seems that…*) (Larsson forthcoming; see Quirk et al. 1985: 1392 and Section 3.2.2 for an overview of the syntactic types). While indications of a form-function correlation have been noted in previous studies (e.g. Römer 2009; Ädel 2014), including my own (e.g. Larsson 2016: 33), these studies left some questions unanswered.

One such unanswered question is to what extent the functional and syntactic distribution of the pattern might be linked, as is put forth by theories such as Pattern Grammar (Hunston and Francis 2000). No previous study has, to the best of my knowledge, investigated this systematically for the pattern using inferential statistics. In order to shed light on these research issues, the present study aims to investigate the interaction between the functional and syntactic characteristics of the pattern in academic writing by non-native-speaker (NNS) and native-speaker (NS) students. Put in another way, the present study investigates whether it is the case that each function has a preferred syntactic type and vice versa. Further knowledge about whether this is the case would benefit both English for Academic Purposes teaching and theory, as such information would enable more targeted teaching of the pattern. As clear cross-disciplinary differences have been noted for both the form and function of the pattern across linguistics and literature (Larsson 2017; Larsson forthcoming), this study also includes discipline as a factor in the analysis, in addition to NS status. The following research questions were used:

- Is there a correlation between the function and the syntactic form of the pattern?
- What differences and similarities can be found when comparing the NNS and NS corpora?
- Are there any disciplinary differences (linguistics vs. literature)?

2 *The introductory* it *pattern and previous research*

In this section, the introductory *it* pattern is defined and delimitations made are discussed in Section 2.1. Previous studies of relevance to the present article are presented in Section 2.2.

2.1 Definition and delimitations

As noted in Section 1, the introductory *it* pattern, as in (4), is here defined as a pattern which contains two subjects: an introductory *it* (which does not have anaphoric reference) and a nominal clause. The two subjects are italicized in the example and will be discussed below.

(4) *It* is obvious *that a lot of thought went into that aspect* […] (ALEC_LIT.126)

The first subject, the introductory pronoun *it* is described as supplying "the structural requirement for an initial subject" (Quirk et al. 1985: 89). As such, introductory *it* thus does not carry much information in itself; however, it is not completely empty of meaning, as it has cataphoric reference to the clausal subject (Quirk et al. 1985: 349). The introductory pronoun *it* differs from the *it* used in *it*-clefts (5), 'prop' *it* (6) or *it* with anaphoric reference (7); all of these are excluded from the analysis.

(5) *It* is Anna who echoes this behaviour […]. (ALEC_LIT.050)

(6) […] *it* is raining right now […]. (ALEC_LIT.037)

(7) […] *it* is reported by informants to have closer ties with the traditional structure […]. (ALEC_LING.011)

The second subject in an instance of the introductory *it* pattern is made up of a nominal clause. Using Quirk et al.'s (1985: 1047ff) terminology, nominal clauses are included in the category of subordinate clauses, along with three other types that are not covered by the definition: *adverbial clauses*, *relative clauses* and *comparative clauses*. There are six subtypes of nominal clauses: *that*-clauses (8), subordinate interrogative clauses (9), subordinate exclamative clauses (10), nominal relative clauses (11), nominal *-ing* participle clauses (12) and *to*-infinitive clauses (13). However, no instances of subordinate exclamative or nominal relative clauses were found in the data. The four remaining subtypes were included in the analysis. Instances of the pattern with a *to*-infinitive clause also include the *for*/*to* construction, as in (14). Quirk et al. (1985: 1061) note that "[t]he presence of a subject in a *to*-infinitive clause normally requires the presence of a preceding *for*".

(8) *It* is possible *that the respondents have answered the questions in the questionnaire* […]. (ALEC_LING.053)

(9) Interestingly, *it* is not clear *how Huddleston & Pullum want to account for these patterns*. (ALEC_LING.083)

(10) *It*'s incredible *how fast she can run*. (Example taken from Quirk et al. 1985: 1055)

(11) Macy's is *where I buy my clothes*. (Example taken from Quirk et al. 1985: 1056)

(12) *It* is more problematic *making a distinction between the terms 'dialect' and 'accent'*. (ALEC_LING.082)

(13) […] *it* might be difficult *to determine at what point a word form becomes lexicalized* […]. (ALEC_LING.029)

(14) *It* might be easier *for the students to see the integrative use of English right away* […]. (ALEC_LING.003)

Constructions that did not have a nominal clause were excluded from the analysis; such constructions include those in which an introductory *it* is followed by an adverbial clause, as in (15) and tokens with nominal extraposition (i.e. where the *it* refers to an NP), as in (16).

(15) *It* is a pity *if teachers do not use this awareness*. (ALEC_LING.078)

(16) *It*'s staggering *the number of books that can pile up*. (Example taken from Michaelis and Lambrecht 1994: 362)

The definition used for the present article is largely in keeping with Quirk et al.'s (1985: 1391ff) definition of what is referred to as *subject extraposition*,[3] although with one main exception: Quirk et al. (1985: 1391) state that extraposition operates "almost exclusively" on subordinate nominal clauses. The definition used in the present study is thus slightly more exclusive, as *only* subordinate nominal clauses are allowed as the second subject here. Furthermore, while the definition used here is very similar to that of Quirk et al. (1985: 1391) where *structure* is concerned, there is an important *conceptual* difference between the two. This has to do with the meaning of the term *extraposition*.

When referred to as (*subject*) *extraposition*, the introductory *it* pattern is commonly discussed in relation to a non-extraposed construction (see, e.g., Miller 2001; Herriman 2013). In more detail, Quirk et al. (1985: 1391) describe extraposition as being derived from sentences with "more orthodox ordering", i.e. from their non-extraposed equivalents. The constructed non-extraposed equivalent to example (4), repeated here as (17), is given below as (18).

(17) *It* is very important *to remain unbiased* [...]. (ALEC_LING.105)

(18) *To remain unbiased* is very important.

The extraposed clausal subject is said to have been "moved to the end of the sentence", with the introductory *it* filling its original slot (Quirk et al. 1985: 1391).

However, there are three main reasons why this conceptualization of the pattern is problematic. First, the use of 'derive' and 'movement' implies movement of the clausal subject in an underlying structure, which is a conceptualization that is arguably closer to a generativist view of grammar than to a descriptive, empirically-based view of grammar (the latter being the approach taken for the present study). Second, viewing non-extraposition as the canonical construction is problematic considering that the extraposed constructions have been found to be significantly more frequent than non-extraposed ones in the present study, as well as in previous studies (e.g. Mair 1990: 30–31; Huddleston and Pullum 2002: 969; 1402; Mukherjee 2006: 348–349; Mindt 2011: 31). Third, this view also makes it difficult to account for the group of tokens for which extraposition is obligatory. An example of such a token is given in (19), along with its constructed (and ungrammatical) non-extraposed equivalent in (20).

(19) *It* seems *that he can read Jane* […]. (ALEC_LIT.108)

(20) **That he can read Jane* seems.

While tokens with obligatory extraposition are included in the category of extraposition in Quirk et al. (1985: 1183, 1392[a]) along with a discussion of how there is no non-extraposed counterpart,[4] the view of extraposition as resulting from movement of the clausal subject from pre-predicate position does not provide a satisfactory explanation for this group of tokens.

In light of these points of criticism, the present study has adopted an approach to this construction where no claims are made about there being movement of sentence constituents. The term *extraposition* is therefore not used; instead, the construction is referred to as the *introductory* it *pattern*, as mentioned above (see, e.g., Hunston and Francis 2000; Groom 2005). Other terms used in previous studies to refer to the pattern include the *anticipatory* it *pattern* (e.g. Hyland 2008b; Ädel 2014), it-*clauses* (e.g. Hewings and Hewings 2002) and it-*extraposition* (e.g. Kaltenböck 2005; Zang 2015).

2.2 Previous studies

The use of the pattern has received a fair amount of attention in an English for Academic Purposes (EAP) context (e.g. Groom 2005; Peacock 2011; Zhang 2015), as well as in corpus studies of a more general character (e.g. Mair 1990; Ramhöj 2016); the focus will, however, be on studies in EAP here.

The use of the pattern has previously been investigated in the same material as is used for the present study with regard to its syntactic distribution and its functional distribution. The results of the syntactic analysis (Larsson forthcoming) showed that the same three syntactic types were the most frequent ones in all subcorpora, suggesting that the use of the pattern is stable across the points of comparison. These were Subject-Verb-Complement (e.g. *it is interesting to note*), Subject-Verb (e.g. *it seems that*) and Subject-Verb$_{\text{pass}}$ (e.g. *it has been noted that*). While there were no major differences across NS status, there were clear differences found between the two disciplines investigated. The article thus concluded that there appear to be discipline-specific conventions with regard to the pattern that students could benefit from being made aware of.

The results of the functional analysis (Larsson 2017) showed that the stance marking function (e.g. *it is interesting to*) is more than three times as frequent as the stance-neutral observations category (e.g. *it can be seen in table 1 that*), making the former the most important overarching function of the pattern. Furthermore, there were clear disciplinary differences. For example, the linguistics students made more frequent use of observations and attitude markers, in particular those expressing *difficulty* (e.g. *it is difficult to*), *expectation* (e.g. *it is surprising that*) and *importance* (e.g. *it is imperative that*). Thus, as was the case for the syntactic types, appropriate use of the pattern is important for academic writers wishing to adhere to discipline-specific conventions. There were also certain noteworthy differences across NS status. For example, the NS students made significantly more frequent use of the pattern to hedge claims, using realizations such as *it seems that* and *it appears that.*

While certain tendencies were noted in the above-mentioned studies that led to the current project to be initiated, these studies did not formally investigate whether there is a correlation between the form and function of the pattern. The pattern has, nonetheless, previously been given as an example of the inseparability of grammar and lexis. A tendency for the meaning of the matrix predicate and the type of clausal subject it occurs with (e.g. *to*-infinitive, *that* or *wh*-clause) to be correlated was found both in Herriman's (2000) study investigating the use of the pattern in the LOB corpus and in Zhang's (2015) study comparing academic writing to popular writing using the ICE-GB corpus. For example, matrix predicates expressing epistemic modality (e.g. *clear*, *likely*) were almost exclusively found to be followed by a *that*-clause (Herriman 2000: 593; see also Mair 1990: 25). In addition, Römer (2009: 159) found "clear associations" between form and function. Similar results were reported in Groom (2005) where a tendency for certain subpatterns to be more strongly associated with certain meaning groups was noted. For example, *difficulty* and *validity* were the most frequently occurring subcategories for *it is* ADJ *that*/*to* in both Groom's (2005) and Römer's (2009) studies.

Furthermore, Ädel (2014) investigated how instances of the introductory *it* pattern containing one of seven different adjectives (*interesting*, *important*, *possible*, *clear*, *evident*, *obvious* and *apparent*) were used for making 'rhetorical moves', such as *comment on specific findings* and *indicating areas for future research* in NNS and NS student writing. Ädel (2014: 78) noted that "most subpatterns were found to be specialised for one or a few rhetorical moves". However, it was also pointed out that there is no one-to-one correspondence between moves and subpatterns (Ädel 2014: 76), as witnessed by the fact that the most common rhetorical move, *comment on specific findings*, was not only found to be realized by all seven adjective patterns, but was also the most common move for six of them.

In the present study, which takes a macro-level approach, the introductory *it* pattern will be analyzed further in order to shed more light on the question of whether there is a correlation between the functional and syntactic categories of the pattern. Unlike previous studies, inferential statistics will also be applied to the results to see whether there is a correlation in the statistical sense.

3 Material and method

3.1 Corpora used in the study

The study uses data from three corpora: the *Advanced Learner English Corpus* (ALEC), the *Michigan Corpus of Upper-Level Student Papers* (MICUSP) and the *British Academic Written English* corpus (BAWE). In total, ALEC comprises 1.3 million words (146 texts) written by university students in English linguistics and English literature. The vast majority of the students in ALEC have Swedish as their first language (L1), but the corpus also includes some texts written by students with other L1s, such as English, Finnish and Spanish. The students were in their third through fifth year of university studies on average when they wrote the texts. MICUSP includes a total of 2.6 million words (approximately 830 texts) written by students at University of Michigan in the USA (Römer and O'Donnell 2011). It spans sixteen disciplines, including psychology, engineering and economics, and a range of different text types, such as proposals and argumentative essays. BAWE includes 6.5 million words (approximately 2,800 texts) from several British universities (Heuboeck et al. 2008). The texts were written by students in 35 different disciplines, ranging from history and philosophy to medicine and mathematics. The texts were divided into 13 different text types, such as case study, essay and research report.

In order to ensure comparability to as large an extent as possible, subsets of these three corpora were used. These subsets comprise texts written by students whose L1 is English (BAWE, MICUSP and ALEC) or Swedish (ALEC) and who are in their third or fourth year of linguistics or literature studies (on average). Upper and lower cut-off points were used to bring the mean number of words contributed by each student closer across the corpora; each student contributed between 2,000 and 15,000 words to the subcorpora (mean length approximately 6,000 words). In total, these articles included investigation of approximately 255,000 words of NS writing and 590,000 words of NNS writing. An overview of the size of the corpora can be found in Table 1.

Table 1: Number of words per subcorpus

Subcorpus	*Number of words*	*Number of texts*
BAWE (NS)	94,345	25
MICUSP (NS)	121,147	37
ALEC (NS)	39,786	4
ALEC (NNS)	587,829	69
Total	**843,107**	**135**

3.2 Method

3.2.1 Elicitation and data management

In order to find all instances of the introductory *it* pattern in the material, the lexical item <*it*> was searched for in all corpora using WordSmith Tools (Scott 2012). The hits were subsequently gone through manually to exclude all instances of the constructions described in Section 2.1. Many of the invalid tokens are very difficult (if not impossible) to distinguish from the valid tokens without manual investigation, as they are superficially similar. Examples of this include tokens with non-conditional (21) versus conditional use of an *if*-clause (22) and tokens where *it* has cataphoric reference to a nominal clause (23) versus tokens where *it* has anaphoric reference (24); despite the surface similarity, only the first instances in these two pairs are counted as valid tokens of the introductory *it* pattern, in accordance with the definition used.

(21) […] *it* can be questioned *if the results are beneficial for either group of students*. (ALEC_LING.084)

(22) *It* becomes even clearer *if in the passage the pronoun 'I' is substituted for 'he'* […]. (ALEC_LIT.059)

(23) […] *it* should also be acknowledged *that the different worldviews between the source and target contexts make it quite difficult* […]. (ALEC_LING.125)

(24) […] while in the past academic writing was categorized as impersonal and lacking in subjectivity, *it* is now widely acknowledged to be a dialogical genre […] (ALEC_LING.104)

The present approach differs from that of many previous studies in that it allows for a wide variety of valid instances of the introductory *it* pattern to be identified and included. For example, the approach enabled the inclusion of tokens with inverted word order (25) and tokens in which the complementizer *that* is omitted (26), which are difficult to find using search patterns, such as *it*+V+ADJ *to*/*that*/*wh*-clause. Being able to include tokens for which the complementizer *that* is omitted is especially important when looking at learner data, as *that* omission has been found to be common in these kinds of data (Biber and Reppen 1998: 155).

(25) Nor is *it* her fault *that she was not fed distinction with her mother's milk.* (ALEC_LIT.001)

(26) […] *it* seems *she is trying to say that she still suffers from them*. (ALEC_LIT.034)

The tokens were subsequently classified into syntactic and functional categories, as described in Sections 3.2.2 and 3.2.3. The software environment *R* (R Core Team 2017) was used to manage the data and test the results for statistical significance using a Multinomial log-linear model. The reason why such a model is needed is that all of the variables are categorical rather than numeric.

3.2.2 The syntactic classification

The data were categorized into syntactic categories using Quirk et al.'s (1985: 1392) classification. The seven syntactic types are listed below (see Larsson forthcoming, for a more detailed discussion).

i. Subject + Verb + Complement (SVC): *it is interesting to*
ii. Subject + Verb + (obligatory) Adverbial (SVA): *it is beyond the study to*
iii. Subject + Verb (SV): *it appears that*
iv. Subject + Verb + Object (SVO): *it involves manual investigation to*
v. Subject + Verb + Object + Complement (SVOC): *it makes him happy to*
vi. Subject + passive Verb (SV_{pass}): *it has been found that*
vii. Subject + passive Verb + Complement ($\text{SV}_{\text{pass}}\text{C}$): *it has been proven useful to*

Four groups of tokens deserve mention from the perspective of the syntactic classification. The first group of tokens to be considered here contains those instances that include BE + a past particle (e.g. *it is suggested that*). While the -*ed* form for many instances of this group of tokens is verbal, as in (27), there were also instances where it is adjectival, as in (28). The former would result in the tokens being classified as belonging to the SV$_{\text{pass}}$ category and the latter would result in the tokens being classified as SVC, as explained below.

(27) *It* could be argued *that we cannot entirely trust the narrator of 'The Lord of the Rings'* […]. (ALEC_LIT.018)

(28) *It* is generally accepted *that different genres show different degree of productivity.* (ALEC_LING.029)

In order to be able to distinguish between these, Quirk et al.'s (1985: 167–171) *passivity gradient* was used. It states that instances that meet the formal criteria (i.e. BE followed by a past participle) can be placed on a continuum where *central passives* and adjectival complements make up the end-points. Functional criteria are given to assist the differentiation process. Instances of the introductory *it* pattern were only classified as SV$_{\text{pass}}$ if they met both the formal and the functional criteria. For example, such instances were classified as SV$_{\text{pass}}$ if they could be paraphrased into an active sentence (not necessarily in the form of an introductory *it* pattern) and if they did not meet the criteria for semi passives or pseudo passives (i.e. if they cannot be modified by degree adverbs, be placed after copular verbs such as APPEAR and/or be coordinated with an adjective).

The second group that deserves mention consists of tokens such as *it is of interest that* and *it is of importance that*. In the present article, these were classified as SVC, rather than SVA. These tokens belong to a group that Quirk et al. (1985: 732) describe as being "best treated though gradience and multiple analysis", since they can be counted either as complements or as obligatory adjuncts. However, since they can be coordinated with adjective phrases that function as complements and be used as "complementation for copular verbs other than BE" (e.g. SEEM, APPEAR), like adjective phrases, but unlike prepositional phrases (Quirk et al. 1985: 732), these tokens are arguably most similar to the ones classified as belonging to the SVC type, and were therefore categorized as such. They are, moreover, semantically very similar to *it is interesting that* and *it is important that* respectively.

The third group includes SEEM+*to* and APPEAR+*to* tokens, such those exemplified in (29).

(29) It seems to be the case that the Swedish NNS are less aware than the NS […]. (ALEC_LING.019)

At first sight, such tokens appear to allow for dual classification. On the one hand, one could argue that the clausal subject of example (29) is *to be the case* […] *than the NS*, which would result in this token being classified as SV with obligatory extraposition, similarly to SEEM/APPEAR directly followed by a *that*-clause (e.g. *it seems that*). On the other hand, the clausal subject could instead be interpreted as being made up of the *that*-clause (*that the Swedish NNS* […] *than the NS*), which would result in this token being classified as SVC. While there thus seem to be two possible classifications for these tokens, the second classification stands out as the preferred one for two main reasons. First, the *that*-clause is a more plausible clausal subject semantically, as this is where the proposition is stated. Second, and more importantly, the *that*-clause is the most logical clausal subject syntactically. Quirk et al. (1985: 137) treat verbs such as SEEM and APPEAR followed by the infinitive marker *to* as *catenatives*. Catenative verb constructions are considered to be different from main verbs and take an intermediate position between main verbs and semi-auxiliaries (Quirk et al. 1985: 137). As the *to*-clause cannot be a subject if it is part of the VP, we get the following syntactic analysis for (29): $[\textit{it}]_{\text{intr.subj}}$ $[\textit{seems to be}]_{\text{Vgrp}}$ $[\textit{the case}]_{\text{complement}}$ $[\textit{that the Swedish NNS are less aware than the NS}]_{\text{clausal subject}}$. Such tokens were, thus, classified as SVC.

The final relatively small group includes tokens that can be seen as idioms (or fixed expressions), as in *it goes without saying that*. Tokens belonging to this group have been analyzed as syntactic strings rather than chunks; the example provided above is thus classified as SVA (Subject-Verb-obligatory Adverbial).

3.2.3 The functional classification

The concept of replicability is central to corpus linguistics studies. In an attempt to develop a classification that more readily yields reproducible results, the functional classification used in the present study has been designed with the aim of limiting the impact of subjective interpretation on the classification. The classification, which was developed in Larsson (2017), makes use of a feature-assigning system that allows the classifier to be less dependent on word semantics as a means of classifying the data. In this system, the features are assigned based mainly on linguistic evidence other than word semantics. The features are binary (+/-), marking either presence or absence of a hedge (+/-H), affective attitude (+/-A) and an emphatic (+/-E) for each token. The features can be combined, or all set to minus; examples include *it appears that* (+H-A-E), *it seems very difficult to* (+H+A+E) and *it is shown in table 2 that* (-H-A-E). There are eight different permutations possible, six of which were found in the data; these are shown in Table 2 below (see Larsson 2017 for a more detailed discussion).

Table 2: Permutations of the different features coded for

+/-H	+/-A	+/-E	Category name	Example
-H	+A	-E	Attitude marker (A)	*It is fascinating that*
-H	-A	+E	Emphatic (E)	[not attested]
-H	+A	+E	Emphatic attitude marker (EA)	*It is very important that*
+H	-A	-E	Hedge (H)	*It seems that*
+H	-A	+E	Hedged emphatic (HE)	[not attested]
+H	+A	-E	Hedged attitude marker (HA)	*It could be interesting to*
+H	+A	+E	Hedged emphatic attitude marker (HEA)	*It appears extremely easy to*
-H	-A	-E	Observation (O)	*It has been found that*

Attitude markers are used to express “the writer’s affective attitude towards what is stated in the clausal subject” (Larsson 2017: 61). *Emphatics* are used to “strengthen the force of the utterance” (Larsson 2017: 61; see also Hewings and Hewings 2002: 373). This category includes *amplifiers* (e.g. *extremely*, *highly*, *very*; see Quirk et al. 1985: 445) and modal auxiliaries expressing deontic modality (e.g. *need to*, *should*, *must*) (Larsson 2017: 61). *Hedges* express “possibility rather than certainty” and indicate “a lack of complete commitment to the truth of a proposition or [...] a desire not to express that commitment categorically” (Hyland 1996: 251). This category also includes modal verbs functioning as *downtoners* (e.g. *could*, *might*) (see Hewings and Hewings 2002: 370; Larsson 2017: 61).

Although the functional classification started out from preconceived categories, the categories have been refined based on the data. These categories were loosely based on those of Hewings and Hewings (2002), Groom (2005) and Herriman (2000). However, while these previous classifications are suitable for the aims of the projects they were developed for, there are two main drawbacks of these previous classifications from the perspective of the present article.

First, none of these classifications cover all the tokens that were included in the present study. For example, none of them include a category for SV$_{\text{pass}}$ tokens with text-organization purposes, such as *it has been shown in table 1 that*. Second, all three classifications discussed in the previous subsection rely heavily on word semantics. For example, in Hewings and Hewings’s (2002) classification, *it is important to* is counted as an attitude marker and *it is appar-*

ent that is counted as an emphatic, based on the word semantics of the adjectives *important* and *apparent*. Similarly, *it is incredible* is counted as epistemic modality in Herriman (2000), whereas *it is astonishing* is classified as evaluation. Relying heavily on word semantics is slightly problematic in two ways. First, unless the classifications allow for dual membership, they do not take polysemy into account (*incredible* can mean either 'hard to believe' or 'amazing'). The decision to place tokens such as *it is incredible* and *it is astonishing* into two different categories based solely on one meaning of *incredible* might be perceived as slightly arbitrary, which brings us to the second point, namely that a classification based heavily on word semantics can be seen as inherently subjective. Herriman (2000: 584) addresses these points when she gives the following caveat:

> There is no set of exhaustive, semantic categories in which meanings may be organised and there are no foolproof, clear-cut criteria by which semantic categories may be clearly distinguished from one another. Inevitably, then, the semantic classification has been based on my own subjective interpretation of the examples and a number of arbitrary decisions have had to be made.

While it is perhaps impossible to base a functional classification solely on objective criteria, and a functional classification necessarily entails a reliance on categories with fuzzy boundaries, the approach developed in Larsson (2017) arguably at least *increases* the replicability of the results. Furthermore, the present functional classification covers the range of functions performed by the tokens found in the data.

4 Results and discussion

In this section, the results of the investigation of possible form-function associations will be presented and discussed first more generally in Section 4.1 and then more specifically with respect to NS status and discipline in Section 4.2.

4.1 An investigation of form-function associations

A total of 1,610 tokens were found in the data. The distribution suggests that there is, indeed, a correlation between form and function. Table 3 shows the raw frequencies of the syntactic types for each function in all the data. The most frequent syntactic type per function is marked in gray; the most frequent function per syntactic type is marked in bold. As can be recalled, the abbreviations used in the table for the forms and functions are discussed in Sections 3.2.2 and 3.2.3 respectively.

Table 3: The raw frequencies of the syntactic types and functions in all the data

	SVC	*SVpass*	*SV*	*SVO*	*SVA*	*SVOC*	*SVpassC*	*OTHER*	Total
A	**711**	7	4	7	3	0	1	0	733
O	112	**209**	17	**12**	**17**	**1**	**7**	**2**	377
H	84	73	**160**	0	0	0	0	0	317
HA	123	6	1	0	1	0	1	0	132
AE	16	32	0	1	0	0	0	0	49
HAE	2	0	0	0	0	0	0	0	2
Total	1048	327	182	20	21	1	9	2	1610

As is shown, the three most frequent syntactic types map onto the three most frequent functions, which suggest that there are specialized form-function pairings. However, most of the forms and functions can be realized by several different categories; for example, the observations category can be realized by all the syntactic types, thereby showing that there is no one-to-one correspondence. Although Ädel (2014) used different classification systems than the ones used in the present study, these findings seem to concur with those of her study of the pattern, in the sense that a specialization rather than a perfect form-function mapping was found.

Further, as can also be seen in Table 3, while there are six functional categories and seven syntactic types[5], the three most frequent functional and syntactic types make up the bulk of the tokens. No less than 89 percent (1,427/1,610) of the tokens belonged to the three most frequently occurring functional categories, namely attitude markers (A), observations (O) and hedges (H), and 97 percent (1,557/1,610) of the tokens are realized through the SVC, SV or SV$_{\text{pass}}$ types. This subsection will therefore henceforth focus on these functional and syntactic categories.

A visual overview of the distribution of the functions across these syntactic categories can be found in Figure 1.

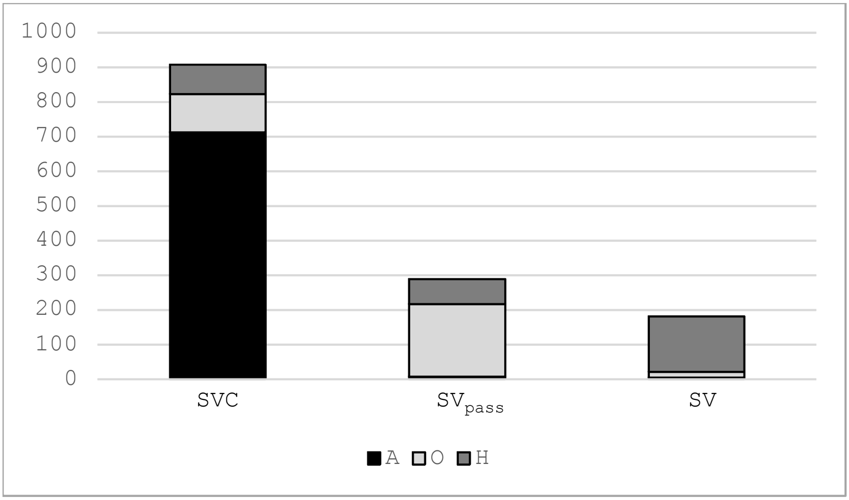

Figure 1: The functional distribution across the three most frequent syntactic categories

The pairings are exemplified in (30) (attitude markers – SVC), (31) (hedges – SV) and (32) (observations – SV$_{pass}$).

(30) […] *it* is important *to look at the refusals of proposal in the novels themselves*. (ALEC_LIT.127)

(31) […] *it* appears *that the verb does not raise to I* […]. (MICUSP_LING.202.1)

(32) […] *it* has been noted *that the extent of a language deficit depends on how much the brain has been damaged* […]. (BAWE_LING.6174d)

In order to test whether there is, indeed, a correlation between form and function of the pattern in the statistical sense, a multinomial log-linear model was fitted onto the data. As the *R* output of such models is extensive, effect plots will be used throughout the article to show the results of the models for clarity; the confidence intervals can, nonetheless, be found in Appendix 1. Figure 2 shows the probabilities for each syntactic type (i.e. the form) given a function.

The functions are displayed on the x-axis and the probability for each syntactic type, given one of the functions, is shown on the y-axis. The probabilities add up to one vertically. The *effects* package (Fox and Hong 2009) is used here to display the results; this package not only provides an overview, but it also displays the confidence intervals, which are marked by the whiskers (the vertical bars extending from the dots). The confidence intervals tell us how likely it is that the true value lies within the interval produced by the model. Since the number of tokens included is comparatively high, the model can predict the outcome relatively reliably, which results in narrow confidence intervals around the probabilities.

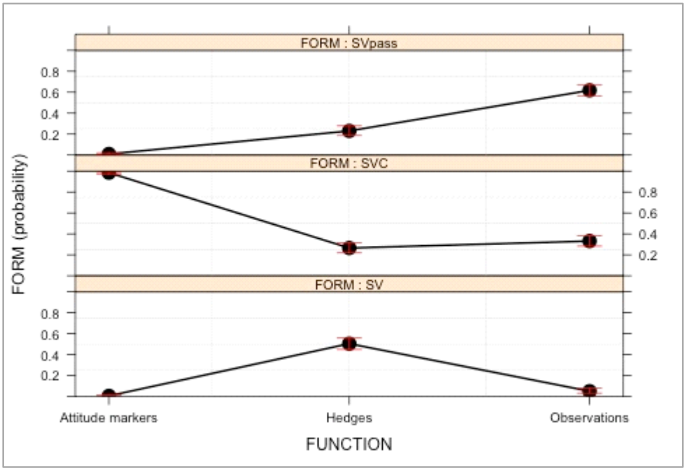

Figure 2: Effect plot showing the probability of each syntactic type given a function

The model shows that there is a correlation between the functional and syntactic categories, in the sense that the three most frequent functions each have a preferred form (model AIC: 1346.6). If we look at the attitude markers column, we can see that this functional category is almost exclusively realized by the SVC type (the likelihood for attitude markers to be realized as SVC is almost one, as can be seen in the middle section of the graph). It is furthermore significantly more likely that any given hedge is realized by the SV type than through any other syntactic type (as can be seen when comparing the bottom section of the graph to the other sections). Finally, the observations category is significantly more likely to be realized by the SV_{pass} type than by the other types (as can be seen when comparing the top section of the graph to the other sections). What this essentially means is that once a writer has decided to express a certain function, s/he is significantly more likely to use a realization that belongs to the corresponding syntactic type (attitude markers→SVC, hedges→SV and observations→SV_{pass}) than some other realization, assuming that the introductory *it* pattern is used. The function of the introductory *it* pattern can thus be used to predict the syntactic types with relatively high accuracy.

Conversely, form (i.e. the syntactic types) can also be used for predicting function (model AIC: 1765.3), as shown in Figure 3.

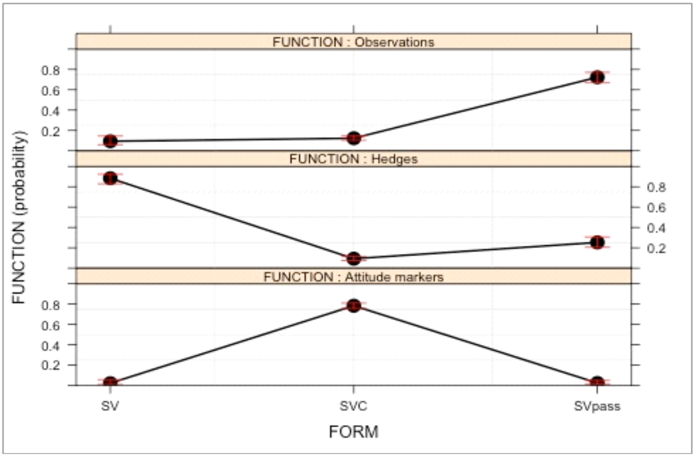

Figure 3: Effect plot showing the probability of each function given a syntactic type

When the probabilities for the SVC column are compared vertically across the different functions, an instance of the pattern that is realized by the SVC type is significantly more likely to be used as an attitude marker than for any other purpose. Similarly, any SV-type token is most likely to be used as a hedge, and any SV_{pass}-type token is most likely to be used to make observations. As can be seen, there is an even clearer difference between the probabilities for these predictions than for the ones presented in Figure 2. These findings could, for example, be used for automatized functional tagging of parsed corpora; however, the less frequent categories would, of course, have to be included in such analyses too.

4.2 A comparison across NS status and discipline

Let us now turn to the second and third research questions pertaining to potential differences and similarities across NS status and discipline. Since it is arguably more reasonable linguistically to expect function to precede form in actual usage, this subsection will focus on the question of which syntactic type is used, given a particular function, rather than the other way around.

With regard to whether there are differences across NS status, the answer appears to be negative. As is clear from Figure 4, there are only very minor differences across NS status, and the predictor itself is not statistically significant (model AIC: 1349.885). The results for the NNS students can be found in the left column and the NS students' results are shown in the right column.

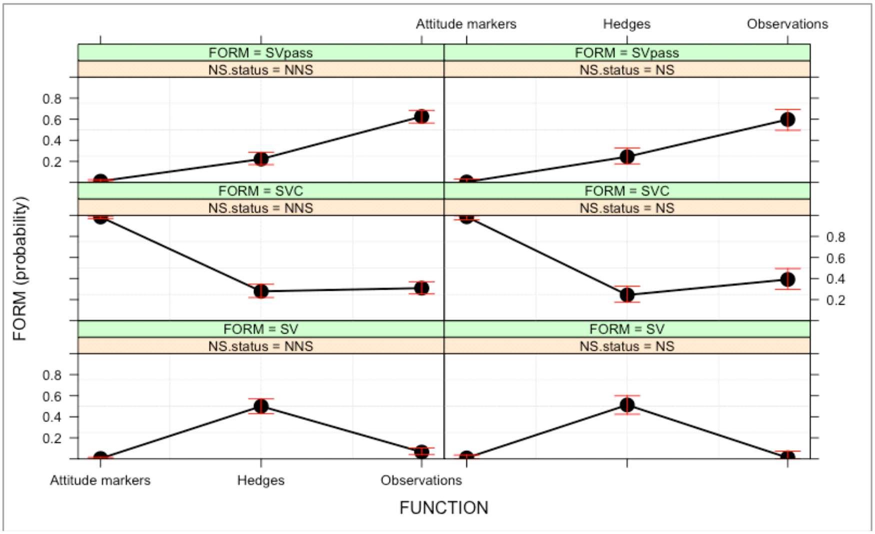

Figure 4: Effect plot showing the probability of each syntactic type given a function across NS status

Since no clear differences were noted across NS status, these results seem to be in line with Larsson (forthcoming), which investigated the frequency of the syntactic types. Interestingly, however, they diverge somewhat from those of Larsson (2017) looking at the functional distribution of the pattern where clear differences were found across NS status; the NS students were, for example, found to make significantly more frequent use of the pattern to hedge claims (Larsson 2017). What this means is that while the NNS students underuse the hedging function of the pattern overall (Larsson 2017), they use this function very similarly to the NS students structurally when they *do* use it.

For both groups, SV is the most likely realization, although it is not uncommon for a hedge to be realized by the other two types, SVC and SV$_{pass}$, as well.

Examples of hedges realized by the three different syntactic types can be found in (33)–(35) below.

(33) SV: *It* seems *that there are so many things happening at once* […]. (ALCE_LIT.058)

(34) SVC: *It* is possible *that they started out under a node within the CP* […]. (ALEC_LING.077)

(35) SV$_{pass}$: […] *it* might be argued *that she utilises music, albeit verbal, as part of that device* […]. (ALEC_LIT.016)

When it comes to the comparison across discipline, more notable differences were found, as shown in Figure 5. The results for the linguistics data is shown to the left, and the literature data can be found to the right.

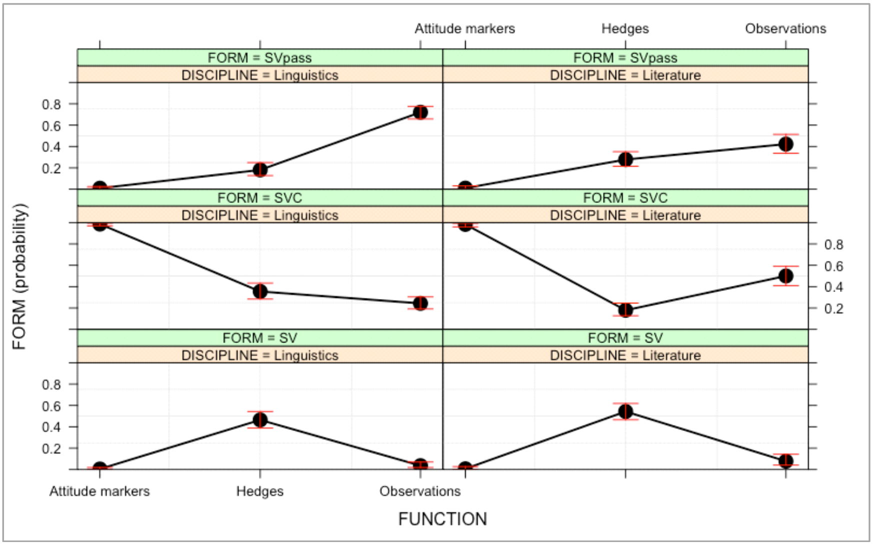

Figure 5: Effect plot showing the probability of each syntactic type given a function across discipline

Although discipline is not a significant predictor in the model overall (model AIC: 1316.352), usage differences, especially for observations, can still be noted. These can be seen most clearly when the first and second rows are compared. While the observations category is most likely to be realized by the SV$_{pass}$ type in the linguistics data, the SVC type is more or less an equally likely realization in the literature data. Examples from the disciplines can be found for literature and linguistics respectively in (36)–(39).

(36) SV$_{pass}$: […] *it* is revealed *that the vocal choices made in depicting the characters are only partly supported in the original text*. (ALEC_NS_LIT.010)

(37) SVC: […] *it* is politically unacceptable *for Satrapi to be the one who tells Americans the story* […]. (MICUSP_LIT.055.2)

(38) SV$_{pass}$: […] *it* is reported *that over 50 % of VLBW children need special assistance in school* […]. (BAWE_LING.6206b)

(39) SVC: *It* is common in Classical Tibetan *to reduce disyllabic words into a single syllable*. (ALEC_LING.125)

Although there were comparatively low frequencies of the observations category in the literature data, which could potentially explain the disparity, we can draw the tentative conclusion that the form-function correlation is discipline-specific, at least to a certain degree. This is in line with previous studies where somewhat differing discipline-specific conventions for the pattern have been noted for both form and function (Larsson 2017; Larsson forthcoming).

5 Conclusion

The present study has investigated whether there seems to be a correlation between the functional and syntactic categories of the introductory *it* pattern in NNS and NS student writing. It has also investigated potential differences and similarities across NS status and discipline (linguistics and literature). With regard to the question of whether form can be predicted based on the function and vice versa, the answer seems to be affirmative, even if there is no perfect one-to-one mapping. When the three most frequent forms and functions were investigated in more detail, this form-function correlation was confirmed. With some slight disciplinary disparities, the three most frequent functional and syntactic categories map onto one another, with attitude markers being most commonly realized by the SVC type (e.g. *it is interesting to note*), hedges by the SV type (e.g. *it appears that*) and observations by the SV$_{pass}$ type (e.g. *it has been found that*). There were no clear differences across NS status, which suggests that the learners are using the pattern in a native-like manner with regard to the pairings.

The fact that there seems to be a correlation between form and function provides valuable insight into how academic writers use the pattern that could be helpful for English for Academic Purposes theorists and practitioners. This fact could also be useful for automatized tagging of parsed corpora, which would increase time-efficiency, thereby allowing for larger data sets to be analyzed. It does, however, remain for future studies to examine the use of such programs. Further avenues for future research also include investigations of form-function pairings in expert writing and in more L1 varieties, across more disciplines.

Notes

1. The introduction and background sections of this article are largely based on the introductory survey chapter of my unpublished doctoral thesis (Larsson 2016); however, the results and conclusion presented in Sections 4 and 5 are new and were not included in the thesis.
2. At Université catholique de Louvain, I am a beneficiary of a "MOVE-IN Louvain" Incoming Post-doctoral Fellowship, co-funded by the Marie Curie Actions of the European Commission. I would also like to express my gratitude to Merja Kytö, Erik Smitterberg and the participants at a higher seminar at Uppsala University for their insightful comments on an earlier draft of this article. Finally, I would like to thank the anonymous reviewers for their valuable comments.
3. The study does not include an investigation of the related, but considerably less frequent, construction referred to as *object extraposition* (cf. Quirk et al. 1985: 1391f; Huddleston and Pullum 2002: 963), as in *the bard takes **it** upon himself **to sing*** [...] ***carols*** (ALEC_LIT.028). Subject extraposition is described as "the most important type of extraposition" (Quirk et al. 1985: 1391).
4. This group includes the following verbs: SEEM, APPEAR, CHANCE, HAPPEN, TRANSPIRE, COME ABOUT, TURN OUT. Following Quirk et al. (1985: 1213[note]), tokens such as *it strikes me that* are also included as an instance of the introductory it pattern.
5. Two tokens were found in the data that did not fit into Quirk et al.'s (1985) categorization, thereby adding two syntactic types to Quirk et al.'s seven types: $SV_{pass}A$ (*It is taken into account that the RP accent is quite difficult to define* (ALEC_LING.099)) and SVOA (*it takes Ashley more than a year to get all the way down to Chile* (ALEC_LIT.068)). These are, however, grouped together and classified as OTHER in the table for clarity.

References

Ädel, Annelie. 2014. Selecting quantitative data for qualitative analysis: A case study connecting a lexicogrammatical pattern to rhetorical moves. *Journal of English for Academic Purposes* 16: 68–80.

Advanced Learner English Corpus (ALEC). Corpus compiled in 2013.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Biber, Douglas and Randi Reppen. 1998. Comparing native and learner perspectives on English grammar: A study of complement clauses. In S. Granger (ed.). *Learner English on computer*, 145–158. London: Longman.

British Academic Written English (BAWE). Corpus compiled at the Universities of Warwick, Reading and Oxford Brookes in 2004–2007. http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/

Fox, John and Jangman Hong. 2009. Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects Package. *Journal of Statistical Software*, 32 (1): 1–24. URL: http://www.jstatsoft.org/v32/i01.

Groom, Nicholas. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4 (3): 257–277.

Herriman, Jennifer. 2013. The extraposition of clausal subjects in English and Swedish. In K. Aijmer and B. Altenberg (eds.). *Advances in corpus-based contrastive linguistics: Studies in honor of Stig Johansson*, 233–260. Amsterdam: John Benjamins.

Herriman, Jennifer. 2000. Extraposition in English*:* A study of the interaction between the matrix predicate and the type of extraposed clause. *English Studies* 81 (6): 582–599.

Heuboeck, Alois, Jasper Holmes and Hilary Nesi. 2008. *The BAWE corpus manual*. Available online from http://www.reading.ac.uk/internal/appling/bawe/BAWE.documentation.pdf, accessed on March 21, 2014.

Hewings, Martin and Ann Hewings. 2002. "It is interesting to note that…": A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes* 21 (4): 367–383.

Huddleston, Rodney D. and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Hunston, Susan and Gill Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hyland, Ken. 1996. Talking to the academy: Forms of hedging in science research articles. *Written Communication* 13 (2): 251–281.

Hyland, Ken. 2008a. Persuasion, interaction and the construction of knowledge: Representing self and others in research writing. *International Journal of English Studies* 8 (2): 1–23.

Hyland, Ken. 2008b. Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18 (1): 41–62.

Kaltenböck, Günter. 2005. *It*-extraposition in English: A functional view. *International Journal of Corpus Linguistics* 10 (2): 119–159.

Larsson, Tove. 2016. The introductory *it* pattern in academic writing by non-native-speaker students, native-speaker students and published writers: A corpus-based study. Unpublished doctoral dissertation, Department of English, Uppsala University, Sweden.

Larsson, Tove. 2017. A functional classification of the introductory *it* pattern: Investigating academic writing by non-native speaker and native-speaker students. *English for Specific Purposes* 48: 57–70.

Larsson, Tove. Forthcoming. A syntactic analysis of the introductory *it* pattern in non-native-speaker and native-speaker student writing. In M. Mahlberg and V. Wiegand (eds.). *Corpus linguistics, context and culture*. Berlin: De Gruyter Mouton.

Mair, Christian. 1990. *Infinitival complement clauses in English: A study of syntax in discourse*. Cambridge: Cambridge University Press.

Michaelis, Laura A. and Knud Lambrecht. 1994. On nominal extraposition: A constructional analysis. In K. E. Moore, D. A. Peterson and C. Wentum (eds.). *Proceedings of the twentieth annual meeting of the Berkeley Linguistics Society: General session dedicated to the contributions of Charles J. Fillmore*, 362–373. Berkeley: Berkeley Linguistics Society.

Michigan Corpus of Upper-level Student Papers (MICUSP). Ann Arbor, MI: The Regents of the University of Michigan. Corpus compiled at the University of Michigan in 2009. http://micusp.elicorpora.info/about-micusp

Miller, Philip H. 2001. Discourse constraints on (non)extraposition from subject in English. *Linguistics* 39 (4): 683–701.

Mindt, Ilka. 2011. *Adjective complementation: An empirical analysis of adjectives followed by* that-*clauses*. Amsterdam: John Benjamins.

Mukherjee, Joybrato. 2006. Corpus linguistics and English reference grammars. In A. Kehoe and A. Renouf (eds.). *The changing face of corpus linguistics*, 337–354. Amsterdam: Rodopi.

Peacock, Matthew. 2011. A comparative study of introductory *it* in research articles across eight disciplines. *International Journal of Corpus Linguistics* 16 (1): 72–100.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London, UK: Longman.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Ramhöj, Rickard. 2016. On clausal subjects and extraposition in the history of English. PhD dissertation: University of Gothenburg, 2016. Gothenburg.

Römer, Ute. 2009. The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics* 7 (1): 140–162.

Römer, Ute and Matthew B. O'Donnell. 2011. From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora* 6 (2): 159–177.

Scott, Mike. 2012. WordSmith Tools version 6 [Computer software]. Liverpool: Lexical Analysis Software.

Thompson, Paul. 2009. Shared disciplinary norms and individual traits in the writing of British undergraduates. In M. Gotti (ed.). *Commonality and individuality in academic discourse*, 53–82. Bern: Peter Lang.

Zhang, Guiping. 2015. *It is suggested that*…or *it is better to*…? Forms and meanings of subject *it*-extraposition in academic and popular writing. *Journal of English for Academic Purposes* 20: 1–13.

Appendix 1

Confidence intervals (confint) for the Multinomial log-linear models fitted.

```
Call: multinom(formula = FORM ~ FUNCTION, data=to)
```

```
> confint(model1)
,, SVC
                                       2.5%           97.5%
(Intercept)                        4.197643        6.163112
FUNCTIONHedges                    -6.842334       -4.807136
FUNCTIONObservations              -4.402356       -2.187827

,, SVpass
                                       2.5%           97.5%
(Intercept)                      -0.6688581      1.78808677
FUNCTIONHedges                   -2.6036048     -0.08505246
FUNCTIONObservations              0.6253121      3.27370257
```

```
Call: multinom(formula = FUNCTION ~ FORM, data=to)
```

```
> confint(model2)
,, Hedges
                                   2.5%             97.5%
(Intercept)                    2.696557        4.68048110
FORMSVC                       -6.841593       -4.80678048
FORMSVpass                    -2.603918       -0.08614569

,, Observations
                                   2.5%             97.5%
(Intercept)                   0.3569902          2.535113
FORMSVC                      -4.4013520         -2.187074
FORMSVpass                    0.6253769          3.273108
```

```
Call: multinom(formula = FORM ~ FUNCTION + NS.status + FUNCTION * NS.status, data =
to)
```

```
> confint(model3)
,, SVC
                                          2.5%          97.5%
(Intercept)                         4.13268635       6.909577
FUNCTIONHedges                     -7.53460827      -4.679073
FUNCTIONObservations               -5.45244093      -2.473572
NS.statusNS                        -2.82888853       1.103772
FUNCTIONHedges:NS.statusNS         -1.33483048       2.747434
FUNCTIONObservations:NS.statusNS    0.04112614       5.735616

,, SVpass
                                          2.5%          97.5%
(Intercept)                      -0.5019330828      2.6983165
FUNCTIONHedges                   -3.5516035940     -0.2717635
FUNCTIONObservations             -0.5147397846      2.8470646
NS.statusNS                      -4.6763341782      1.0933658
FUNCTIONHedges:NS.statusNS       -1.0763986003      4.8024858
FUNCTIONObservations:NS.statusNS -0.0007043501      7.0704334
```

```
Call: multinom(formula = FORM ~ FUNCTION + DISCIPLINE + FUNCTION*DISCIPLINE,
data = to)
```

```
> confint(model4)
,, SVC
                                                    2.5%         97.5%
(Intercept)                                      3.960996      6.732035
FUNCTIONHedges                                  -7.045124     -4.186548
FUNCTIONObservations                            -5.007822     -1.863581
DISCIPLINELiterature                            -2.334365      1.595522
FUNCTIONHedges:DISCIPLINELiterature             -2.511043      1.568333
FUNCTIONObservations:DISCIPLINELiterature       -1.893796      2.535975


,, SV_pass
                                                    2.5%         97.5%
(Intercept)                                     -1.0052945     2.3829160
FUNCTIONHedges                                  -3.3826078     0.1162672
FUNCTIONObservations                             0.4711224     4.1452297
DISCIPLINELiterature                            -2.7433029     2.1869271
FUNCTIONHedges:DISCIPLINELiterature             -1.9772189     3.0809815
FUNCTIONObservations:DISCIPLINELiterature       -3.6867997     1.6373085
```