Department of English

The introductory *it* pattern in academic writing by non-native-speaker students, native-speaker students and published writers

A corpus-based study

TOVE LARSSON

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Ihresalen, Thunbergsvägen 3, Uppsala, Saturday, 17 December 2016 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Randi Reppen (Northern Arizona University).

Abstract
Larsson, T. 2016. The introductory *it* pattern in academic writing by non-native-speaker students, native-speaker students and published writers. A corpus-based study. 85 pp. Uppsala: Department of English, Uppsala University. ISBN 978-91-506-2606-3.

The present compilation thesis investigates the use of a pattern that is commonly found in academic writing, namely the introductory *it* pattern (e.g. *it is interesting to note the difference*). The main aim is to shed further light on the formal and functional characteristics of the pattern in academic writing. When relevant, the thesis also investigates functionally related constructions. The focus is on learner use, but reference corpora of published writing and non-native-speaker student writing have also been utilized for comparison. The thesis encompasses an introductory survey (a "kappa") and four articles.

The material comes from six different corpora: ALEC, BATMAT, BAWE, LOCRA, MICUSP and VESPA. Factors such as native-speaker status, discipline, level of achievement (lower-graded vs. higher-graded texts) and level of expertise in academic writing are investigated in the articles. In more detail, Articles 1 and 2 examine the formal (syntactic) characteristics of the introductory *it* pattern. The pattern is studied using modified versions of two previous syntactic classifications. Articles 3 and 4 investigate the functional characteristics of the pattern. In Article 3, a functional classification is developed and used to categorize the instances. Article 4 examines the stance-marking function of the pattern in relation to functionally related constructions (e.g. stance adverbs such as *possibly* and stance noun + prepositional phrase combinations like *the possibility of*).

The introductory *it* pattern was found to be relatively invariable in the sense that a small set of formal and functional realizations made up the bulk of the tokens. The learners, especially those whose texts received a lower grade, made particularly frequent use of high-frequency realizations of the pattern. The thesis highlights the importance of not limiting investigations of this kind to comparisons across native-speaker status, as this is only one of the several factors that can influence the distribution. By exploring the potential importance of many different factors from both a formal and a functional perspective, the thesis paints a more complete picture of the introductory *it* pattern in academic writing, of use in, for instance, second-language instruction.

Keywords: The introductory it pattern, stance markers, non-native-speaker students, native-speaker students, published expert writing, learner language, corpus linguistics

Tove Larsson, Department of English, Box 527, Uppsala University, SE-75120 Uppsala, Sweden.

ISBN 978-91-506-2606-3
urn:nbn:se:uu:diva-305735 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-305735)

For my grandmothers, mormor Märtha and farmor Dagny, who never had the chance to study at university

List of papers

This thesis is based on the following papers, which are referred to in the text by their Arabic numerals.

1. Larsson, Tove (forthcoming). A syntactic analysis of the introductory *it* pattern in non-native-speaker and native-speaker student writing. In M. Mahlberg & V. Wiegand (Eds.), *Corpus linguistics, context and culture*. Berlin: De Gruyter Mouton.
2. Larsson, Tove (2016). The introductory *it* pattern: Variability explored in learner and expert writing. *Journal of English for Academic Purposes*, 22, 64–79.
3. Larsson, Tove (under review a). A functional classification of the introductory *it* pattern: Investigating academic writing by non-native-speaker and native-speaker students.
4. Larsson, Tove (under review b). *The importance of*, *it is important that* or *importantly*? The use of morphologically related stance markers in learner and expert writing.

Acknowledgements

Although it has certainly been stressful at times, I have thoroughly enjoyed my time as a PhD student, very much thanks to certain people to whom I would like to express my heartfelt gratitude.

First of all, I am deeply grateful to my main supervisor, Professor Merja Kytö. With all her experience, expertise, dedication and patience, Merja not only made the very abstract PhD process into something concrete, but also made it a rich learning experience. Her sharp editorial eye and helpful comments have been instrumental to the process. I want to thank Merja for always taking the time to discuss my project, and for being an unfailing optimist. Second, I would like to extend my warmest thanks to my secondary supervisor, Docent Erik Smitterberg. His kindness and encouragement have helped me immensely, and so have his excellent ideas and suggestions. With an outstanding eye for detail, Erik has provided me with invaluable feedback throughout the project.

I would also like to express my gratitude to the members of the higher seminar in linguistics. I feel very lucky to have been a PhD student in a department where everyone takes time out of their busy schedules to give such helpful feedback. Dr. Linnéa Anglemark, Erika Berglind Söderqvist, Dr. Irina Frisk, Dr. Gregory Garretson, Docent Christer Geisler, Dr. Angela Hoffman, Dr. Christine Johansson, Dr. Ewa Jonsson, Henrik Kaatari, Edward Long, Dr. Pia Norell, Göran Rönnerdal, Sarah Schwarz, Professor Terry Walker and Dr. Ying Wang – thank you all very much!

I would like to extend a special thanks to my officemates Sarah and Erika. Without their enthusiastic support and encouragement, being a PhD student would not have been anywhere near as enjoyable as it has been. I am also very grateful to Henrik for taking me under his wing when I had just started the PhD program and everything was new and confusing.

Moreover, I would like to thank my fellow PhD students Sally, Elisabeth, Leonard and Sindija for many interesting and rewarding conversations over numerous lunches. Thank you also to the rest of the PhD/postdoc community – Mike, Tasnim, Ryan, Kristen, Ola, Elsa and Julie – for creating a friendly and inspirational working environment. I am also very grateful to Ruth Hvidberg, Lóa Kristjánsdóttir and Dr. Åke Eriksson for all their help and guidance during my time at Uppsala University.

In addition, I would like to thank the following grant foundations whose generous support enabled me to travel to conferences, attend summer

schools and go on two research stays: *Anna Maria Lundins Resestipendier*, *Erik Tengstrands Stipendiestiftelse*, *Kungl. Humanistiska Vetenskaps-Samfundet i Uppsala* and *Petra and Karl Erik Hedborgs stiftelse.*

Furthermore, I would like to express my gratitude to some people outside Uppsala University. First, I would like to thank my former teachers at Stockholm University for introducing me to the joy of studying English. Second, a warm thank you to Professor Sylviane Granger and her team in Louvain-la-Neuve for their very helpful feedback during my research stay at the Centre for English Corpus Linguistics at Université catholique de Louvain. Thank you also very much to Dr. Signe-Anita Lindgrén and colleagues for their kind feedback during my research stay at Åbo Akademi University. Third, I am very grateful to the compilers of the corpora that I have very kindly been allowed to use: Dr. Magali Paquot (VESPA), Professor Sylviane Granger and colleagues (LOCRA), Professor Hilary Nesi and colleagues (BAWE), Dr. Ute Römer and colleagues (MICUSP) and Dr. Signe-Anita Lindgrén (BATMAT). Fourth, I would like to thank the students who contributed their texts to ALEC and VESPA.

Finally, I am immensely grateful to my friends and family who made sure that I also thought about things other than my dissertation. Thank you to Sara and Josefine for always being there for me; I cannot find the words to describe how important you both are to me. My heartfelt thanks also to Niamh, Roberto, Sol, Christer L, Elin, Åke and Nikolaus for widening my horizons (literally and figuratively) and for making my life in Uppsala such a pleasure. I am grateful to Daniel for always being willing to discuss life and for providing pertinent analyses. Thank you very much to Gunnel, Johanna, Filippa and Marika for stimulating pedagogical discussions. Furthermore, I am very grateful to my parents, Lisa and Stefan, for always believing in me and supporting me, without ever putting pressure on me. You are the best parents anyone could ever wish for. Thank you also to Charlie and Ingmarie and to my aunts and uncles, Sonja, Bosse, Anita, Björn and Lilian, for taking an interest in my thesis and my life in Uppsala. And last but not least: thank you, Gregory, for caring, for understanding and for always discussing topics great and small with equal passion.

Uppsala, October 2016

Tove Larsson

Contents

Abbreviations

AdjP	Adjective phrase
ALEC	Advanced Learner English Corpus
AmE	American English
BA	Bachelor's degree
BAWE	British Academic Written English
BNC	British National Corpus
BrE	British English
CIA	Contrastive Interlanguage Analysis
COCA	Corpus of Contemporary American English
EAP	English for Academic Purposes
ICE-GB	International Corpus of English (British component)
ICLE	International Corpus of Learner English
ILV	Interlanguage variety
L1	First language
L2	Second language
LCR	Learner Corpus Research
LOCRA	Louvain Corpus of Research Articles
MA	Master's degree
MICASE	Michigan Corpus of Academic Spoken English
MICUSP	Michigan Corpus of Upper-level Student Papers
NNS	Non-native speaker
NP	Noun phrase
NS	Native speaker
POS	Part of speech
RLV	Reference language variety
SEPC	Swedish-English Parallel Corpus
SV	Subject + Verb
SVA	Subject + Verb + obligatory Adverbial
SVC	Subject + Verb + Complement
SVO	Subject + Verb + Object
SVOA	Subject + Verb + Object + obligatory Adverbial
SVOC	Subject + Verb + Object + Complement
SV_{pass}	Subject + Passive verb
$SV_{pass}A$	Subject + Passive verb + obligatory Adverbial
$SV_{pass}C$	Subject + Passive verb + Complement
VESPA	Varieties of English for Specific Purposes dAtabase

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 project.

(1) […] *it* is interesting *to consider the three metaphors as a rhetorical sequence* […]. (ALEC_LING.3.075)
(2) […] *it* could be the case *that the issues addressed are so oblique, so illusive that you cannot approach them gradually* […]. (ALEC_LING.3.011)
(3) […] *it* appears *that DeLillo's perspective on subjectivity has shifted traumatically* […]. (ALEC_LIT.3.037)

The introductory *it* pattern (also referred to as *subject extraposition*) is here defined as a pattern that has two subjects: introductory *it*,[1] which does not have anaphoric reference, and a clausal subject. The pattern is described in more detail in section 5.

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 & Hewings, 2002; Römer, 2009).[2] The present thesis sets out to further investigate the use of

[1] Other terms that are sometimes used are 'anticipatory *it'* and 'preparatory *it*' (e.g. Swan, 2005:446–447).
[2] As is commonly the case in studies of this kind, the term *learners* is used to refer to non-native speaker students. Following Scott and Tribble (2006:133), *apprentice writers* are

the pattern, in particular in learner academic writing. However, unlike many previous studies which have focused on tokens containing an adjective phrase (as in Römer, 2009) or tokens that have an interpersonal function (as in Hewings & Hewings, 2002), the approach taken in the present thesis allows for a much broader range of tokens to be investigated. The thesis can thereby yield a more comprehensive description of the use of the pattern in academic discourse than has been provided in previous work.

In addition to the present introductory survey (the "kappa"), the thesis encompasses four articles (referred to as Articles 1–4 and summarized in section 7), each contributing to painting a more complete picture of the introductory *it* pattern and its use in academic discourse. The thesis has focused on two academic disciplines that are commonly housed in the same university department in a European context, namely (English) linguistics and literature, sometimes contrasting the two and other times focusing on one of them. Due to the diversity of the material, the present format of a compilation thesis was deemed especially suitable. The thesis makes use of qualitative and quantitative corpus methods as well as of inferential statistics. Furthermore, methodological approaches such as *Contrastive Interlanguage Analysis* (Granger, 2015) and *Pattern Grammar* (Hunston & Francis, 2000) have informed the thesis.

The aim of this "kappa" is to provide an overview of and a background to the thesis, thereby tying the articles together. In the subsection below, the overall aim of the thesis is presented first, after which the more detailed aim for each article is addressed. The more specific research questions found in each of the articles will not be repeated here, however.

1.1 Aims

The overall aim of the present thesis is to investigate the use of the introductory *it* pattern in academic writing, and thereby gain insights into its formal and functional characteristics; when relevant, the pattern will also be examined in relation to other, related constructions. While reference corpora consisting of expert writing and native-speaker student writing have been used for comparison, the main focus is on learner use of the pattern. An overview of the more detailed (and, to a certain degree, overlapping) aims of each of the four articles will now follow.

Articles 1 and 2 contribute to the overall aim by studying the formal, i.e. syntactic, characteristics of the pattern. The pattern is analyzed using modified versions of Quirk et al.'s (1985:1392) syntactic classification and the COBUILD grammar classification (Francis et al., 1996, 1998) respectively

viewed to be the authors of "unpublished pieces of writing that have been written in educational or training settings". The term apprentice is also used in Römer (2009).

(cf. section 6.1). Thus, while the first two articles build on existing classifications, they extend the classifications and use them to empirically investigate the use of the pattern across several different parameters.

In more detail, Article 1 provides an overview of the syntactic characteristics of the pattern in non-native-speaker (NNS) and native-speaker (NS) student writing, and has two main aims. First, it aims to investigate what constitutes the full inventory of the pattern, and does so by quantifying the syntactic types described by Quirk et al. (1985:1392). Second, it aims to examine in what way(s) the use of the pattern is affected by the three factors *NS status* (NS vs. NNS student writing), *academic discipline* (linguistics vs. literature) and *level of achievement* (higher-graded papers vs. lower-graded papers). Although the first factor has been explored in previous studies on the pattern, the material used in the present thesis enables investigation of more comparable groups than previous material has allowed for. The second and third factors have received very limited attention in the literature on learner writing.

However, while the general syntactic examination in the first article provides an important overview, it cannot satisfactorily answer questions about the lexico-grammatical make-up of the introductory *it* pattern. Article 2 therefore starts out from the more fine-grained syntactic classification presented in the COBUILD grammars (Francis et al., 1996, 1998). This article uses apprentice learner writing and published expert writing and has two main aims. First, the article considers the *degree of variability* of the pattern in the data; that is, it investigates to what extent the different subpatterns that make up the pattern are fixed or variable. Second, it explores whether the variability and/or frequency of realizations of the pattern found in expert writing can also be found in learner writing.

Through Articles 3 and 4, the complementary functional aspect is brought into the thesis. Whereas Articles 1 and 2 investigate the *syntactic* (and, to a certain extent, lexical) variability of the pattern, Article 3 aims to develop a classification that captures the *functional* heterogeneity of the pattern in NS and NNS student writing. It subsequently uses this classification to map out the functional distribution across the same factors as were explored in Article 1: NS status, academic disciplines and levels of achievement.

Article 4 builds on Article 3 and focuses on the function of the pattern that the third article found to be most frequent, namely stance marking. It uses linguistics texts written by experts and apprentice learners to investigate not only the introductory *it* pattern, but also constructions that are functionally very similar to the pattern. In doing so, Article 4 widens the scope in order to investigate what alternatives there are to the pattern. In more detail, the aim of this article is to examine what influences the use of a certain grammatical realization of stance in learner and expert writing and thereby to detect problem areas for learners. The material consists of two subcorpora of learner data: one with Swedish texts and one with Belgian French texts. The

grammatical realizations considered come from Biber et al.'s (1999) framework of stance and comprise the stance complement clause construction (e.g. *the possibility that this will happen*), stance adverbs (e.g. *importantly*) and stance noun + prepositional phrase (e.g. *the possibility of rain*). The article also carries out a more detailed analysis of two syntactically and semantically very similar constructions, the introductory *it* pattern (e.g. *it is surprising that*) and disjuncts (e.g. *surprisingly*).

1.2 Outline

The present "kappa" is structured as follows: the background of the thesis, including previous research and relevant methodological approaches, will be presented in sections 2 and 3. After that, the material is presented in section 4. A more detailed discussion of the introductory *it* pattern and relevant related constructions is provided in section 5, after which the classification schemes used in the four articles are discussed in section 6. The four articles that form the basis of this project are summarized in section 7. Finally, an overview of the results, teaching implications resulting from the investigations and some concluding remarks are given in section 8.

2 Previous research

The present thesis project, which is situated in the subfield of learner corpus research (LCR), makes use of several different corpora (see section 4.1) to investigate the use of the introductory *it* pattern and related constructions in learner corpora, as well as in expert and NS reference corpora. This section presents previous research carried out on the use of the introductory *it* pattern (section 2.1), as well as on its most frequent function: stance marking (section 2.2).

2.1 Previous research on the introductory *it* pattern

In this subsection, previous studies on the introductory *it* pattern (whether as the main focus or an ancillary focus) will be presented. Studies of a more general character will be discussed in section 2.1.1, in order to provide a background to the thesis. Studies on apprentice use of the pattern, which are of more immediate relevance to the thesis, will be addressed in section 2.1.2.

2.1.1 General studies

Many previous studies, especially outside an English for Academic Purposes (EAP) context, have referred to the introductory *it* pattern as (*subject*) *extraposition*; the reasons why the latter term is not used in the present thesis will be discussed further in section 5.1. The term *extraposition*, albeit with a slightly different meaning, was first used by Jespersen (e.g. Jespersen, 1927:356–357; Jespersen, 1937:45–46), but received more detailed attention and was given a more specific definition in transformational studies of syntax in the 1960s, for example in Rosenbaum (1967) (Seppänen, 1999:51; Seppänen & Herriman, 2002:30; Miller, 2001:683). Many studies since then have discussed the status and/or the use of the construction that is here referred to as the introductory *it* pattern. Although the introductory *it* pattern is sometimes proscribed in style guides (cf., e.g. *The Bedford Handbook*, Hacker & Sommers, 2014), with the claim that it makes writing unnecessarily wordy, it is still very common, in particular in academic writing (e.g. Zhang, 2015; see also Biber et al., 1999:722). While it is beyond the scope of the present thesis to investigate prescriptive stance, the possible influence

of prescriptive norms on expert writers' use of the pattern is discussed briefly in Article 4, section 3.2.2.

Linguists in a wide variety of subfields have investigated the use of the pattern. In fact, several different frameworks have been applied, such as Pattern Grammar (e.g. Groom, 2005), Construction Grammar (Mak, 2005) and Lexical Functional Grammar (e.g. Ramhöj, 2016; cf. also Kaltenböck, 2005). Furthermore, while some of these studies are firmly grounded in a corpus-based tradition (e.g. Herriman, 2013; Kaatari, forthcoming),[3] others make use of other methods (e.g. Seppänen, 1999). The focus of the present subsection will, however, be on the corpus-based studies.

The use of the pattern has received a fair amount of attention in an 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). There is a certain degree of overlap with regard to research focus between the two; three such foci will be addressed here. First, a number of studies have established that there seems to be an association between the matrix predicate and the clause type; i.e. matrix predicates tend to co-occur with a certain type of clausal subject (e.g. *likely* and a *that*- clause in *it is likely that*, or *easy* and a *to*-infinitive clause in *it is easy to*) (Collins, 1994; Dixon, 1991; Herriman, 2000; Mair, 1990; Quirk et al., 1985:1225[b]; Zhang, 2015:10).

Second, other studies have found that some instantiations of the pattern can be placed towards the "fixed" end of a continuum of level of variability. For example, in studies taking a lexical-bundles approach it has been found that certain instances of the pattern (e.g. *it is clear that*; *it is interesting to note*; *it is easy to*) show up among the top high-frequency four-word, five-word and/or six-word bundles in multi-genre corpora (Biber et al., 1999:1019–1020) or corpora of academic writing (Hyland, 2008b:48; Pan et al., 2016).[4] In light of Biber et al.'s findings, Rowley-Jolivet & Carter-Thomas (2005:51) note that "[e]xtraposition is a frequent structure in the RA [research article], where it regularly occurs as semi-formulaic 'lexical bundles'".

Third, a large number of studies have established that there appear to be clear differences between the contexts in which the introductory *it* pattern is used and the contexts in which the corresponding non-extraposed construction is utilized (*it is important to sing* vs. *to sing is important*) (e.g. Mukherjee, 2006; Mair, 1990; Ramhöj, 2016). Among other things, such studies have argued for the importance of *information packaging* (cf. Biber et al.,

[3] Although most of these studies have investigated the use of the pattern in written corpora, the introductory *it* pattern has also been examined in spoken data (see, e.g., Calude, 2008; Couper-Kuhlen & Thompson, 2008).

[4] While Biber et al. (1999) investigated all three types of lexical bundles, Hyland (2008b) and Pan et al. (2016) restricted their analysis to 4-word bundles. The results of Hyland's study will be discussed in more detail in section 2.1.2 in relation to apprentice writing.

1999:42) when explaining this distribution; using the introductory *it* pattern instead of a non-extraposed construction enables writers to adhere to the principles of *end-weight* (cf. Quirk et al., 1985:1361f) and *end-focus* (cf. Quirk et al., 1985:1357). These principles state that new information requires a heavier structure and that elements with high information value come towards the end of a sentence (Herriman, 2013:237–238; see also Huddleston & Pullum, 2002:1403; Rowley-Jolivet & Carter-Thomas, 2005:43). Using the introductory *it* pattern thus increases the ease of processing compared to sentences with non-extraposition (Huddleston & Pullum, 2002:1405–1406; see also Ramhöj, 2016), as the latter retain the "heavy" clausal subject, thereby violating these principles.

Furthermore, studies of direct relevance to an EAP context have investigated the possible influence on the use of the pattern of factors such as mode (speech vs. writing) (e.g. Rowley-Jolivet & Carter-Thomas, 2005), genre (e.g. Kaltenböck, 2005; Zhang, 2015), discipline (e.g. Biber et al., 1999; Groom, 2005; Peacock, 2011; Hyland & Tse, 2005) and NS status (e.g. Herriman, 2013).[5] The relevant findings of these studies will be summarized briefly below. EAP studies of apprentice writing have also investigated similar factors; these studies will be presented in more detail in section 2.1.2.

Differences across modes (spoken vs. written material) were found in Kaltenböck's (2005) study using material from the British component of the *International Corpus of English* (ICE-GB), as well as in Rowley-Jolivet & Carter-Thomas' (2005) study looking at conference proceedings and conference presentations by NNS and NS scientists. In Kaltenböck's (2005) study, mode-specific (as well as genre-specific) functional differences were noted. For example, instances of the pattern containing *given information*[6] in the extraposed clause were more than twice as common in the spoken mode as in the written mode (Kaltenböck, 2005:128–129). Rowley-Jolivet & Carter-Thomas (2005:46) found the pattern to be more frequent overall in the written than in the spoken material. Furthermore, clear differences across NS status were noted: the authors concluded that while "NS writers and speakers appear to adapt their information packaging strategies in response to the genre, this is not necessarily the case for NNS" (Rowley-Jolivet & Carter-Thomas, 2005:60).

With regard to differences across genres, further functional and frequency-based differences were found in Zhang (2015), Biber et al. (1999) and Groom (2005). Zhang (2015:10–11) investigated the introductory *it* pattern

[5] Some studies (e.g. Groom, 2005; Herriman, 2000; Hewings & Hewings, 2002) have also developed functional classifications to further investigate the use of the introductory *it* pattern. These classifications along with the syntactic classifications outlined in Quirk et al. (1985:1392) and Francis et al. (1996, 1998) respectively will be presented in greater detail in section 6.

[6] *Given information* denotes information that is retrievable "from the preceding co(n)text" (Kaltenböck, 2005:126).

using the academic and popular writing subcorpora of ICE-GB and found that the pattern was more frequent in academic writing than in popular writing.[7] While Biber et al.'s (1999) study is not as firmly based in an EAP context, the findings are still relevant to the present thesis and will be included here. Among other things, Biber et al. (1999:674, 722) looked at the overall frequency of the introductory *it* pattern in a large corpus comprising conversation, fiction, news and academic prose. They found that instances of the pattern with *to*-infinitive or *that*-clauses as the clausal subject were moderately common to common in academic writing (Biber et al., 1999:674, 722); this was especially the case for those instances that included an adjective and a *to*-infinitive clause (e.g. *it is interesting to*), as these were more frequent in academic prose than in any other genre (Biber et al., 1999:722). Further differences were noted in Groom's (2005:266) study comparing research articles and book reviews. He found that subpatterns such as ***it is* LIKELY/OBVIOUS *that*** were more frequent in research articles than in book reviews, whereas subpatterns such as ***it is* EASY/DIFFICULT *to*-inf** were prevalent in book reviews. Groom furthermore investigated the use of the pattern across two disciplines – history and literary theory – and found clear differences between these as well. The findings led him to conclude that the pattern (with *to*-infinitive or *that*-clauses) varies "in systematic ways across the four corpora studied" (Groom, 2005:272).

Other studies looking at the use of the pattern across disciplines include Peacock (2011) and Hyland & Tse (2005). In Peacock's (2011) study, instances of the pattern with *to*-infinitive or *that*-clauses were investigated in research articles from eight disciplines: biology, chemistry, physics, environmental science, business, language and linguistics, law, and public and social administration. As was the case in Groom's (2005) study, clear interdisciplinary differences were found. For example, the students in the "hard sciences" were found to use the pattern statistically significantly less frequently than the students in the "soft sciences" (Peacock, 2011:82); the latter group furthermore made use of a wider range of different forms of the pattern (Peacock, 2011:86). When it comes to individual disciplines, the pattern was particularly frequent in language and linguistics (Peacock, 2011:82). Hyland & Tse (2005:40–41) looked at "evaluative *that*-structures" (i.e. matrix clause [evaluation] + *that*-clause [evaluated entity]), as in *we believe that* or *it is unclear that* in research-article abstracts from six disciplines: applied linguistics, biology, business studies, computer science, electrical engineering, and public administration. In general, the authors found clear differences across the disciplines for both form and function of the evaluative *that* structures. Furthermore, researchers in the hard sciences were found to make more frequent use of the verb *show*, whereas researchers in the soft sciences

[7] It should be noted, however, that Zhang's (2015:4–5) conclusions are based on analyses of very few tokens (158 and 183 tokens for popular and academic writing respectively).

predominantly used more tentative verbs, such as *suggest* (Hyland & Tse, 2005:58). It was also noted that up to 20 percent of all the valid tokens were instances of the introductory *it* pattern; applied linguistics showed the second highest frequencies after electrical engineering (Hyland & Tse, 2005:54).

Some further differences have been identified with regard to NNS (L1 Swedish)[8] vs. NS users of the pattern. Herriman (2013) looked at a corpus of translations, the English-Swedish Parallel Corpus (ESPC), and observed that while there can be certain differences in use, instances of the introductory *it* pattern are "translationally equivalent in English and Swedish" (Herriman, 2013:236); that is, the constructions exist in both languages. Examples include *it is important to* and *it is likely that* (in Swedish: *det är viktigt att* and *det är troligt att* respectively) (Herriman, 2013:237). The pattern was also found to be more frequently used in Swedish than in English (Herriman, 2013:256).

In sum, it can be concluded from a review of the literature that factors such as mode, genre, discipline and NS status are important to take into account in the present thesis, as they have all been found to affect the use of the pattern. For this reason, two measures were taken. First, the thesis controls for mode by only including writing. Second, the material is sampled so that any genre differences are minimized (see section 4.2). The two remaining factors – discipline and NS status – will be further explored in the present thesis. We will now turn to previous studies of more direct relevance to the thesis, which investigate the possible impact of these two (and other) factors in apprentice writing.

2.1.2 Apprentice writing

There has been increasing interest in corpus-based approaches to the study of second language (L2) writing in recent decades. In fact, since learner corpora started being used in the late 1980s (Granger, 2015:7), a large number of studies have been carried out on a wide range of topics, such as lexis (especially phraseology; see, e.g., Paquot & Granger, 2012; Nesselhauf, 2005), grammar (e.g. Gilquin, 2002) and discourse (e.g. Müller, 2005). Several studies have looked at the use of the introductory *it* pattern in NS and NNS apprentice writing. While some have compared learners and expert writers (Hewings & Hewings, 2002, 2004; Hyland, 2008b) or learners and NS students (Boström Aronsson, 2005; Ädel, 2014; Ädel & Erman, 2012), others have looked at learners, experts *and* NS students (Hasselgård, 2009; Römer, 2009) or at NS apprentice writing across disciplines (Thompson, 2009; Charles, 2000, 2006). These studies will be presented in more detail below.

One of the first corpus studies of the use of the introductory *it* pattern in learner writing is Hewings & Hewings (2002; see also 2004). In this study,

[8] L1 stands for *first language*.

the authors compared published journal articles from the field of Business Studies to MBA theses written by learner writers of English (with various L1s). Although the material used in the study is relatively limited,[9] the authors made an important contribution to the field by outlining a corpus-driven functional classification of the instances of the introductory *it* pattern that have an interpersonal function (see section 6.2.1), such as tokens used for "commenting on, evaluating, or hedging the following clause" (Hewings & Hewings, 2002:372).[10] The authors found that the pattern sometimes causes problems for the learners, as they tended to overstate the validity of their claims (Hewings & Hewings, 2002:378) or use the pattern infelicitously (Hewings & Hewings, 2002:369). Furthermore, compared to the experts, the learners were not only found to make more frequent use of the pattern overall, but also to use all functional categories except for hedges (e.g. *it seems that*) more frequently (Hewings & Hewings, 2002:374).

Although the main focus of Hyland's (2008b) study was not the introductory *it* pattern, the study offers findings of clear relevance to the present thesis. Hyland took a similar approach with regard to the types of texts included for analysis (published journal articles and MA/PhD theses) as Hewings & Hewings; however, he had a broader scope with regard to the number of different patterns studied, as well as with regard to the number of different disciplines included (business studies, electrical engineering, applied linguistics and microbiology). Taking a lexical-bundles approach, Hyland (2008b) compared writing by published writers to theses written mainly by L1 Cantonese students to see which 4-word clusters were most frequent across the different groups. He found that the introductory *it* pattern made up no less than six percent of all the 4-word clusters (n=12,000 tokens) (Hyland, 2008b:48); these fixed instances of the pattern thus seem to be important in academic writing (cf. also Biber et al., 1999:1019–1020). He also noted that both the MA students and the PhD students made more frequent use of the pattern than the experts (Hyland, 2008b:53). Similarly, in Ädel & Erman's (2012:87) study looking at four-word lexical bundles in learner and NS student writing, the learners were found to use the pattern "to a greater extent" than the NS students.

The extent to which learners use the pattern in a native-like manner was investigated in two studies looking at NNS and NS student writing: Boström Aronsson (2005) and Ädel (2014). As was the case in Hyland's (2008b) study, the learners (L1 Swedish) were found to overuse the pattern in Bos-

[9] The study included investigation of 15 student theses and 28 journal articles; the total number of words included in the corpus used was approximately 330,000 (Hewings & Hewings, 2002:371).

[10] The authors thereby leave out tokens with "a predominantly ideational function", such as *it is possible to* (Hewings & Hewings, 2002:371); they also exclude tokens with "a text-organising purpose", such as *it was pointed out in the table that* (Hewings & Hewings, 2002:372).

tröm Aronsson's (2005) study, although this time in comparison to NS apprentice writers, rather than expert writers. The authors stated that L1 transfer could explain this difference, as the pattern was found to occur very frequently in Swedish (Boström Aronsson, 2005). Ädel (2014) qualitatively examined a selection of high-frequency subpatterns of the introductory *it* pattern that were overused, underused and used equally frequently by NNS students in relation to NS students.[11] In more detail, she explored what rhetorical moves (e.g. *indicating future research* and *commenting on method*) these instances of the introductory *it* pattern performed. Linguistics essays written by learners in Sweden and NS students who were in their second and third year of university studies were compared. In general, the learners were found to use the rhetorical moves similarly to the NS students (Ädel, 2014:76); however, for instances of the pattern such as those including the adjectives *important* or *clear*, the learners exhibited a comparatively broader repertoire of moves, which could indicate that the learners' usage was "a bit 'all over the place'" and that they were not using the pattern in a fully target-like manner (Ädel, 2014:76).

Römer (2009) and Hasselgård (2009) examined learner writing (L1 German and L1 Norwegian respectively), NS student writing *and* published expert writing, thereby identifying further differences and similarities across NS status and level of expertise with regard to the use of the pattern. In Hasselgård's (2009) extensive study of stance marking, instances of the pattern with *that*-, *to*-, *wh*- and -*ing*-clauses were included. She compared argumentative writing and conversation from corpora of learner writing and NS student writing; she also used a translation corpus of English and Norwegian original texts and their translations. Using Herriman's (2000) functional classification (cf. section 6.2.1), Hasselgård (2009:130) found that the learners use the introductory *it* pattern mainly for "the thematization of evaluation or opinion" in order to evaluate difficulty, importance, expectedness and appropriateness (e.g. *it is difficult to*; *it is important to*). The Norwegian learners also overused the pattern compared to the NS writers (Hasselgård, 2009:137). Since Hasselgård's (2009) study included investigation of stance markers more broadly, too, this study will be returned to when the previous literature on stance marking in apprentice writing is discussed in section 2.2.2.

Römer (2009) used the pattern as an example of the inseparability of lexis and grammar, and included tokens with the following structure: *it is* (ADVERB) ADJECTIVE *to*/*that*/OTHER-clause. Texts from a relatively broad range of genres and disciplines were included (Römer, 2009:148–149).[12]

[11] The following instances were included: *it* BE *interesting/of interest to/that*; *it* BE *important/of importance to/that*; *it* BE *evident/clear/apparent/obvious that*; *it* BE (*im*)*possible to/that* (Ädel, 2014:73).

[12] The corpus includes argumentative essays written by L1 German undergraduates; linguistics and literature essays written by L1 German final-year undergraduates and first-year grad-

Whereas no clear general differences across NS status were reported (Römer, 2009:158), certain non-target-like uses were noted, such as the learners' occasional use of "extreme" adjectives, such as *amazing*, *stupid* and *ridiculous* inside the pattern (Römer, 2009:156). The less advanced learners were furthermore found to use the pattern to "express strong emotions" in a way that the experts and the NS students did not do (Römer, 2009:156). It was, nonetheless, concluded that factors such as general language proficiency and expertise in academic writing seemed to account for more of the differences found than NS status (Römer, 2009:158–159).

Thompson (2009) investigated the use of the introductory *it* pattern (of the kind *it* + BE/SEEM/APPEAR+ ADJ + *to*/*that*) in NS student writing, across two disciplines: history and engineering. The study is a cross-sectional investigation of students who were in their first through third year of university studies. Differences were found across the disciplines investigated: the pattern was commonly used to comment on measurement, calculations and testing in the engineering texts, whereas it was used to examine, note and argue in the history texts (Thompson, 2009:70). Furthermore, as has been noted in previous studies of a more general character (e.g. Herriman, 2000; Mair, 1990), an association between meaning and type of clausal subject is found, where *that*-clauses were very often used to express "judgement about a proposition" (e.g. "from the graph it is evident that") (Thompson, 2009:69), and *to*-clauses were commonly used to "comment on a process" (e.g. "it is very simple to change the input conditions…") (Thompson, 2009:70). Overall, an increase in the use of the pattern was found when the first-year students were compared to the third-year students (coupled with a decrease in the use of the pronoun *I*). Thompson (2009:79) concluded that "it is clear that the use of the two patterns [*it* + BE/SEEM/APPEAR + ADJ + *to* and *it* + BE/SEEM/APPEAR + ADJ + *that*] increases over time, which may be taken to be an indication of a growing ability to express judgements within one's writing in an authoritative manner."

Disciplinary differences in NS student writing were also found in Charles' (2000) relatively small-scale study focusing on instances of the introductory *it* pattern with an adjective phrase and a *to*-infinitive clause (*it* + linking verb + ADJ + *to*-infinitive clause). She compared eight MPhil theses in politics/international relations to eight PhD theses from materials science to see whether the pattern contributed to creating an appropriate academic persona (Charles, 2000:46). While the two groups used the matrix-clause adjectives very similarly (Charles, 2000:51), certain functional differences were found across the disciplines. Examples include the politics/international

---

uates; linguistics, philosophy, psychology and sociology papers written by NS final-year undergraduates and first through final-year graduates; and published articles from linguistics, psychology and social sciences (Römer, 2009:149).

relations students' tendency to use the *to*-infinitive in these instances of the pattern for *mental* or *verbal processes*, whereas the materials science students used these more frequently for *material processes* (e.g. *to evaporate*), in Halliday's (1994:108–109) terminology (Charles, 2000:49). Charles (2000:56) concluded that the introductory *it* pattern is used for creating an appropriate academic persona, although one that differs somewhat across the two disciplines investigated. Further disciplinary differences were found in Charles (2006), where reporting clauses followed by a *that*-clause (of the form **V that** and ***it be* V-ed that**; e.g. *Jones argues that* and *it has been reported that*) were investigated in apprentice NS writing across two disciplines: politics/international relations and materials science. For example, instances of the introductory *it* pattern with a passive verb were found to occur more frequently in the materials science corpus (16.3 per 100,000 words, compared to 4.2 per 100,000 words for politics/international relations) (Charles, 2006:313). More generally, however, an integral citation with a human subject[13] proved to be predominant in both disciplines and ARGUE was found to be the most frequent verb used (Charles, 2006:326).

In sum, the survey of previous research presented above showed that there are some factors that ought to be taken into account in studies of the introductory *it* pattern. For example, many of the studies have found that learners tend not to use the pattern in a fully expert-like or native-like manner, as they occasionally use it infelicitously; this is investigated further in the present thesis. Furthermore, the fact that there might be L1 transfer involved has also been taken into consideration by including only one L1 group in each subcorpus used in the thesis. The possibility of L1 transfer is addressed in all four articles when relevant, but investigated in most detail in Article 4, where the stance marking function of the pattern is examined in relation to similar constructions (see section 7.4). Moreover, since disciplinary differences have been noted for NS students, this factor is controlled for (and investigated) in the present thesis.

2.2 Stance marking

As shown in Article 3 of the present thesis, the most important function of the pattern is to express writer stance; Article 4 further explores this function by comparing the pattern to other, similar constructions. As background for Article 4, a brief overview of studies on stance will now follow. Following Gray & Biber (2012:15), the term *stance* (as realized through *stance markers*) is used to denote "the linguistic mechanisms that convey a speaker or

[13] In an integral reference the author's name is incorporated in the sentence (e.g. "AUTHOR (YEAR) suggests that…" or "According to AUTHOR (YEAR), we can…") (Charles, 2006:314; see also Swales, 1990).

writer's personal attitudes and assessments". The most pioneering work within the paradigm of *stance* has arguably been carried out by Biber and colleagues (e.g. Biber et al., 1999; Biber 2006a, 2006b) and by Hyland (e.g. 1996, 2000, 2005, 2008a). Evaluative language has also been investigated in paradigms such as *appraisal* (e.g. Martin & White, 2005), *intensity* (Labov, 1984), *evaluation* (e.g. Hunston & Thompson, 2000) and *attitude* (e.g. Halliday, 1994). Although these paradigms all have a common core in that they focus on the encoding of assessment and opinions in language, their focus and approach differ somewhat (Gray & Biber, 2012:15). The present section will be limited to studies carried out within the paradigm of stance. This section provides a brief overview of such studies using general data (2.2.1) and apprentice data (2.2.2). The purpose of the section is not to provide a comprehensive account of the vast field of stance, but rather to give a brief overview, as background for Article 4.

2.2.1 General studies

As is the case for the introductory *it* pattern in particular, differences across mode, discipline and genre have been found for stance marking in general as well. A brief overview of some of the studies that have investigated these and other factors will now follow. The object of study in these studies varies considerably and ranges from specific groups of stance markers, such as reporting verbs (e.g. Thompson & Ye, 1991; Hyland 2000) and hedging devices (Hyland, 1996), to broader investigations of stance in large-scale studies (e.g. Biber et al., 1999; Biber 2006a, 2006b; Hyland, 2008a).

In Biber et al. (1999, chapter 12), the use of stance was investigated across modes and genres. While the authors noted that stance can be expressed by means of paralinguistic devices (e.g. pitch, loudness) and lexical choices (e.g. *the cats are* ***nice***), they focused on *grammatical stance marking*. Grammatical stance markers are grammatical devices that "express a stance relative to another proposition", such as stance complement clauses (e.g. *I am happy that she showed up*; *it is interesting that she did*) and stance adverbs (e.g. *kind of*, *importantly*) (Biber et al., 1999:966). The corpus used includes subsets of one spoken genre (conversation) and three written genres (fiction, news and academic texts). The results showed that while stance markers are common in all four genres, they are "considerably more common" in the spoken material than in the written material (Biber et al., 1999:979). Nonetheless, Biber et al. (1999:980) noted that "it is not at all uncommon to find personal attitudes and estimates of likelihood expressed in academic writing through impersonal stance devices such as modal verbs, adverbials, and extraposed complement clauses [i.e. the introductory *it* pattern]". There were also certain similarities found across the modes; for example, both academic prose and conversation showed "heavy reliance" on single adverbs (Biber et al., 1999:983). With regard to genre differences,

stance complement constructions were found to be less common in academic prose than in the other genres (Biber et al., 1999:984). However, one subcategory of stance complement clauses, namely extraposed complement clauses (i.e. the introductory *it* pattern), was found to be frequent in academic prose (especially those with a *to*-infinitive clause) (Biber et al., 1999:984).

Biber (2006a:100) further examined three of the structural categories investigated in Biber et al. (1999), namely modal verbs (including semi-modals) (e.g. *might*, *has to*), stance adverbs (e.g. *unfortunately*, *possibly*) and the stance complement clause construction (e.g. *I hope that…*; *it is amazing that…*). The study used a corpus consisting of four different genres, two spoken (classroom teaching, class management talk) and two written (textbooks and syllabi) (Biber 2006a:100). As was the case in Biber et al. (1999), clear differences across both genre and mode were found. For example, all categories of stance investigated were more frequent in speech than in writing; stance devices were especially rare in textbooks compared to the other genres (Biber, 2006a:114). One explanation suggested for the scarcity of stance devices in textbooks is that "[t]he general pattern is to emphasize the factual nature of the information in textbooks, with comparatively little attention to the assessment of likelihood" (Biber, 2006a:114).

Hyland (2008a) investigated whether there are disciplinary differences in the use of stance marking, where interviews were complemented by examination of research articles from eight disciplines (mechanical engineering, electrical engineering, marketing, philosophy, sociology, applied linguistics, physics and microbiology). The study covered "320 potentially productive items based on previous research, grammars and the most frequently occurring items in the texts themselves" (Hyland, 2008a:4); these include hedges (e.g. *possible*, *may*), boosters (e.g. *sure*, *prove*) and attitude markers (e.g. *remarkable*, *unexpected*).[14] The results showed that stance markers occurred very frequently; this was especially the case for the hedges category (Hyland, 2008a:11). With regard to the cross-disciplinary comparison, the "soft sciences" (philosophy, marketing, sociology and applied linguistics) displayed particularly high frequencies of the items investigated (Hyland, 2008a:13). The author explained these results by stating that "[p]ersonal credibility, getting behind your arguments, plays an important part in creating a convincing discourse in the humanities and social sciences" (Hyland, 2008a:13). This reasoning was echoed in the interviews, where researchers in the "soft sciences" emphasized the need to present ideas in a confident manner in order to be taken seriously (Hyland, 2008a:14).

[14] This study also looked at *engagement*; Hyland (2008a:5) draws a distinction between *stance*, which refers to "the writer's textual 'voice'", and *engagement*, which addresses the "ways writers rhetorically recognise the presence of their readers to actively pull them along with the argument […], and guide them to interpretations." Examples of the latter (which is not investigated in the present thesis) include imperatives and questions.

In sum, stance markers have been found to be used frequently in all genres and disciplines investigated. However, certain differences across genres, modes and disciplines have been noted, which is something that will be further discussed in the next subsection, which focuses on studies of apprentice writing.

2.2.2 Apprentice writing

The factors reported on above, discipline and genre, have also been investigated in apprentice writing with similar results (e.g. Charles, 2003; Reilly et al., 2005), as discussed below.[15] After this initial discussion, an overview of studies looking at additional factors of more direct relevance to Article 4 will be presented. These factors include level of expertise in academic writing (e.g. Neff et al., 2003) and L1 transfer (e.g. Hasselgård, 2009).

Charles (2003) and Reilly et al. (2005) both looked at NS apprentice writing. Charles (2003:315) investigated nouns preceded by sentence-initial deictic *this* to explore in what way "retrospective labels are used to construct stance". The material comprised graduate theses from two disciplines: politics/international relations and material science. She found that these constructions "contribute significantly to the construction of stance in both disciplines examined" (Charles, 2003:324).[16] There was also a certain degree of variation with discipline. For example, the politics corpus exhibited higher frequencies of metalinguistic nouns (e.g. *distinction*, *argument*) and stance nouns (e.g. *confusion*); this was explained in terms of cross-disciplinary differences in the construction of knowledge and research practices (Charles, 2003:324). Reilly et al. (2005) looked at a wider range of constructions, including attitudinal markers such as modal verbs, across two genres: expository texts and written narratives. The material was collected from American students who were in fourth grade, junior high, high school and graduate school (Reilly et al., 2005:192). Differences across genres were noted; for example, all age groups included more deontic and epistemic attitudinal markers (e.g. *must*, *should* and *might*, *could* respectively) in their expository texts than in their narratives (Reilly et al., 2005:201). Furthermore, the youngest writers were found to use proportionally more deontic than epistemic markers compared to the other age groups (Reilly et al., 2005:202).

In what follows, a brief summary will be provided of previous research looking at two factors of direct relevance to Article 4, namely level of expertise in academic writing and L1 transfer. Differences across at least one of

[15] Stance marking has also been investigated in spoken apprentice data, for example using the Michigan Corpus of Academic Spoken English (MICASE) to explore the use of the hedges *sort of* and *kind of* (Poos & Simpson, 2002) and evaluative adjectives and intensifiers (Swales & Burke, 2003). These studies will, however, not be discussed further here.

[16] The token frequencies are, however, relatively low, as only 146 and 180 tokens respectively are found in the two corpora (see Charles, 2003:116).

these factors have been found in Petch-Tyson (1998), Neff et al. (2003), Hasselgård (2015) and Hasselgård (2009). In particular, these studies have found overuse or slightly marked use of stance markers of different kinds by learner writers. Petch-Tyson looked at features of involvement, such as emphatics (e.g. *really*), "fuzziness words" (e.g. *kind of*) and first-person pronouns (*I*, *we*) in NNS and NS student writing. Using the Swedish, Finnish, Dutch and French subcorpora of ICLE, Petch-Tyson (1998:112) found that the L1 Swedish and Finnish students showed more interpersonal involvement than the other L1 groups. Similarly, Neff et al. (2003) found NNS students to make frequent use of stance markers. The authors looked at NNS and NS student writing and expert writers. The NNS material was culled from the ICLE corpus and comprises the following L1s: Dutch, Belgian-French, Italian, and peninsular Spanish. They investigated "devices used to construct writer stance", such as *it is* + (adverb) adjective + *that* and conjuncts (*conjunctions* in their terminology), such as *however* and *nevertheless* (Neff et al., 2003:562). A certain degree of overuse of most of the stance markers investigated was noted in the learner data compared to the NS students and expert writers (Neff et al., 2003:565). In a study comparing L1 Norwegian learners' use of *-ly* adverbs to that of NS students (in BAWE and the Norwegian component of VESPA), a certain degree of overuse was reported, although only for certain categories, such as frequency adjuncts (e.g. *frequently*, *usually*) (Hasselgård, 2015:187). Other examples of overused constructions can also be found if we return to Hasselgård's (2009:121) study, which was discussed in relation to the introductory *it* pattern in section 2.1.2. As can be recalled, the study compared discourse patterns produced by Norwegian learners of English (from the Norwegian subcorpus of ICLE) to those produced by NS students and expert writers of English and Norwegian (mainly from the ENPC and ICE-GB corpora). Among other things, the results showed that the L1 Norwegian learners tended to overuse clause-initial adverbials. As Hasselgård (2009:126) noted, while such uses are not ungrammatical in English, they are more marked in English than in Norwegian. Furthermore, Hasselgård (2009:126) reported that it seems that "the learners are using Norwegian patterns in their written English". The Norwegian learners' English therefore seemed to have been affected by L1 transfer. Overuse of clause-initial adverbials has also been found in studies looking at other L1 varieties, such as Danish (see, e.g., Shaw, 2004:79 for a study on published articles in economics).

The extent to which there might be L1 transfer involved in the use of other constructions as well was further explored in Hatzitheodorou & Mattheoudakis (2009) and Altenberg & Tapper (1998). Hatzitheodorou & Mattheoudakis (2009) looked at stance markers in the Greek subcorpus of ICLE compared to NS student writing and found certain signs of L1 transfer. The learners' use of boosters in particular led the authors to conclude that the learners' "choices seem to be culturally induced and, therefore, it is possible

that learners may be misled into believing that they can transfer Greek rhetorical conventions to L2 writing" (Hatzitheodorou & Mattheoudakis, 2009:175). By contrast, Altenberg & Tapper (1998) did *not* find any clear evidence of L1 transfer in their relatively small-scale study of conjuncts (e.g. *in any case*, *moreover*) in NNS (L1 Swedish and L1 French) and NS student writing.[17] The L1 Swedish learners in particular used conjuncts similarly to the NS students; Altenberg & Tapper (1998:92) note that

> [t]he English and Swedish use of connectors is evidently not different enough for transfer to play an important role, and even in areas where there are cross-linguistic differences, such as the position of conjuncts, the learners seem to have little difficulty in conforming to the target norm.

There were, nonetheless, certain differences across NS status, such as both student groups' tendency to use less formal connectors. The learners (especially the L1 Swedish learners) were reported to exhibit a "lack of register awareness" (Altenberg & Tapper, 1998:92).

All in all, although the above-mentioned studies have explored a very broad range of different stance-marking topics in a wide variety of different L1 groups, the tendencies noted (such as the learners' propensity to overuse stance markers) help provide the backdrop for Article 4. Since these studies, along with the ones of a more general character presented in the previous subsection, emphasize the need for taking discipline, mode and genre differences into account, the number of disciplines, modes and genres has been restricted in Article 4. Furthermore, as can be noted, many of the studies reported on above use material from the ICLE family (Granger et al., 2009). While the ICLE corpus is uncontestably the largest international corpus of learner writing, the material used for the present thesis allows for investigation of texts that not only are longer than the ICLE texts on average, but also are more like research articles in their structure, thereby being more comparable to the expert texts investigated in Article 4.

[17] Subcorpora no larger than approximately 50,000 words were used (see Altenberg & Tapper, 1998:82).

3 Methodological considerations

There are three methodological approaches that have served to inform the thesis: corpus linguistics (see, e.g., McEnery & Hardie, 2012), Pattern Grammar (Hunston & Francis, 2000) and Contrastive Interlanguage Analysis (CIA) (Granger, 1996, 2015). Since one of the approaches, corpus linguistics, provides a background for the other two, it will be presented first, in section 3.1. Pattern Grammar will subsequently be introduced in section 3.2, followed by CIA in section 3.3.

3.1 Corpus linguistics

While the first uses of corpora for linguistic research date back to the first half of the 1900s (cf., e.g., Boas, 1940), the term *corpus linguistics* was not used until the 1980s (McEnery et al., 2006:3; see also Leech, 1992). Since then, a vast number of corpus linguistics studies have been carried out on a wide variety of topics. Instead of relying on introspection, which was the main method just after the mid-1900s, researchers could now use large computerized corpora consisting of authentic texts (sometimes in combination with introspection) to answer questions about language use, in a way that increases the reliability of the results (McEnery et al., 2006:6–7).

In addition to *reliability*, other central concepts of corpus linguistics include *representativeness* and *replicability*. According to Leech (1991), a corpus is representative if the finding (based on its texts) can be deemed generalizable to the language variety that the corpus represents. The material used for the thesis has been sampled with the aim of being representative of its respective variety. Furthermore, using corpus-linguistics methods also increases the replicability of the results, as any trained researcher using the same material and the same method is likely to achieve very similar (if not the same) results; this is not necessarily the case for studies that are based solely on introspection, as the latter rely more heavily on processes that are inherently subjective. Nonetheless, replicability is not achieved automatically by using a corpus. For example, categorization based on unclear criteria can have a negative effect on the replicability of the results. As will be discussed in more detail in section 6.2, the functional classification developed in Article 3 in the present thesis attempts to attain replicability to as great an extent as possible by not relying heavily on word semantics.

While certain characteristics are common to all areas of corpus linguistics, the views on whether to take a corpus-based or a corpus-driven approach differ (Tognini-Bonelli, 2001). In short, corpus-based studies make use of corpus data primarily to test pre-existing theories (McEnery & Hardie, 2012:6). By contrast, corpus-*driven* studies "start with as few preconceived theoretical concepts as possible" (Lindquist, 2009:26); the corpus itself is seen as providing the source of information about language (McEnery & Hardie, 2012:6). However, in practice, the divide between the two views is not as clear as it may seem at first sight (e.g. Meyer, 2015); instead, they can be seen as representing the poles of a continuum on which corpus studies can be placed. In the present thesis, the two views are seen as complementary. While pre-existing theories form the basis of all four articles, elements of a corpus-driven approach are incorporated to a varying degree.

3.2 Pattern Grammar

Although the Pattern Grammar framework is not applied in its entirety to the material, it is still of relevance to the thesis, as discussed below. In this framework, the object of study is the surface form of recurrent clusters of words. A pattern is defined as "a phraseology frequently associated with (a sense of) a word, particularly in terms of the prepositions, groups and clauses that follow the word" (Hunston & Francis, 2000:3). Unlike approaches to the study of recurrent clusters such as *lexical bundles* studies (e.g. Biber et al., 1999:990), Pattern Grammar looks not only at lexical items, but also (and primarily) at word classes. This approach allows for a large set of similar tokens to be grouped under the same syntactic heading. For example, the pattern **V n -ing**, can be realized as *kept her waiting*, but also as *kept him wondering* etc., both of which would be subsumed under the same pattern heading (cf. Hunston & Francis, 2000:53). In a lexical bundles approach, the two realizations would typically have been counted as two separate instances and only detected if each of the two realizations was sufficiently frequent.

In the Pattern Grammar framework, the introductory *it* pattern is considered to be many separate patterns (e.g. ***it* V ADJ to-inf**, ***it* V that**, etc.). Nonetheless, these patterns are viewed as having a common core (Hunston & Francis, 2000:156–157). The conceptualization of the introductory *it* pattern in the present thesis as a single pattern with subpatterns does not, then, fully correspond to that of Pattern Grammar; nonetheless, both conceptualizations share the view that the focus should be on the surface form, i.e. what is visible in a given text (Hunston & Francis, 2000:37f). Thus, no assumptions are made about any underlying structures or movement of constitu-

ents.[18] The Pattern Grammar framework forms the basis for the classification scheme used in Article 2 (see section 7.2).

Furthermore, Pattern Grammar perceives lexis and grammar as being closely related (cf. also Sinclair, 1991).

> Patterns and lexis are mutually dependent, in that each pattern occurs with a restricted set of lexical items, and each lexical item occurs with a restricted set of patterns. In addition, patterns are closely associated with meaning, firstly because in many cases different senses of words are distinguished by their typical occurrence in different patterns; and secondly because words which share a given pattern tend also to share an aspect of meaning. (Hunston & Francis, 2000:3)

With regard to the instantiations of the introductory *it* pattern that include an adjective, it is noted that such instances contain adjectives that "fall into a limited number of meaning groups" (Hunston & Francis, 2000:29). Another implication that this has is that different meanings of polysemous words can be distinguished with the help of the pattern they occur in (Römer, 2009:143); for example, *it is possible that* and *it is possible to* make clear the two senses of *possible* (cf. section 6.2.3). A certain tendency for a correlation between form and function in more general terms has been noted in the material. For example, whereas instances of the pattern with no adjective or noun phrase (SV: *it seems that*) are most often used for hedging, passive instances of the pattern (SV_{pass}: *it has been shown that*) are frequently used to make neutral observations. However, it falls outside the scope of the thesis to investigate this in more detail.

Whereas Pattern Grammar is firmly situated in a corpus linguistic context, it is not based on (or explicitly designed for) learner language. In order to add this perspective, let us now turn to CIA, which is especially created for studies of learner language.

3.3 Contrastive Interlanguage Analysis

CIA (Granger, 1996, 2015) was developed within a project that resulted in one of the most well-known learner corpora to date, the *International Corpus of Learner English* (ICLE; Granger et al., 2009). CIA was later updated, and the new version is referred to as CIA^2 (Granger, 2015). Many studies in the field of learner corpus research (LCR) make use of this approach. The material and method used in the present thesis are compatible with CIA^2.

The method is outlined in Granger (2015). In short, the model includes two sets of language varieties: reference language varieties (RLVs) and in-

[18] The introductory *it* pattern (or "subject extraposition") is sometimes viewed as resulting from movement of the clausal subject to the end of the sentence (***to be on time*** *is important* → *it is important* ***to be on time***) (see, e.g., Quirk et al., 1985:1391). This view will be discussed further in section 5.1.

terlanguage varieties (ILVs). Any one of the varieties can be compared to any other variety. Granger (2015:17) stresses that the RLV(s) do not have to be NS varieties, and brings up Lee & Chen's (2009) study of function words in published expert writing as an example of a study where NS status is not considered. The ILV(s) can be investigated separately, or compared to the RLV(s). Granger (2015:18) furthermore urges researchers to take certain other variables into account, such as dialectal variables and task variables.

Traditionally, two kinds of comparisons have been carried out in LCR studies: ILV (learner) use vs. RLV (typically NS) use of a language, and ILV (learner:$L1_A$) use vs. ILV (learner:$L1_B$) use (Granger, 2015:8; cf. also Granger, 1996). Both of these approaches have been described as valuable "to meet the theoretical and applied objectives of LCR, namely to gain a better understanding of the mechanisms of foreign or second language acquisition and to design more efficient language teaching tools and methods" (Granger, 2015:9). However, in order to be able to systematically investigate L1 transfer, Granger (2015:11) stresses the need for the less common second comparison (i.e. comparing ILVs – learners groups with different mother tongues).

One point of criticism that has been voiced against studies using the earlier CIA model is that the terms *underuse* and *overuse* seem to presuppose that "the learner should at all times attempt to conform to native-speaker norms" (Aston, 2008:343 cited in Granger, 2015:18). While the terms *underuse* and *overuse* are kept for pragmatic reasons in CIA^2, it is emphasized that the terms are to be understood as neutral, i.e. as descriptive rather than prescriptive (Granger, 2015:18; see also Gilquin & Paquot, 2008 and Hasselgård, 2015:165). This is the case in the studies included in this thesis as well, where the terms are used as statistical terms (see Lee & Chen, 2009).

In the present thesis, the ILV (learner) vs. RLV (native-speaker) comparison (using both British English, BrE, and American English, AmE, as RLVs) can be found in Articles 1 and 3. In Article 2, the reference corpus comprises expert writing instead (regardless of NS status). In Article 4, comparisons are made between an RLV of expert writing and two different ILVs (L1 Swedish and L1 French). More information about the corpora used will be presented in the next section.

4 Material

The corpora used in the present thesis are presented in section 4.1, followed by a description of the sampling process in section 4.2.

4.1 Corpora used

Data from a total of six different corpora were used in the present thesis: ALEC, BATMAT, BAWE, LOCRA, MICUSP and VESPA. These six corpora will be presented below in sections 4.1.1 (ALEC) and 4.1.2 (the remaining corpora). The full corpora are described here; however, as detailed in the articles, subsets were used from each of the corpora.

4.1.1 ALEC

The *Advanced Learner English Corpus* (ALEC) was compiled by the author as part of the present thesis project in 2013 (see also Larsson, 2014).[19] It comprises a total of 1.3 million words (146 texts) written by 143 different students. The texts are Bachelor's (BA), one-year Master's (Magister) and Master's (MA) theses[20] in English linguistics and English literature. The texts were written by students at Stockholm University in Sweden between 2004 and 2013. While the vast majority of the texts were written by students whose L1 is Swedish, the corpus also includes some texts written by students who have English or some other language as their L1. Based on the forms used to gather metadata for the compilation of BAWE and MICUSP, a form was put together and used to elicit information about the students whose texts were included in the corpus. As summarized in Table 1, the metadata include information about the student's grade, discipline, level of study, age, sex, self-reported native language(s), language(s) spoken at home, primary school language(s), secondary school language(s) and time spent in an English-speaking country.

[19] I am very grateful to Gregory Garretson at Uppsala University for his advice and technical support in the corpus-compilation process.

[20] The majority of the students were in their third year of studies when they wrote their BA thesis, in their fourth year when they wrote their Magister's thesis and in the fifth year when they wrote their MA thesis.

Table 1. Information provided for all texts included in ALEC.

Variable	*Possible values*
Term	Fall 2004 – Spring 2013
Grade	A – E
Discipline	English linguistics; English literature
Genre	Essays
Level	BA; Magister; MA
Age	[Open value]
Sex	Male, Female
Self-reported native language(s)	Swedish; English; Swedish + English; OTHER
Childhood language(s)	Swedish; English; Swedish + English; OTHER
Main primary school language(s)	Swedish; English; Swedish + English; OTHER
Main secondary school language(s)	Swedish; English; Swedish + English; OTHER
Time spent in English-speaking countries	< 3 months – [open value]
Time spent in Sweden[21]	[open value]

In ALEC, the body of the text was marked up for title, main sections, paragraphs, quotes and italics. All graphs, tables, references and tables of contents were excluded. The corpus is tokenized and encoded in Unicode UTF-8. It comes in XML, HTML and plain TXT formats.

4.1.2 Other corpora

The BATMAT corpus, which is under compilation at Åbo Akademi University in Finland, comprised 119 texts (34 BA theses and 85 MA theses) written by students in realia, English linguistics and English literature (Lindgrén, 2015) at the time it was used for the present thesis. The BA theses were written between 2002 and 2015, while the MA theses (which were not investigated in the present thesis) were written between 1972 and 2015. The corpus includes detailed metadata about, among other things, the student's official mother tongue, home language(s), foreign languages studied and time spent in an English-speaking country. The corpus comes in RTF and (annotated) plain TXT formats.

The *British Academic Written English corpus* (BAWE) is a 6.5-million-word corpus (approximately 2,800 texts) compiled in 2004–2007 at the universities of Warwick, Reading and Oxford Brookes in the UK (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

[21] This question was only used for the small number of students who were brought up in an English-speaking country.

were divided into 13 different text types, such as case study, essay and research report. Only papers that were awarded a high grade were included. Information about the student's first language, year of birth, gender and education is provided, along with information about the text itself (e.g. number of paragraphs, lists and tables, etc.). The corpus comes in XML, plain TXT and PDF formats.

The *Louvain Corpus of Research Articles* (LOCRA) is a 3-million-word corpus of published research articles that is under compilation at the Centre for English Corpus Linguistics at Université catholique de Louvain in Belgium (http://www.uclouvain.be). The corpus includes articles from peer-reviewed, top-rated journals in linguistics, business and medicine. The corpus is not restricted to native-speaker writing.

The *Michigan Corpus of Upper-Level Student Papers* (MICUSP) is a 2.6-million-word corpus (approximately 830 texts) of student writing from the University of Michigan in the USA (Römer & O'Donnell, 2011; O'Donnell & Römer, 2012). It spans sixteen disciplines, including psychology, engineering and economics, and a range of different text types, such as proposals and argumentative essays. Only high-graded papers were included. The metadata include information about the student's gender, age and language background. The corpus is available in XML and plain TXT formats.

The *Varieties of English for Specific Purposes dAtabase* (VESPA) is a corpus of learner writing that is administered at the Centre for English Corpus Linguistics at Université catholique de Louvain in Belgium (Paquot et al., 2013). The corpus includes texts from disciplines such as business, linguistics and literature; the texts have been compiled by collaborators in several European countries, including Belgium, Norway and Sweden.[22] Among other things, the metadata include information on the student's age, gender and language background. The corpus is available in XML format.

4.2 The sampling procedure

The sampling procedure and the subsets used for each of the four articles will be presented in brief below; more detailed information can be found in the corresponding articles.

The same corpora were used for Articles 1 and 3, namely subsets of ALEC, BAWE and MICUSP. The articles compare texts written by students who are, on average, in their third or fourth year of linguistics or literature studies. Only L1 Swedish student texts were included in the learner subcorpus and only L1 English student texts were included in the NS-student refer-

[22] I have contributed 650,000 words of L1 Swedish learner writing to VESPA. While some of the texts in VESPA are also included in ALEC, the vast majority of the texts are not; the new texts were collected from students at Uppsala University and Stockholm University in 2014.

ence subcorpus. 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 (135 texts in total). A more detailed description of the subcorpora can be found in section 3.1 in Article 1.

In Articles 2 and 4, the reference corpus was made up of published expert writing, rather than NS student writing, as was the case in Articles 1 and 3. While NS student writing is oftentimes considered a good yardstick for NNS students, expert writing is often viewed as being the ultimate goal for students (NNS and NS alike). The present thesis therefore used NS student writing as the reference for two of the articles (Articles 1 and 3) and expert writing as the reference for the two remaining articles (Articles 2 and 4). Article 2 made use of the linguistics subcorpus of LOCRA and the highest-graded NNS linguistics texts (i.e. those that were awarded an A or a B) that were written by students who are in their third or fourth year of studies on average[23] and whose L1 is Swedish. A total of approximately 1 million words (109 texts) of expert writing and approximately 170,000 words (21 texts) of learner writing were included in the analysis. Section 2.1 in Article 2 provides a more detailed description. Article 4 included the largest number of different corpora. It compared a subset of LOCRA texts to subsets of texts from ALEC, VESPA and BATMAT. The expert subcorpus was made up of the same subset of LOCRA as for the second article (approximately 1 million words and 109 texts). The learner subcorpora comprised linguistics texts written by students whose L1 is Swedish (from ALEC, VESPA and BATMAT)[24] and by students whose L1 is (Belgian) French (from VESPA). The vast majority of the texts included were written by students who were in their third year of study. The L1 Swedish data comprised approximately 775,000 words (94 texts) and the L1 French data were made up of 105,000 words (20 texts). A more detailed description can be found in section 2.1 in Article 4.

[23] Due to the relatively limited size of the subcorpora, level of study (i.e. investigations of third-year vs. fourth-year students, etc.) falls outside the scope of the study.

[24] The students whose texts are included in the BATMAT corpus have "Finland Swedish" as their L1, whereas the other students have "Sweden Swedish" as their L1.

5 The introductory *it* pattern and related constructions

In this section, the introductory *it* pattern and relevant related constructions will be further examined. The full definition is given below in 5.1, followed by a discussion of particularly interesting and/or problematic tokens in 5.2. Constructions that are related to the introductory *it* pattern but that were excluded from the thesis will be reviewed in section 5.3. After that, the functionally related constructions investigated in Article 4 will be presented in section 5.4. Finally, the data retrieval process will be explained in section 5.5.

5.1 Definition and discussion of the introductory *it* pattern

As was mentioned 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 important *to look at the interaction* […] (LOCRA_022-02)

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. The excluded constructions will be discussed in separate subsections in 5.3.1–5.3.6.

(5) *It* was this ambiguity *that left room for the social positioning events below*. (LOCRA_021-01)
(6) *It* is early in the morning at home. (LOCRA_014-02)
(7) Students were told the test was not going to be marked and that *it* was meant to help the teacher evaluate his own course. (LOCRA_018-01)

The second subject in an instance of the introductory *it* pattern is made up of a nominal clause.[25] Using Quirk et al.'s (1985:1048ff) terminology, 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.4.053)
(9) Interestingly, *it* is not clear *how Huddleston & Pullum want to account for these patterns*. (ALEC_LING.4.083)
(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.4.082)
(13) Nonetheless, *it* is worthwhile *to discuss briefly some of the key linguistic structures* […]. (LOCRA_005-02)
(14) Rather, *it* is necessary *for researchers to understand the motivations and effects of variation* […]. (LOCRA_021-05)

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) (see section 5.3.4) and tokens with nominal extraposition (i.e. where the *it* refers to an NP), as in (16) (see section 5.3.5).[26]

(15) *It* is a pity *if teachers do not use this awareness*. (ALEC_LING.3.078)
(16) *It*'s staggering *the number of books that can pile up*. (Example taken from Michaelis & Lambrecht, 1994:362)

[25] 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*.
[26] Unlike the present thesis, Huddleston & Pullum (2002:1407–1408) include nominal extraposition (what they call "extraposition of NPs") in the category of extraposition. While some empirical studies (such as Kaltenböck, 2005) include nominal extraposition in the category of extraposition, many others (such as Hasselgård, 2009) explicitly state that they do not.

The definition used for the present thesis is largely in keeping with Quirk et al.'s (1985:1391ff) definition of what is referred to as *subject extraposition*,[27] 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 thesis is thus slightly more exclusive, as *only* subordinate nominal clauses are allowed as the second subject here. Furthermore, while the definition used in the present thesis is very similar to that of Quirk et al. (1985:1391) where *structure* is concerned,[28] 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., Herriman, 2013; Miller, 2001). 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 important *to look at the interaction* […] (LOCRA_022-02)
(18) *To look at the interaction* is 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).[29]

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 thesis). 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 thesis, as well as in previous studies (e.g.

[27] The thesis 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 & Pullum, 2002:963), as in *a common language makes **it** easier **to gain employment*** (LOCRA_021-02). Subject extraposition is described as "the most important type of extraposition" (Quirk et al., 1985:1391). Object extraposition will be further discussed in section 5.3.6.

[28] As will become clear, the definition of what constitutes an instance of the introductory *it* pattern in the present thesis does, however, differ somewhat from the definitions of (*subject*) *extraposition* proposed by Biber et al. (1999:155) and Huddleston & Pullum (2002:1403ff).

[29] It has, however, been argued that the clausal subject does not remain unchanged when extraposed, as can be seen, for example, when co-ordinated clauses are extraposed: *to X and to Y **are** important ~ *it **are** important to X and to Y* (see Seppänen, 1999:61, and Seppänen et al. (1995) for a more detailed discussion).

Huddleston & Pullum, 2002:969; 1402; Mair, 1990:30–31; Mindt, 2011:31; Mukherjee, 2006:348–349).[30] 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.3.108)
(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,[31] 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 thesis 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;[32] instead, the construction is referred to as the *introductory* it *pattern*, as mentioned above (see, e.g., Groom, 2005; Hunston & Francis, 2000). Other terms used in previous studies to refer to the pattern[33] include the *anticipatory* it *pattern* (e.g. Ädel, 2014; Hyland, 2008b), it-*clauses* (e.g. Hewings & Hewings, 2002) and it-*extraposition* (e.g. Kaltenböck, 2005; Zang, 2015).[34]

5.2 Further discussion of relevant tokens

This subsection will consider some tokens that merit further discussion, either because they were particularly interesting syntactically or semantically,

[30] Similar frequency differences have also been noted in Quirk et al. (1985) for subject *that*-clauses (p. 1049), and for subject *to*-infinitive clauses (p. 1062). By contrast, Quirk et al. (1985:1064) also state that "subject *-ing* clauses are not normally extraposed"; this claim seems to be supported by the data in this thesis, where such instances of the pattern were very infrequent.

[31] 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. However, it can be noted in passing that unlike the view taken in the present thesis and in Quirk et al. (1985) and Biber et al. (1999:733), Huddleston & Pullum (2002:906) do not count these tokens as extraposition. Instead they group them under the heading "the impersonal construction with *it* as subject", for example arguing that the subject is semantically empty.

[32] Even though the term *extraposition* is not used in the thesis, the term *non-extraposed* is, for clarity. *Non-extraposed* is used to denote the construction in which the clausal subject is placed before the verb (i.e. in pre-predicate position).

[33] The definition of what constitutes an instance has, however, varied somewhat.

[34] Of the other terms that have been used to describe the pattern, *it*-clauses and *it*-extraposition can be seen as misleading, since both of them suggest, using Quirk et al.'s (1985) terminology, that it is the introductory pronoun *it* that is extraposed and not the clausal subject.

or because they were problematic to classify. Tokens that deserve more attention in relation to one of the classification systems will be discussed in the relevant subsection – 6.1.3 for the syntactic classification or 6.2.3 for the functional categorization.

Although the definition outlined above made for a relatively straightforward classification process, there are some tokens that deserve further attention. Two such groups of tokens will be discussed here, one of which was excluded and the other included. While both groups are marginal in terms of frequency, they are still interesting enough to merit specific mention.

The first group, which was excluded, comprises *it says that* tokens, as in (21) and (22) below.

(21) What is important is the first of these sentences because in essence *it* says *that Stephen is able to perceive his identity as baby tuckoo as an infant*. (ALEC_LIT.3.059)

(22) Gibb's proposal calls to mind the lexicalization hypothesis but can be said to be a more extreme version of it, as *it* says *that the literal meaning will be neglected if the first few words are recognized as part of a known expression*. (ALEC_LING.3.066)

While these tokens are structurally similar to the introductory *it* pattern,[35] they are potentially ambiguous: *it* can be interpreted either as being introductory or as having anaphoric reference. In the first interpretation, such tokens can be perceived as being *functionally* similar to valid tokens of the pattern, such as *it is said that*, where *it* does not have anaphoric reference. In the second interpretation, *it* refers back to a previously mentioned entity X (*the first of these sentences* and *Gibb's proposal* respectively), thus making the sentence functionally similar to *X reads as follows* (or, possibly, *X makes the claim that*).

The fact that the second interpretation can be arrived at without alteration of the examples, whereas an adverbial has to be added to clearly disambiguate the two possibilities in favor of the first interpretation (e.g. *it says* ***on page 2*** *that*), could suggest that *it* does, in fact, have anaphoric reference. Regardless, however, it is very difficult, if not impossible, to *rule out* the possibility that *it* has anaphoric reference for the *it says that* tokens, as none of the instances included an adverbial of that kind.

As can be recalled from section 5.1, for the token to be a valid instance of the introductory *it* pattern, the definition stipulates that *it* cannot have anaphoric reference (cf. also section 5.3.2). The definition of the pattern does not take function into account, nor does it take structural similarity into consideration, as witnessed by the fact that tokens such as (23), which are struc-

[35] The fact that there is no non-extraposed equivalent to *it says that* is not in and of itself a reason for disqualifying the token (cf. the discussion of obligatory extraposition in section 5.1, and Quirk et al., 1985:1183, 1392[a]).

turally similar, but where *it* has anaphoric reference (referring back to *the recipient*), are excluded.

(23) Furthermore, since the recipient is typically given, *it* is likely to be pronominal and hence light. (LOCRA.LING.013-03)

Therefore, since a (vague) anaphoric reference cannot be ruled out for *it* in *it says that* tokens, these tokens were not included in the analysis.

The second group to be discussed contains tokens that are covered by the definition. Nonetheless, these tokens form a group that differs slightly from prototypical instances of the pattern, as they allow for the complement of the matrix clause to become the subject, as exemplified in (24) and (25).[36]

(24) *It* is my aim *to shed light on the phenomena in the novel from a phenomenological viewpoint* […]. (ALEC_LIT.4.004)

(25) […] *it* has thus been an aim of feminism *to "recover a neglected history"* […]. (BAWE_LIT.3.3006k)

Unlike the more prototypical instances of the pattern, tokens belonging to this group can be reconstructed in two different ways, as shown in the two constructed examples for (24) in (26) and (27).

(26) *To shed light on the phenomena in the novel from a phenomenological viewpoint* is my aim.

(27) My aim is *to shed light on the phenomena in the novel from a phenomenological viewpoint*.

In (26), *it* is interpreted as anticipating the *to*-infinitive clause, with the resulting non-extraposed sentence being structured as *to*-infinitive clause+V+NP. By contrast, (27) shows a construction where *it* is interpreted as anticipating the NP instead of the clause, thus resulting in an alternative construction NP+V+*to*-infinitive clause. Nevertheless, tokens belonging to this group have been included in previous studies of the pattern (e.g. Herriman, 2000), as *it* is seen to anticipate the clausal subject, even though an alternative restructuring is possible (Jennifer Herriman, personal communication). Moreover, even if the second interpretation could make such tokens seem similar to the excluded group of nominal extraposition (see section 5.3.5), the two groups are clearly distinct, as, unlike tokens with nominal extraposition, the tokens attested in the corpus have a clausal subject, which is what is required by the definition. In sum, these tokens have been included as valid instances of the introductory *it* pattern in this thesis.

[36] In form, these tokens seem to be at least remotely related to tokens such as *she's a pleasure to teach*, which exhibit so-called *tough movement* (cf. Quirk et al., 1985:1394).

5.3 Delimitations: Excluded constructions

In this subsection, six constructions that were not covered by the definition stated in section 5.1 (and that were therefore excluded) will be discussed further in relation to the introductory *it* pattern: constructions with a 'prop' *it* (5.3.1) and *it* with anaphoric reference (5.3.2), *it*-clefts (5.3.3), introductory *it* followed by an adverbial clause (5.3.4) and nominal extraposition (5.3.5). The closely related, but excluded, construction *object extraposition* will be discussed in 5.3.6.

5.3.1 Constructions with 'prop' *it*

'Prop' *it* (also referred to as 'expletive' *it*, 'empty' *it* or 'pleonastic' it) is placed in subject position in clauses in which no participant is required. 'Prop' *it* predominantly occurs in clauses commenting on time, atmospheric conditions and distance (Quirk et al., 1985:748f; Huddleston & Pullum, 2002:1482), as illustrated in (28)–(30), using examples from Quirk et al. (1985:748).

(28) *It*'s ten o'clock precisely.
(29) *It* was sunny yesterday.
(30) *It*'s just one more stop to Toronto.

The 'prop' subject *it* is considered to be clearly different from the introductory subject *it*, as the latter, unlike the former, has "cataphoric reference to a postponed clausal subject" (Quirk et al., 1985:749[b]).

There are slightly divergent views on whether introductory *it* carries more or less semantic content than 'prop' *it*. With regard to semantic content, Quirk et al. (1985:349) place 'prop' *it* in-between introductory *it* and *it* in cleft sentences, on the one hand, and referring *it*, on the other hand, whereas Kaltenböck (2003) argues for a model where introductory *it* is placed between 'prop' *it* and referring *it* (cf. also Kaltenböck, 1999; 2002).[37] However, as the relative order of the different kinds of *it* is not of importance to the present thesis, it will not be discussed further here.

5.3.2 Constructions in which *it* has anaphoric reference

Most instances of the pronoun *it* have anaphoric reference; that is, they refer back to a previously mentioned item. While there are tokens where the pronoun *it* has clear anaphoric reference that bear no resemblance to the intro-

[37] Biber et al. (1999:125) treat introductory *it* as a subcategory of what they call "dummy subject *it*", along with 'prop' *it*, the *it* used in existential clauses and *it*-clefts. Similarly, Huddleston & Pullum (2002:1403) refer to the introductory pronoun *it* as "a dummy subject". However, in the present thesis, the term 'dummy' *it* is not considered an accurate description of the function performed by introductory *it* and will therefore not be used.

ductory *it* pattern, such as (31), other instances of referring *it*, such as (32), are seemingly more similar, at least structurally.

(31) This study has found some interesting differences between the two branches of literary criticism, but since **it** is limited in scope, further research is needed to determine whether the findings can be generalised. (ALEC_LING.3.011)

(32) This instrument was considered a valid measure of participants' vocabulary knowledge for this study because **it** is not designed to estimate general vocabulary knowledge but rather to track the early development of specific word knowledge […] (LOCRA_018-04)

In the second example, *it* is followed by a predicate and a *to*-infinitive clause, making it superficially similar to instances of the introductory *it* pattern. However, in both examples, the pronoun *it* has anaphoric reference (to *this study* and *this instrument* respectively), and is thus not an introductory *it*.

The fact that there are invalid tokens that seemingly contain the same identifying features as the pattern strongly argues in favor of manual screening of the results in studies of this kind; this is something that will be returned to in section 5.5, where the data retrieval process is described.

5.3.3 *It*-clefts

It-clefts, as in (33) and (34), are used for information focus; whatever is placed in the empty slot in *it* BE ___ *that*/*who* becomes highlighted (Quirk et al., 1985:89; Huddleston & Pullum, 2002:1414ff).

(33) […] *it* is Van Helsing who first informs the reader of this […]. (ALEC_LIT.3.121)

(34) […] *it* is the construction as a whole that makes a contribution to the progressive aspect meaning. (LOCRA_021-03)

A cleft sentence is divided up into two parts: the initial focused element and a back-grounded clause that "resembles a relative clause" (Quirk et al., 1985:89). The back-grounded clause is described as "quite distinct" from a nominal *that*-clause (Quirk et al., 1985:1384[a]).[38] Other differences between the constructions include the fact that unlike the nominal clause of an introductory *it* pattern, the clause in a cleft sentence cannot be considered the subject of the sentence. Furthermore, while the introductory *it* has cataphoric reference to the nominal clause in the introductory *it* pattern, the *it* of an *it*-cleft refers to the focused element (what comes after BE) and therefore not to

[38] While the present thesis shares Quirk et al.'s (1985) view, there are slightly divergent views on the matter: Huddleston & Pullum (2002:962) refer to the back-grounded clause as a (full) relative clause.

the clause (see also Calude, 2008:13ff). The introductory *it* pattern and *it*-clefts are thus clearly distinct constructions.

5.3.4 Introductory *it* followed by an adverbial clause

Another structurally related, albeit comparatively infrequent, construction is made up of an introductory *it*, a predicate and an adverbial clause. Two examples of constructions in which an introductory *it* is followed by an adverbial clause are shown in (35) and (36).

(35) […] *it* would be better *if there were no wars*. (ALEC_LING.3.141)
(36) […] *it* seems *as if the secrecy is there in order to keep magic exclusive*. (ALEC_LIT.3.052)[39]

However, despite an apparent resemblance to certain valid instances of the introductory *it* pattern (most notably tokens with non-conditional uses of *if*, i.e. where *if* can be exchanged for *whether*, as in (37)), tokens with an adverbial clause behave differently from tokens with a nominal clause in several ways.

(37) […] *it* is unclear *if this type of critique has reached a wider circle* […] (ALEC_LING.3.135)

For example, nominal clauses can perform a wide variety of functions in a sentence, including subject, object, complement and prepositional complement; this is not the case for adverbial clauses, which mainly function as adjuncts or disjuncts (Quirk et al., 1985:1047–1048). In contrast to non-extraposed nominal clauses, non-extraposed adverbial clauses still require *it* to be retained as a subject; this becomes clear in the non-extraposed version of (35), namely *if there were no wars, it would be better*.

5.3.5 Nominal extraposition

The construction that has been referred to as *nominal extraposition* or *extraposition of NPs* is made up of the pronoun *it*, a predicate and an NP with a relative clause (Huddleston & Pullum, 2002: 1407–1408; Michaelis & Lambrecht, 1994). It is most often associated with spoken language (Michaelis & Lambrecht, 1994), which might explain why this construction is very infrequent in the written material used for the present thesis. Two examples from Michaelis & Lambrecht (1994:362) can be found in (38) and (39).

(38) *It*'s staggering the number of books that can pile up.
(39) *It*'s amazing the access he got!

[39] Huddleston & Pullum (2002:962) note that since such clauses cannot function as subject, "there is no question of an extraposed subject analysis for this construction"; instead, they classify *as if*-clauses as an "internal complement" contained in an "impersonal construction".

Nominal extraposition may appear similar to the introductory *it* pattern in that the non-extraposed equivalent of the sentence does not retain *it* (unlike constructions with an introductory *it* that are complemented by an adverbial clause; see section 5.3.4). However, there is one very important difference, namely that *it* in instances of nominal extraposition does not refer to a nominal clause, but rather to a phrase: a noun phrase (i.e. *the number of books that can pile up* and *the access he got* respectively in the examples above). Nominal extraposition and the introductory *it* pattern are thus considered distinct constructions in the present thesis.

5.3.6 Object extraposition

Object extraposition is commonly included in the category of *extraposition* (see Quirk et al., 1985:1391ff; Huddleston & Pullum, 2002:963). However, due to the comparably very low frequencies with which this construction occurred in the data,[40] this construction has not been included in the present thesis. The construction includes an introductory *it* which is co-referential with an objective clause. Examples are given in (40) and (41), as well as in (42) (which is taken from Quirk et al., 1985:1393).

(40) […] Miss Crawley's fortune makes *it* difficult *for her relatives to reciprocate in equal manners* […]. (ALEC_LIT.3.008)
(41) […] the scope of this study makes *it* impossible *to develop this issue further*. (ALEC_LING.4.027)
(42) You must find *it* exciting *working here*.

Extraposition is obligatory in SVOC and SVOA clause types for instances where the object is a *to*-infinitive clause, as in the first two examples above, or a *that*-clause; by contrast, there is a non-extraposed equivalent for instances with an *-ing* clause (Quirk et al., 1985:1393). There were no instances of object extraposition with an *-ing* clause attested in the data.

In addition to the type of clause element that is extraposed, object extraposition and the introductory *it* pattern exhibit further differences. A closer look at the instances identified in the data used for the present thesis shows that object extraposition is particularly frequently used together with the main verbs MAKE and FIND, as in the examples above; by contrast, the most frequent verb in the introductory *it* pattern is BE. Furthermore, there is a difference with regard to voice: while active clauses such as *X makes it clear that* take object extraposition, the corresponding passive clause, *it is made clear by X that*, takes subject extraposition (i.e. it is an instance of the introductory *it* pattern).

[40] For example, while there were 1,650 valid instances of the introductory *it* pattern in the 1-million-word corpus of expert writing used for Article 2, there were only just over 60 tokens of what Quirk et al. (1985:1391ff) refer to as *object extraposition*.

Now that the *structurally* related constructions that were excluded have been discussed, we will turn to the *functionally* related constructions that were included in Article 4.

5.4 Related stance marking constructions

The fourth article widens the scope of the thesis to also include investigation of grammatical realizations that are functionally related to the introductory *it* pattern. The focus of this article was on the stance marking function of the introductory *it* pattern and these constructions. The article used Biber et al.'s (1999:969–970) framework of grammatical stance as a first step, as well as Quirk et al.'s (1985:612ff) subdivision of adverbials as a second step.

As described in more detail in section 7.4, Article 4 included investigation of the following grammatical realizations of stance (Biber et al., 1999:969–970): stance adverbials (43),[41] the stance complement clause construction (44) and stance noun + prepositional phrase (45).

(43) **Interestingly**, there is an apparent tension between these two criteria […]. (ALEC_LING.3.125)
(44) **It is possible that** Ahab here makes a mistake when trusting the mysterious fire […]. (ALEC_LIT.3.002)
(45) **The importance of** time and temporal structure is further enhanced by the use of temporal words as 'now', […], and 'previous'. (ALEC_LIT.5.028)

The stance complement clause construction, which in Biber et al.'s (1999) terminology subsumes extraposition (i.e. the introductory *it* pattern), includes instances of complement clauses controlled by a verb (46), an adjective (47) or a noun (48).

(46) This considered, **I hope that** the findings would be representative enough to at least give us a hint of what are the most popular English textbooks in Sweden […]. (ALEC_LING.4.013)
(47) **I am not sure that** the metaphor is the best one to describe our professional activity. (ALEC_LING.3.135)
(48) **The possibility that** it was coincidence can be excluded […]. (ALEC_LING.3.044)

The article focused on those instances where one base form, such as IMPORTAN*, can be instantiated through all three of these main grammatical realizations of stance (e.g. *importantly*, *it is important that*, *the importance*

[41] *Single adverbs*, which is what the study focuses on, is by far the most frequently occurring member of the *stance adverbial* category (Biber et al., 1999:982).

of), in order to investigate what might have affected their distribution (see section 7.4).

Using Quirk et al.'s (1985:612ff) more fine-grained categorization of adverbials as a second step, the article also looked more closely at the introductory *it* pattern in relation to disjuncts (49). Compared to the other subcategories of adverbials – adjuncts (50), subjuncts (51) and conjuncts (52) – disjuncts is the subcategory that is most similar to the introductory *it* pattern semantically and syntactically.

(49) **Clearly**, this corpus cannot be the bases of a diachronic study of productivity […]. (ALEC_LING.3.029)
(50) **Because of all these characteristics**, MUDs are seen as the foundation for modern role-playing games […]. (ALEC_LING.3.014)
(51) […] I **kindly** explained that I needed to interview students with Swedish as a mother tongue. (ALEC_LING.3.016)
(52) […] they are **nonetheless** distinct markers of what it is to be non-Hobbit […]. (ALEC_LIT.4.018)

Conjuncts are different from the other three subcategories in that they "have the function of conjoining independent units rather than one of contributing another facet of information to a single integrated unit" (Quirk et al., 1985:631). Adjuncts are "similar in weight and balance of their sentence role to other sentence elements" and subjuncts have "a lesser role than the other sentence elements". Disjuncts, by contrast, have "a superior role as compared with the sentence elements" (Quirk et al., 1985:613), which makes this category structurally similar to the introductory *it* pattern. This similarity also extends to semantics. For example, Quirk et al. (1985:623) note that the disjunct *evidently* corresponds to *it is evident that* (as well as to the non-extraposed equivalent of the introductory *it* pattern). There is, however, only a select group of disjuncts for which this correspondence holds true; Quirk et al. (1985:623) mention *perhaps* as an example that does not have such a correspondence. Using a bottom-up approach, five such disjuncts with equivalent instances of the introductory *it* pattern were identified and investigated in Article 4,[42] as further discussed in section 5.5.2.

5.5 Data retrieval and processing

The same method was used to retrieve the tokens in the first three articles; this method will be described and discussed in 5.5.1. The method used for

[42] These were INTEREST* (*interestingly, it is interesting that*), IMPORTAN* (*importantly, it is important that*), POSSIB* (*possibly, it is possible that*), SURPRISING* (*surprisingly, it is surprising that*), and CLEAR* (*clearly, it is clear that*).

Article 4 will be described in 5.5.2. Finally, the data processing for all four articles will be discussed in 5.5.3.

5.5.1 Articles 1, 2 and 3

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 5.3. As has already been mentioned, 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 (53) versus conditional use of an *if*-clause (54) and tokens where *it* has cataphoric reference to a nominal clause (55) versus tokens where *it* has anaphoric reference (56); despite the surface similarity, only the first instances in these two pairs are counted as valid tokens of the introductory *it* pattern; see sections 5.3.4 and 5.3.2 for a more detailed discussion of these groups.

(53) […] *it* can be questioned *if the results are beneficial for either group of students*. (ALEC_LING.3.084)
(54) *It* becomes even clearer *if in the passage the pronoun 'I' is substituted for 'he'* […]. (ALEC_LIT.3.059)
(55) […] *it* should also be acknowledged *that the different worldviews between the source and target contexts make it quite difficult* […]. (ALEC_LING.5.125)
(56) […] 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.5.104)

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 (57) and tokens in which the complementizer *that* is omitted (58), 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 & Reppen, 1998:155).

(57) Nor is *it* her fault *that she was not fed distinction with her mother's milk.* (ALEC_LIT.3.001)
(58) […] *it* seems *she is trying to say that she still suffers from them*. (ALEC_LIT.3.034)

The tokens were subsequently classified into syntactic and/or functional categories, as described in section 6.

5.5.2 Article 4

For the fourth article, the scope was broadened to also include constructions whose functions are similar to the stance-marking function of the introductory *it* pattern (see section 7.4); therefore, a slightly different method was adopted. As mentioned earlier, this article investigates what factors affect the distribution of stance markers that are morphologically related, such as *importantly*, *it is important that* and *the importance of*. In order to identify such stance markers, a bottom-up approach was adopted.

In short, the following steps were taken. First, since the corpora used were not part-of-speech (POS) tagged, search patterns were used to identify potential stance markers that can be realized through three different grammatical realizations from Biber et al.'s (1999:969–970) framework of grammatical stance marking: stance adverbs (*possibly*), stance noun + prepositional phrase (*the possibility of*) and the stance complement clause construction (*it is possible that*) (see section 5.4). Second, the software environment R (R Core Team, 2015) was used to take the intersection of the first four letters of each of these potential stance markers to find tokens for which there was overlap that is retrievable through search patterns across the three categories and across the investigated corpora. Five such triplets were found: POSSIB* (*possibly, possibility, possible*), INTEREST* (*interestingly, interest, interesting*), IMPORT* (*importantly, importance, important*), PRESUM* (*presumably, presumption, presumed*) and PROBAB* (*probably, probability, probable*). The article also included investigation of the two most similar constructions: the introductory *it* pattern and disjuncts. SURPRISING* (*surprisingly*, *it is surprising that*) and CLEAR* (*clearly*, *it is clear that*) were here substituted for PRESUM* and PROBAB*. A more detailed description of the method used can be found in section 2.2 in Article 4 (i.e. in Larsson, under review b).

5.5.3 Data processing

The software environment R (R Core Team, 2015) was used to summarize and manage the data in all four articles; R was also used to test the results for statistical significance using a wide array of tests, depending on the type of data. For the nominal variables, four different tests were used: a Chi-squared test (to compare two sets of frequencies), a multinomial log-linear model for multivariate analyses with a dependent nominal variable with more than two levels, and two kinds of generalized linear models (GLMs): a log-rate GLM (for multivariate analyses) and a binomial GLM (for binomial analyses). The log-rate generalized linear model also takes size of subcorpora into account

(Powers & Xie, 2000:154ff). For the per-text frequencies, Kruskal-Wallis rank sum test was used to test the differences between the medians of these frequencies for statistical significance. Furthermore, in Article 2, a hierarchical agglomerative cluster analysis was carried out, using the UCREL Clustertool (http://corpora.lancs.ac.uk). The tests used are further discussed in the corresponding article.

6 Classification schemes used

Three different classification schemes were used to subcategorize the instances of the introductory *it* pattern in the thesis: two syntactic classifications and one functional classification. The syntactic classifications used are slightly modified versions of pre-existing classifications developed by Quirk et al. (1985:1392) and Francis et al. (1996, 1998) respectively. The functional classification, by contrast, was developed as part of the current thesis project. The two syntactic classifications will be addressed in 6.1, whereas the functional classification will be discussed in 6.2.

6.1 The syntactic classification schemes

The two syntactic classifications are compared and contrasted in section 6.1.1. The classifications themselves are described in Articles 1 and 2, in sections 3.2.2 and 2.2.1 respectively. This section also includes a discussion of how the classifications were modified and developed in the present thesis in 6.1.2. Finally, the tokens that were problematic to categorize using these schemes will receive more attention in 6.1.3.

6.1.1 Quirk et al.'s (1985) classification vs. the COBUILD classification

There are seven categories in Quirk et al.'s (1985:1392) syntactic classification. Examples include SVC (Subject-Verb-Complement: *It is a pleasure to teach her*) and SV (Subject-Verb: *It doesn't matter what you do*) (Quirk et al., 1985:1392). In the COBUILD classification (Francis et al., 1996, 1998), by contrast, the instances of the pattern are classified into surface-level subpatterns, as exemplified below in (59) and (60). Together, the COBUILD grammars list just under 50 different categories (not including those where the introductory *it* is in object position). Further examples are provided later in this subsection. A list of the abbreviations used in the COBUILD grammars can be found in the Appendix.

(59) ***it* V ADJ to-inf** (e.g. *it is important to notice the differences*)
(60) ***it* V be V-ed that** (e.g. *it has been noted that she is often late on Tuesdays*)

There are three main differences between the two classification schemes. The first one pertains to the level of detail aimed at in the classification. Whereas Quirk et al. (1985) list seven different categories, the COBUILD grammars together list about seven times as many; the COBUILD classification is thus much more fine-grained. For example, unlike Quirk et al.'s classification, it encodes information about what type of clause is included in the pattern (i.e. *to*-infinite clauses, *that*-clauses and *wh*-clauses are counted as belonging to separate subpatterns).

Second, in one respect, the COBUILD classification is more limited than that of Quirk et al. (1985): one of the syntactic types listed in Quirk et al. (1985) is not included in the COBUILD classification. While the COBUILD classification includes subpatterns with a past participle (i.e. ***it* be V-ed to-inf**), it does not include any subpattern that would correspond to Quirk et al.'s $SV_{pass}C$ type, where the past participle is followed by a complement. An example from the data is provided in (61).

(61) *It* must be made clear at the outset *that the original purpose of the research was not to elicit metaphors in women's talk [...]* (LOCRA_020-05)

The third difference has to do with how the two classification schemes treat instances of the pattern of the *for/to* and *for/that* types, as in (62) and (63).

(62) […] *it* is easy *for her to accept the principles* […]. (ALEC_LIT.3.106)
(63) *It* is crucial for the present analysis *that the nouns just given are indeed formed from the roots* […]. (LOCRA_022.03)

Whereas Quirk et al. (1985) treat ***for* n** in the first example as the subject of the *to*-infinitive clause, and thus as part of the clausal subject, they treat ***for* n** in the second example as an adverbial. No such distinction is made in the COBUILD classification, where the examples would be classified as ***it* V ADJ *for* n to-inf** and ***it* V ADJ *for* n that** respectively and where ***for* n** is thereby seemingly not counted as part of the clausal subject.

Apart from the three differences discussed above, the subpatterns from the COBUILD classification map onto the syntactic types in Quirk et al.'s (1985) classification relatively straightforwardly. The majority of the subpatterns from COBUILD that contain a noun (64) or an adjective (65) would be classified as SVC in Quirk et al.'s classification. Instances of the pattern that contain only a verb and an obligatory adverbial (66) map onto Quirk et al.'s SVA type.

(64) ***it* V det N that**: *it* must be the case *that the lower head will not undergo movement* […] (LOCRA_010-03)
(65) ***it* V ADJ to-inf**: Of course, *it* is difficult *to generalize from a case study of only three subjects to a broader population*. (LOCRA_022-01)

(66) ***it* V prep to-inf**: *It* is beyond the scope of the present study *to undertake a comprehensive critique of the studies mentioned in 7.* (LOCRA_021-04)

The rest of the subpatterns from the COBUILD classification map onto the remaining four syntactic types from Quirk et al. Instances classified as ***it* be V-ed clause**,[43] as in (67), correspond to the SV_{pass} type; the ***it* V clause** and ***it* V to n clause** subpatterns, as in (68) and (69) correspond to the SV type; and the ***it* V n ADJ clause** subpattern, as in (70), corresponds to SVOC.[44]

(67) *It* might be argued *that this is an epistemic confirmation* […]. (LOCRA_010-02)
(68) However, *it* seems *that LSM F-hs is rare in most varieties of modern-day ASL*. (LOCRA_017-02)
(69) *It* seems to me *that these are useful themes to look at* […]. (ALEC_LIT.3.017)
(70) As her parents usually do not talk to each other*, it* makes her happy *to hear them quarrel.* (ALEC_LIT.3.038)

However, although the vast majority of the subpatterns map onto only one syntactic type, there is one exception. Instances classified as ***it* V n clause**, such as (71) and (72) below, can belong to either the SVC or the SVO type, depending on the function of the noun phrase.

(71) *It* is also the job of the other participants *to co-construct these relevancies*. (LOCRA_012-01)
(72) […] *it* would require moral fortitude *to say no*. (LOCRA_004-02)

In the first example of the two, the noun phrase (*the job of the other participants*) functions as a complement, which makes it SVC; in the second example the noun phrase (*moral fortitude*) is an object of the verb (*require*), making it SVO.

Quirk et al.'s classification scheme has the advantage that it covers all instances of the introductory *it* pattern, as defined in the current thesis; however, the classification is not very detailed. Conversely, while the COBUILD classification is very detailed, it was not found to cover all instances of the pattern attested in the material of this thesis. The two classifications thus provided complementary perspectives on the data of great use for the present thesis. However, as mentioned earlier, certain adjustments were made to the classifications based on the material used in the thesis; this will be addressed in the next subsection below, when the additions made are discussed.

[43] Following Francis et al. (1996, 1998), *clause* is here used as an umbrella term to cover all kinds of clausal subjects.
[44] When a subpattern is described in both grammars (Francis et al., 1996 and 1998), as is the case for many of the noun and adjective patterns, the newest, more detailed version (i.e. that of Francis et al., 1998) is used here and in Article 2. The verb is, however, always referred to as V (rather than *v-link*, which is used in Francis et al., 1998).

6.1.2 Additions to the classification schemes

A few additions were made to the classification schemes in the present thesis: four subcategories, and an OTHER group were added to Quirk et al.'s (1985:1392) classification and 14 subpatterns and a discussion of additional features were added to the COBUILD classification scheme (Francis et al., 1996, 1998). These additions will be described in more detail below.

As mentioned in the previous section, Quirk et al.'s (1985:1392) classification is very general; there was therefore room for further subclassification. When the data had been categorized in accordance with this classification, it became clear that there were enough tokens to merit further subcategorization of the syntactic types SVC and $SV_{pass}C$: one for tokens for which the complement contains a noun phrase, as in (73) and (74), and one for those tokens for which the complement contains an adjective phrase, as in (75) and (76).

(73) SVC:NP *It* is a common belief *that extensive reading in any language will result in a wider vocabulary* […]. (ALEC_LING.4.105)
(74) $SV_{pass}C$:NP […] *it* can be seen as an advantage *for one person to carry out the interviews* […]. (ALEC_LING.3.041)[45]
(75) SVC:AdjP *It* is clear *that the answer to this questions is 'no'* […]. (ALEC_LING.4.130)
(76) $SV_{pass}C$:AdjP However, *it* was found necessary *to broaden the selected definition* (ALEC_LING.3.112)

Of these four subtypes, the SVC:AdjP proved to be by far the largest subcategory in terms of number of tokens encompassed, followed by SVC:NP and the two much less frequent $SV_{pass}C$ subcategories. However, the differences in terms of use of these subcategories between the populations investigated in Article 1 did not prove to be statistically significant; larger corpora than the ones used for Article 1 seem to be needed to investigate these subcategories in greater detail, which will have to be left for future studies to explore.

Furthermore, two groups of tokens that were found in the data were not included in Quirk et al.'s (1985:1392) classification of instances of the introductory *it* pattern, namely subject + passive verb + obligatory adverbial ($SV_{pass}A$), as in (77), and subject + verb + object + obligatory adverbial SVOA,[46] as in (78).

[45] The *as*-phrase is here counted as a complement, rather than as an (obligatory) adverbial, due to its semantic similarity to the corresponding complement and since it can be used with verbs other than BE. However, distinguishing between complements and obligatory adverbials is not unproblematic. Quirk et al. (1985:732–733) argue for treating some of these instances "through gradience". Another problematic case of this sort will be discussed in section 6.1.3.

[46] The possibility of posing an adverbial question ("how long did it take?") rather than a *what*-question in addition to the fact that there is no possibility of passivization strongly suggests that the structure is SVOA rather than SVOO. Both SVOA and SVOO are included as clause types in Quirk et al. (1985:56).

(77) *It* is taken into account *that the RP accent is quite difficult to define* […]. (ALEC_LING.3.099)
(78) […] *it* takes Ashley more than a year *to get all the way down to Chile* […]. (ALEC_LIT.3.068)

Only a handful of tokens fall into these two categories. The categories were therefore merged into an OTHER group.

With regard to the more fine-grained COBUILD classification, three main modifications were made. First, since it did not cover all the instances found in the material, a few additional realizations of the introductory *it* pattern were identified and added. These realizations comprise variants of passive verb patterns,[47] such as (79), variants of noun patterns, such as (80), and one adjective pattern (81). However, the low frequencies that these exhibited in the data might explain why they are not addressed in the COBUILD grammars.

(79) ***it* be V-ed ADJ to-inf**: *It* was not deemed appropriate *to match the Dutch and English texts in terms of linguistic difficulty […]*. (LOCRA_004-04)
(80) ***it* V n for n that**: It is an even greater source of chagrin for me that I produced the original transcript […]. (LOCRA_021-05)
(81) ***it* V ADJ to n what/how**: *It* is not clear to me *how any constraint to prevent such an unnatural process could be built into a model of language change* (LOCRA_015-03)

Second, unlike the present thesis, the COBUILD classification does not distinguish between conditional and non-conditional uses of *if*-clauses, as in (82) and (83) respectively. Without any modification to the classification, these instances would both be classified as ***it* V ADJ when/if**.

(82) *It* would be lovely *if there was more peace in the world*. [italics added] (Francis et al., 1998:491)
(83) […] *it* is not clear *if their proficiency arose from their explicit knowledge, or vice versa.* (LOCRA_017-01)

However, since the *if*-clause including the former is adverbial and thus not counted as an instance of the introductory *it* pattern (see section 5.3.4), a distinction between these two types of *if*-clauses was added for the purpose of the present thesis. Any instances with an adverbial clause were excluded.

Third, a discussion of features that were not included in the COBUILD classification was added. These features were *tense marking* (a token with past tense is shown in (84)), occurrence of *negation* (85), *modal verbs* (86) and *optional adverbials*, realized through a prepositional phrase (87) or an adverb (88).

[47] In Quirk et al.'s (1985:1392) terminology, many of these would belong to the $SV_{pass}C$ category, which is not accounted for in the COBUILD classification.

(84) […] *it* **was** important *to persuade the voters to stand on his side* […]. (ALEC_LING.3.025)
(85) […] *it* is **not** possible *to include and compare all of the information gathered* […]. (ALEC_LING.3.077)
(86) […] *it* **would** be interesting *to know what Berman would prescriptively suggest* […]. (ALEC_LING.3.125)
(87) […] *it* is, **according to this theory**, appropriate *to refer to it as a global error*. (ALEC_LING.3.084)
(88) […] *it* is **thus** interesting *to understand the history and the political or ideological struggle behind their speeches*. (ALEC_LING.3.075)

These features proved to be distributed very differently across the subpatterns. For example, the subpattern ***it* V ADJ what/how** (e.g. *it is not clear how to view this*) behaves very differently from the other subpatterns with regard to these features. In the LOCRA corpus, this subpattern is much more likely to be negated or to include an adverbial, but much less likely to include a modal verb or be used in the past tense compared to the other subpatterns, as shown in section 3.1.3 in Article 2; the features are described in section 2.2.4 of the same article. The outlier status of this subpattern vis-à-vis the other subpatterns with regard to all the investigated features can be visualized through a cluster analysis;[48] the dendrogram, which was not included in Article 2, is shown in Figure 1. The subpatterns that behave most similarly with regard to the investigated features cluster together; the higher up the branching, the more dissimilar the branches are.

[48] Following Gries & Otani (2010:131), the similarity measure used is Canberra, and the amalgamation rule used is Ward's method.

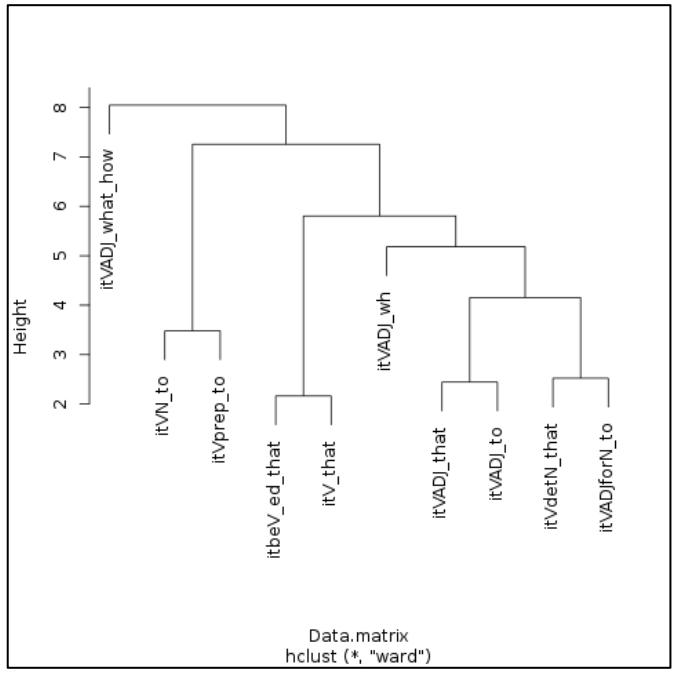

Figure 1. Cluster dendogram of the ten most frequent subpatterns in LOCRA for all the investigated features.

The ***it*** **V ADJ what/how** subpattern can be found in the top-left corner of the graph; it is clearly distinct from the other subpatterns with regard to the investigated features, as witnessed by the fact that it does not cluster with any of the other subpatterns.

More generally, the distribution of the features differs between the different subpatterns (as can be seen from the branching), which suggests that these features are important to take into account for investigations of the pattern. Sections 3.1.3 and 3.2.4 in Article 2 offer more detailed discussions of how the features patterned across the subpatterns, and what effect these have on the use of the introductory *it* pattern.

6.1.3 Further discussion of tokens of relevance to the syntactic classification

Four groups of tokens deserve mention from the perspective of the syntactic classifications used for Articles 1 and 2. The decisions made for these tokens are discussed briefly here in relation only to Quirk et al.'s (1985:1392) classification, as it subsumes the COBUILD classification, due to its more general nature.

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 (89), there were also instances where it is adjectival, as in (90). The former would result in the tokens being classified as belonging to the SV_{pass} category and the latter would result in the tokens being classified as SVC (cf. section 6.1.1), as explained below.

(89) *It* could be argued *that we cannot entirely trust the narrator of 'The Lord of the Rings'* […]. (ALEC_LIT.4.018)
(90) *It* is widely accepted *that sufficient linguistic knowledge is a vital requisite for successful reading in a L2*. (LOCRA_001-04)

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_{pass} if they met both the formal and the functional criteria. For example, such instances were classified as SV_{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 thesis, 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 complement 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 (91) and (92).

(91) It seems to be the case that the Swedish NNS are less aware than the NS […]. (ALEC_LING.3.019)
(92) It in fact appears to be possible to account for a range of sequences […]. (LOCRA_016-03)

At first sight, such tokens appear to allow for dual classification. On the one hand, one could argue that the clausal subject of example (91) 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 (91): [*it*]$_{\text{intr.subj}}$ [*seems to be*]$_{\text{Vgrp}}$ [*the case*]$_{\text{complement}}$ [*that the Swedish NNS are less aware than the NS*]$_{\text{clausal subject}}$. Similarly, the analysis for (92) would be [*it*]$_{\text{intr.subj}}$ [in fact]$_{\text{adverbial}}$ [*appears to be*]$_{\text{Vgrp}}$ [*possible*]$_{\text{complement}}$ [to account for a range of sequences]$_{\text{clausal subject}}$. 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).

6.2 The functional classification

In this subsection, functional classifications used in previous studies will be addressed in section 6.2.1. In section 6.2.2, this will be followed by a discussion of the classification scheme developed for the current thesis. A full description of the classification can be found in section 2.2.3 in Article 3. Finally, a few problematic groups of tokens will receive further attention in section 6.2.3.

6.2.1 Previous functional classification schemes

There have been several functional classifications of the pattern proposed in previous studies; three central ones – those of Herriman (2000), Hewings & Hewings (2002) and Groom (2005) – will be discussed briefly below. Herriman included tokens where *it* is followed by a predicate and a *to*-infinitive

clause, a *that*-clause, a *wh*-clause or an *-ing* clause; only subject extraposition in the active voice was included (Herriman, 2000:583). The tokens were classified into four categories: epistemic modality (e.g. *it is clear*), deontic modality (e.g. *it is your duty*), dynamic modality (e.g. *it is impossible*) and evaluation (e.g. *it is interesting*) (Herriman, 2000:584f).

The classification developed by Hewings & Hewings (2002) covered tokens that have an interpersonal function, i.e. it excluded tokens that were used solely for text-organizing purposes (e.g. *it was pointed out in chapter one that*) or that had an ideational function (e.g. *it is possible to*) (Hewings & Hewings, 2002:371–372). There are four categories: hedges (e.g. *it is likely*), attitude markers (*it is worth pointing out*), emphatics (e.g. *it is important to stress*) and attribution (e.g. *it is estimated*) (Hewings & Hewings, 2002:372).

A slightly different approach was taken by Groom (2005), who investigated tokens that included an adjective followed by a *to*-infinitive or a *that* clause (***it* v-link ADJ that/to**; e.g. *it is interesting to/that*). Groom (2005:261) classified the tokens into six meaning groups, based on Francis et al.'s (1998:480–484, 494–498) classification: adequacy (e.g. *it is enough that*), desirability (e.g. *it is fitting that*), difficulty (e.g. *it is difficult to*), expectation (e.g. *it is interesting that*), importance (e.g. *it is significant that*) and validity (e.g. *it is likely that*). This classification also formed the basis for the categorization used by Römer (2009). In the present thesis, Groom's classification is used for subcategorization of the *attitude markers* category (see Article 3, section 2.2.3).

The three classifications do not always map onto one another, as there are some important differences between them. For example, while Hewings & Hewings (2002:372) counted tokens such as *it is likely that* and *it is clear that* as belonging to two different functional categories (hedges and emphatics respectively), such tokens were perceived as having the same function in Herriman's (2000:585) and Groom's (2005:261) classifications (epistemic modality and validity respectively). Furthermore, tokens such as *it is good to*, *it is important to*, *it is surprising that* were counted as belonging to three different categories in Groom's classification (desirability, importance and expectation respectively); in Herriman's classification, by contrast, all such tokens would be classified as evaluation.

Although all three classifications have many strengths, they also have certain weaknesses from the perspective of the present thesis, as will be discussed further in the next subsection in relation to the classification developed as part of the thesis project.

6.2.2 The functional classification developed

As was mentioned in section 3.1, 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 thesis has been designed with the aim of limiting the impact of subjective interpretation on the classification. The classification 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 full classification can be found in Article 3, section 2.2.3. 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 (see section 2.2.3 in Article 3). The method of assigning binary features in a definition-based classification[49] has been used previously for example in Quirk et al. (1985:206–207) to classify situation types. However, to the best of my knowledge, it has not previously been used to classify instances of the introductory *it* pattern or similar constructions.

Although the functional classification of the present thesis started out from preconceived categories, the categories have been refined based on the data. These categories were loosely based on those of the previous classifications discussed in the preceding subsection. 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 thesis.

First, none of these classifications cover all the tokens that were included in the present thesis. For example, none of them include a category for SV_{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 & Hewings's (2002) classification, *it is important to* is counted as an attitude marker and *it is apparent 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

[49] A definition-based classification can be contrasted to classifications used in *prototype theory*, where the features are weighted (i.e. where certain features are more important than others) (see, e.g., Rosch, 2009).

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 used for the present thesis 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 for Article 3, as well as in the data of the rest of the thesis. For these reasons, this functional classification was deemed suitable for categorizing the material in the thesis. It is hoped that the classification can be of help to researchers wishing to investigate the introductory *it* pattern or similar constructions in the future.

The system proved to work well for classifying the data. Among other things, the results show that the majority of tokens have at least one feature set to positive, as discussed in Article 3. In fact, less than 25 percent of the tokens (n=1,610) had none of the features set to positive. The Attitude marker category (-H+A-E), as in *it is interesting to note*, proved to be the largest category, covering almost 46 percent of the tokens. The results of this study will be returned to in section 7.3. There were, however, a few slightly problematic groups of tokens; these will be addressed in the next section.

6.2.3 Further discussion of tokens of relevance to the functional classification

There were three groups of tokens that deserve mention; the decisions made for these will be discussed in turn below. First, there were a few tokens which contain the adverb *quite*. Since the adverb can function both as a hedge (typically in BrE) and as an emphatic (typically in AmE) (Quirk et al., 1985:446[a]), it was very difficult to determine which meaning is intended. Therefore, a decision was made not to count *quite* as either a hedge or an emphatic, but instead classify tokens such as (93) and (94) simply as *attitude markers*, as all tokens of this kind included an attitude marker.

(93) *It* is quite difficult *to distinguish between these approaches empirically* […]. (LOCRA_010-03)

(94) […] *it* was quite clear *that Ndebele used this particular story to, as he expresses it, "build an interesting and provocative (and superficially plausible) case"* […]. (ALEC_LIT.3.080)

Second, for the group of tokens that include the adjective *possible*, it was particularly important to take the clause type into account. As has been observed in previous studies (e.g. Mair, 1990:49–50; Groom, 2005:259), *possible* has different meanings depending on the clause type. With a *to*-clause, *possible* expresses "theoretical possibility" (i.e. that something can be done), as in (95), whereas *possible that* is used to express "factual possibility" (i.e. that something is probable), as in (96). The former were categorized as *attitude markers*, whereas the latter was categorized as *hedges*.

(95) […] *it* is possible *to reduce their cognitive input* […]. (ALEC_LING.3.101)

(96) […] however, *it* is possible *that students occasionally practiced the type of inferential processing measured in this study during some class periods*. (LOCRA_013-04)

The third group contains tokens with a non-compositional use of the introductory *it* pattern, as in (97).

(97) On the other hand, *it* is tempting *to agree with Quirk (1991), who would argue that the most important goal* […] *is proficiency* […]. (ALEC_LING.4.013)

While this token can be interpreted literally, as in *I am actually tempted to agree with Quirk*, it can also be interpreted as the writer having taken a stance (*you should not agree with Quirk, even if it may seem tempting*). The first interpretation would result in the example being classified as a value-neutral *observation*, whereas the second interpretation would make the token an *attitude marker*. However, in order not to attempt analyses of the perlocutionary act of such instances, these were classified as *observations*.

7 Summaries of the articles

In the subsequent four subsections, 7.1–7.4, the four articles included in the present thesis will be summarized and discussed in turn. The articles contributed to the overall aim by approaching the use of the introductory *it* pattern from slightly different angles, comparing and contrasting different parameters and materials. An overview of the main results can be found in the next section, section 8.

7.1 Article 1. "A syntactic analysis of the introductory *it* pattern in non-native-speaker and native-speaker student writing" (Larsson, forthcoming)

The first article takes a formal, syntactic perspective and investigates what constitutes the full inventory of the pattern at a macro level, starting out from Quirk et al.'s (1985:1392) syntactic types (see section 6.1.1). It also investigates how these realizations of the pattern are distributed across NNS and NS student writing in two disciplines (linguistics and literature) and across two levels of achievement (higher-graded papers vs. lower-graded papers). The material used comes from one corpus of learner writing, ALEC, and two corpora of NS student writing: BAWE (for BrE) and MICUSP (for AmE).

The results showed that the same three syntactic types – SVC, SV and SV_{pass} – were the most frequent ones in all subcorpora, suggesting that the use of the pattern is stable across the points of comparison. Nonetheless, while the SVC type (e.g. *it is interesting to*) proved to be most frequent, the fact remains that the SV and SV_{pass} types (e.g. *it seems that* and *it has been noted that* respectively) also make up a considerable proportion of the data. Excluding these from analysis, as has been done in some previous studies, thereby does not seem to give the full picture. Furthermore, there were clear differences found between the two disciplines investigated. The article thus concludes that there appear to be discipline-specific conventions with regard to the pattern that students could benefit from being made aware of. For example, the pattern was used to make discipline-specific moves, such as objectively commenting on empirical findings (for linguistics) or discussing characters in a work of fiction (for literature).

While there were no major differences across NS status, there were noteworthy differences across levels of achievement. For example, the NNS students whose papers received a lower grade made significantly more frequent use of the most frequent syntactic type, SVC, in a way that suggests that this group is more prone to clinging to what can be referred to as *lexico-grammatical teddy bears* (cf. Hasselgren's (1994) use of "lexical teddy bears"). What follows from this is that frequent use of the pattern does not necessarily entail proficient use of it; this is something that was returned to in the third article where the functions of the pattern were investigated.

7.2 Article 2. "The introductory *it* pattern: Variability explored in learner and expert writing" (Larsson, 2016)

The second article also looks at the pattern from a formal, syntactic angle; however, this article takes a complementary, micro-level approach, making use of a more detailed syntactic classification of the pattern, namely that of the COBUILD grammars (Francis et al., 1996, 1998) in expert and learner writing (see section 6.1.1). The article investigates the extent to which the pattern is fixed or variable, using the type-token ratio and the relative frequency of the different realizations, as well as occurrence of additional features (i.e., tense and presence of negation, modal auxiliaries and/or optional adverbs). It also investigates whether the use of the pattern in expert data is matched in learner data. The material is culled from a corpus of published expert data (LOCRA) and from a corpus of learner data (ALEC).

The results show that while the pattern as a whole can be considered to be relatively invariable in the sense that a small set of realizations make up the bulk of the tokens for each subpattern of the introductory *it* pattern, there is still a range of variability between the subpatterns. In addition, four of the subpatterns were particularly frequent compared to the other ones; these were ***it* be V-ed that** (e.g. *it has been said that*), ***it* V ADJ that** (e.g. *it is clear that*), ***it* V ADJ to-inf** (e.g. *it is possible to*) and ***it* V that** (e.g. *it seems that*). The subpatterns also differed with regard to the additional features, with certain subpatterns showing a preference for one or several of the features, whereas others showed a strong *dis*preference for these features. The subpatterns thus appear to each have their own behavioral profile.

When the results of the investigation of expert writing were compared to those of learner writing, it became clear that the same four subpatterns were the most frequent ones in the learner writing as well. The learners did, however, use these subpatterns more frequently in relation to the other subpatterns than the experts. In addition, there were clear differences between the groups with regard to relative frequencies, in particular for the most frequent subpattern in the expert data, namely ***it* V ADJ to-inf**. The fact that this sub-

pattern was so frequent in the expert data makes it a likely candidate for a frequency effect, thus resulting in very frequent use in the learner data. Since the learners overused not only the subpattern as a whole, but also its most frequent realization, ***it* V POSSIBLE to-inf**, compared to the experts, there seems to be such an effect in play.

7.3 Article 3. "A functional classification of the introductory *it* pattern: Investigating academic writing by non-native-speaker and native-speaker students" (Larsson, under review a)

In the third article, a functional perspective is added. The article develops a functional classification for the introductory *it* pattern that describes all instances of the pattern covered by the definition. The classification is especially designed to increase the replicability of the results, as it mainly uses linguistic evidence other than word semantics to categorize the tokens, unlike most previous studies. This classification is subsequently used to investigate the functions of the pattern across the same three parameters as were explored in the first article: NS status, discipline and levels of achievement. The article uses data from BAWE (BrE student writing), MICUSP (AmE student writing) and ALEC (learner data).

The results show 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. It was furthermore concluded that the stance marking function is not monolithic, as it comprises at least five different combinations of features from the classification. 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 and levels of achievement. 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*. Furthermore, the lower-graded texts contain a significantly higher number of instances of attitude markers than the higher-graded texts. The article thus concludes that frequency of use and proficient use of the pattern are not necessarily correlated. Finally, a certain overreliance on high-frequency realizations was also noted in the lower-graded texts, similarly to what was found in Article 1. All in all, it was concluded that none of the

three factors investigated can be disregarded when trying to understand what mechanisms underlie the differences and similarities between the groups included for investigation.

7.4 Article 4. "*The importance of*, *it is important that* or *importantly*? The use of morphologically related stance markers in learner and expert writing" (Larsson, under review b)

In the fourth article, the scope is broadened to also include constructions that are functionally related to the instances of the introductory *it* pattern that have a stance marking function. As was found in the third article, stance marking is the most important overarching function of the introductory *it* pattern. This article helps situate the pattern vis-à-vis these other constructions in order to gain more insight into the use of the pattern in academic discourse. In more detail, the aims of the article are to investigate what influences the use of a certain grammatical realization of stance and to detect potential challenges for L1 Swedish and L1 (Belgian) French learners. Starting out from Biber et al.'s (1999:969–970) framework of grammatical stance marking,[50] the article examines the use of morphologically related stance markers (such as POSSIB*: *the* ***possibility*** *of*, ***possibly*** and *it is* ***possible*** *that* and IMPORTAN*: *the* ***importance*** *of*, ***importantly*** and *it is* ***important*** *to*) across three factors: level of expertise in academic writing, L1 transfer and lexis. It also looks at the two syntactically and semantically most similar subcategories, namely the introductory *it* pattern and disjuncts. The material used comes from a corpus of expert writing (LOCRA) and three learner corpora (ALEC, VESPA and BATMAT).

The results show that all three of the investigated parameters seem to affect the use of one grammatical realization over another. For example, there was clear inter-lexical variability between the base forms (e.g. POSSIB*, IMPORTAN*, as above), which suggests that each base form has a lexico-grammatical preference. The learners furthermore overused almost all stance markers, in particular INTEREST* (*interest*, *interestingly*, *interest*) and PROBAB* (*probability*, *probably*, *probable*), which stood out as especially problematic for the learners.

Further differences were found when instances of the introductory *it* pattern and disjuncts were compared. The introductory *it* pattern was found to enable inclusion of additional information, such as pre-clausal hedges (e.g. *it seems interesting that* vs. *interestingly*) and certain adverbs (*it is extremely*

[50] Three of Biber et al.'s (1999) constructions are used: stance noun + prepositional phrase, stance adverb and the stance complement clause construction.

important that vs. *importantly*). As this is not possible for disjuncts, this finding suggests that seemingly similar instances of the two constructions cannot always be used interchangeably. In addition, two of the other three factors, lexis and level of expertise in academic writing, were shown to be important.

All things considered, there seem to be principled explanations for why one grammatical realization of stance is used instead of another one. The article provided a macro perspective on some of the realizations of stance markers that are available to writers and detected problem areas for learners with different L1s. The article also developed a method for investigating such realizations of stance in corpora that are not POS tagged, taking a computational approach to identify four-letter items that can be found in all of the relevant constructions.

8 Overview of the main results and concluding remarks

In this section, I return to the overall aim of the thesis, in order to discuss the key findings and contributions of the thesis more from a bird's-eye perspective. In section 8.1 an overview of the main results of the thesis will be provided. In section 8.2, some concluding remarks will be given.

8.1 Overview of the key findings and contributions of the thesis

The present thesis has investigated the use of the introductory *it* pattern and, when relevant, similar constructions in academic discourse. The main focus has been on learner use of the pattern and these constructions, but expert data and NS student data have been used as a point of reference. The thesis was designed to give a more in-depth understanding of the pattern in academic writing through investigations of both its formal and its functional characteristics, and by comparing it to other functionally related constructions. An overview of the ways in which the thesis has added to the field or extended previous research will now follow.

Two main findings have emerged from the thesis. Firstly, in mapping out the frequency distribution of both syntactic and functional categories of the introductory *it* pattern across NS status, academic discipline and levels of achievement, the thesis has investigated the relative importance of these factors. It found that the main differences reside in the comparison of academic disciplines and levels of achievement. The discipline-specific uses noted suggest that the pattern is important for adhering to the conventions of each discipline. The differences found between the lower-graded and higher-graded papers will be discussed in more detail below. Furthermore, the investigation of the stance-marking function of the pattern and other constructions highlighted the importance of also considering factors such as lexis and level of expertise in academic writing in studies of this kind.

Secondly, the investigations have enabled a clearer picture to be painted of how the investigated groups use the introductory *it* pattern and some related constructions in academic writing. The following five groups have been studied: low-achieving learners (L1 Swedish), high-achieving learners

(L1 Swedish), learners more generally (L1 Swedish and L1 French), high-achieving NS students and published expert writers. Both similarities and differences were noted.

With regard to similarities, a limited set of categories included the majority of the tokens in all the investigated groups. Syntactically, the three higher-level categories SVC (e.g. *it is possible to*), SV (e.g. *it appears that*) and SV_{pass} (e.g. *it can be seen that*) are the most frequent categories in the low-achieving learner data, the high-achieving learner data, as well as in the NS student data. If we add the more fine-grained COBUILD categories, we see that SVC can be expected to be realized mainly as ***it* V ADJ that** or ***it* V ADJ to-inf** (e.g. *it is interesting to/that*), SV as ***it* V that** (e.g. *it seems that*) and SV_{pass} as ***it* be V-ed that** (e.g. *it has been shown that*), since these four subpatterns showed the highest frequencies in both the high-achieving learner data and the expert data. Functionally, the attitude marker category (e.g. *it is important to*) was the uncontested most frequent category in all three of the investigated groups (low-achieving learners, high-achieving learners and NS students). Because of these similarities found across the investigated groups, the use of the pattern can be concluded to be relatively stable overall both syntactically and functionally.

There were, however, also differences between the groups. For example, the low-achieving learners tended to make particularly frequent use of the most frequent syntactic and functional category (SVC and attitude markers respectively). The learners' especially frequent use of certain instances of the introductory *it* pattern suggests that they tend to rely on *lexico-grammatical teddy bears* (cf. Hasselgren's (1994) use of "lexical teddy bears"). The learners' overuse of high-frequency realizations could also suggest that these are learned as wholes. Furthermore, while the learner-specific problems found should not be exaggerated, there were some clear differences between the learners and the reference groups (published writers or NS students) even for the most advanced learners, with regard to frequency, form and function. For example, the high-achieving learners were found to underuse the pattern to hedge claims (e.g. *it seems that*) compared to the NS students. Frequency differences were also noted for the other constructions. For example, the learners overused the investigated stance markers compared to the published writers. In particular, the learners overused stance markers that are associated with informal use, such as *probably*. Some evidence of L1 transfer was also found for high-frequency realizations such as *interestingly* and *it is interesting that*. The learners might thus benefit from some explicit teaching of these constructions; this will be returned to in section 8.2.

Furthermore, the thesis has extended previous research in four main ways. First, the thesis has investigated the frequency distribution of the syntactic types and subpatterns for which no frequencies have previously been presented, for example showing that the majority of the instances belong to a small set of different syntactic types or subpatterns, as discussed above. Se-

cond, at a more detailed level, the thesis has shown that the introductory *it* pattern is relatively invariable in the sense that a small number of high-frequency realizations made up the bulk of the tokens for each subpattern in the data (i.e. ***it* SEEM/APPEAR/BE/FOLLOW that** make up the vast majority of the ***it* V that** tokens, for example). Third, in also including investigation of the relatively understudied functional category *observations* (e.g. *it has been shown that*), the thesis could compare the relative frequencies of this category in relation to the stance-marking category (e.g. *it is surprising that*). Fourth, in investigating the pattern in relation to other, functionally similar constructions, the thesis has shown that there seem to be clear reasons why one construction is used instead of another. For example, unlike disjuncts, the introductory *it* pattern gives writers the option of including additional information, such as certain adverbs (*it is certainly interesting that* vs. **certainly interestingly*).

Finally, the thesis has made some methodological contributions, which will hopefully be useful not only for future studies investigating the pattern, but also for studies looking at other constructions. The thesis has added to existing syntactic classifications by using empirical data to further investigate and subclassify the instances of the introductory *it* pattern in (learner) academic writing. Furthermore, a new functional classification was developed with the aim of limiting the dependence on word semantics for classification, thereby increasing the replicability of the results. Finally, a model for identifying morphologically related stance markers in corpora that are not POS-tagged was designed.

8.2 Concluding remarks

This thesis has aimed to expand our knowledge of the introductory *it* pattern and some related constructions in academic discourse, with a particular focus on learner use. The conclusions drawn can serve to inform future research, as well as teaching practices.

With regard to future studies, the findings of the thesis offer several avenues for research. For example, the fact that there were discipline-specific uses of the pattern and the related constructions suggests that it would prove fruitful to take a broader genre and discipline perspective. Such studies could investigate what other factors might affect the use of the pattern and other constructions, for instance by applying frameworks such as *Construction Grammar* (e.g. Goldberg, 1995, 2006). Furthermore, future studies could explore whether the degree of variability of the introductory *it* pattern reported in the present thesis holds more generally for other constructions and patterns, possibly also outside the academic realm.

As regards teaching implications, the thesis has identified high-frequency realizations of the introductory *it* pattern that can be considered especially

important for EAP instruction. At a more general level, these include instances that contain an adjective phrase, especially those that have a stance-marking function, such as *it is interesting to note* and *it is important to remember*. Instances used for hedging claims, such as *it seems that*, *it appears that* and *it* [modal verb] *be the case that* also stood out as important for learners to be made aware of, as the learners made comparably infrequent use of these types. However, since the thesis has also shown that frequent use of the pattern does not necessarily entail proficient use of it, the main focus should, perhaps, be directed towards helping students make both appropriate and varied use of the pattern and other similar constructions.

The thesis has also detected problem areas for learners at different levels of achievement and with different L1s. The learners whose papers received a lower grade appeared to struggle more with balancing the frequency with which to use the pattern, as they tended to make especially frequent use of a small set of high-frequency tokens. Furthermore, as the thesis has identified L1-specific difficulties with regard to the pattern as well as to related constructions, it would be preferable to take a contrastive perspective on some of these forms, pointing out differences and similarities between English and the student's L1, whenever possible. Examples include the L1 Swedish students' tendency to place adverbs such as *possibly*, *presumably* and *probably* sentence-initially.

Finally, it is hoped that the results of this thesis will be useful for the subfield of Learner Corpus Research, as well as for EAP theory and practice. At a more general level, the results could also be helpful for grammars describing the introductory *it* pattern in academic writing.

Summary in Swedish / Sammanfattning på svenska

Denna sammanläggningsavhandling behandlar en engelsk konstruktion som här benämns ”the introductory *it* pattern”. Konstruktionen har två subjekt: ett pronomen (*it*) som inte har anaforisk referens som formellt subjekt och ett egentligt subjekt som ofta är en infinitivfras eller en *that*-sats. Konstruktionen har rapporterats vara speciellt vanlig i akademisk text och finns både i engelska och svenska (t.ex. *it is interesting to note the difference*; *det är intressant att notera skillnaden*). Avhandlingens huvudsyfte är att undersöka konstruktionens syntax och funktioner i akademiska texter. Även relevanta relaterade konstruktioner analyseras. Fokus ligger på att undersöka texter skrivna av studenter som inte har engelska som förstaspråk. Publicerade texter samt texter skrivna av studenter som har engelska som förstaspråk har använts som jämförelse. Avhandlingen består av en inledande kappa och fyra artiklar.

I avhandlingen har material från följande korpusar använts: ALEC, BATMAT, BAWE, LOCRA, MICUSP och VESPA. Flera olika faktorer har undersökts; bland annat kan nämnas modersmål (engelska eller ”annat”), vetenskapsgren (lingvistik eller litteratur) och åstadkommande (texter med lägre vs. högre betyg). Artikel 1 och 2 undersöker konstruktionens syntax, medan Artikel 3 och 4 studerar dess olika funktioner. I mer detalj vidareutvecklar de två första artiklarna tidigare syntaktiska klassificeringssystem med syfte att undersöka konstruktionens uppbyggnad och användning. Artikel 1 syftar till att ge en mer generell översikt och Artikel 2 ger en mer detaljerad bild. I Artikel 3 utvecklas en funktionell klassificering som sedan används för att kategorisera konstruktionens olika användningsområden. I Artikel 4 studeras även relevanta besläktade konstruktioner som används för att uttrycka en ståndpunkt (adverb såsom *importantly* och substantiv + prepositionsfras-konstruktioner såsom *the importance of*).

Studierna visade att när det gäller såväl syntax som funktion är huvudkonstruktionen, the introductory *it* pattern, relativt invariabel då huvuddelen av beläggen tillhörde ett mindre antal kategorier. Studenterna som inte hade engelska som modersmål, speciellt de studenter vars texter fick ett lägre betyg, använde särskilt många högfrekventa realisationer av konstruktionen. Resultaten visar dock att modersmålet endast är en av flera viktiga faktorer som kan påverka konstruktionens användning; vikten av att inte begränsa

undersökningar av detta slag till jämförelser mellan modersmål lyfts därför fram i avhandlingen. Avhandlingen har, genom att undersöka många olika faktorer och hur de påverkar konstruktionens uppbyggnad och funktion, resulterat i en mer komplett bild av hur konstruktionen används i akademiska texter. En tillämpning av resultaten är att de skulle kunna användas för andraspråksinlärning.

References

Corpora

ALEC (Advanced Learner English Corpus). Corpus compiled by Tove Larsson at Uppsala University, Sweden.

BATMAT. Corpus compiled by Signe-Anita Lindgrén at Åbo Akademi University, Finland.

BAWE (British Academic Written English). 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/

BNC (the British National Corpus). Available online from http://www.natcorp.ox.ac.uk/.

COCA (Corpus of Contemporary American English). Available online from: http://corpus.byu.edu/coca/

ICE-GB (British component of the International Corpus of English). Corpus coordinated by the Survey of English Usage. http://www.ucl.ac.uk/english-usage/projects/ice-gb/

ICLE (International Corpus of Learner English). Corpus coordinated at Université catholique de Louvain. https://www.uclouvain.be/en-cecl-icle.html

LOCRA (Louvain Corpus of Research Articles). Corpus under compilation at the Centre for English Corpus Linguistics at Université catholique de Louvain. https://www.uclouvain.be/en-cecl-locra.html

MICASE (Michigan Corpus of Academic Spoken English). (2002). Rita C. Simpson, Sarah L. Briggs, Janine Ovens & John M. Swales. Ann Arbor, MI: The Regents of the University of Michigan. Available online from http://quod.lib.umich.edu/m/micase/

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

SEPC (Swedish-English Parallel Corpus). (2001). Corpus compiled by Bengt Altenberg, Karin Aijmer and Mikael Svensson at the universities of Lund and Gothenburg. http://www.sol.lu.se/engelska/corpus/corpus/espc.html

VESPA (Varieties of English for Specific Purposes dAtabase). Corpus administered at the Centre for English Corpus Linguistics at Université catholique de Louvain. http://www.uclouvain.be/en-cecl-vespa.html

Works cited

Ädel, A. (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.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, 31 (2), 81–92.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer* (pp. 80–93). London: Longman.

Aston, G. (2008). "It's only Human…". In A. Martelli & V. Pulcini (Eds.), *Investigating English with corpora: Studies in honour of Maria Teresa Prat* (pp. 343–354). Monza: Polimetrica International Scientific Publisher.

Biber, D. (2006a). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5 (2), 97–116.

Biber, D. (2006b). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

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

Boas, F. (1940). *Race, language and culture*. New York: Macmillan.

Boström Aronsson, M. (2005). *Themes in Swedish advanced learners' written English*. PhD dissertation, University of Gothenburg.

Calude, A. S. (2008). Clefting and extraposition in English. *ICAME Journal*, 32, 7–34.

Charles, M. (2000). The role of an introductory *it* pattern in constructing an appropriate academic persona. In P. Thompson (Ed.), *Patterns and perspectives: Insights into EAP writing practices* (pp. 45–59). Reading: University of Reading, CALS.

Charles, M. (2003). 'This mystery…': A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes*, 2 (4), 313–326.

Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes*, 25 (3), 310–331.

Collins, P. (1994). Extraposition in English. *Functions of Language*, 1, 7–24.

Couper-Kuhlen, E., & Thompson, S. A. (2008). On assessing situations and events in conversation: 'Extraposition' and its relatives. *Discourse Studies*, 10 (4), 443–467.

Dixon, R. M. W. (1991). *A new approach to English grammar, on semantic principles*. Oxford: Clarendon.

Francis, G., Hunston, S., & Manning, E. (1996). *Collins COBUILD grammar patterns 1: Verbs*. London: HarperCollins.

Francis, G., Hunston, S., & Manning, E. (1998). *Collins COBUILD grammar patterns 2: Nouns and adjectives*. London: HarperCollins.

Gilquin, G. (2002). Automatic retrieval of syntactic structures: The quest for the Holy Grail. *International Journal of Corpus Linguistics*, 7 (2), 183–214.

Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction*, 1 (1), 41–61.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), *Languages in contrast: Papers from a symposium of text-based cross-linguistic studies, Lund 4–5 March 1994* (pp. 37–51). Lund: Lund University Press.
Granger, S. (2015). Contrastive Interlanguage Analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1 (1), 7–24.
Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International Corpus of Learner English. Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.
Gray, B., & Biber, D. (2012). Current conceptions of stance. In K. Hyland & C.S. Guinda (Eds.), *Stance and voice in written academic genres* (pp. 15–33). Basingstoke: Palgrave Macmillan.
Gries, S. Th., & Otani, N. (2010). Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal*, 34, 121–150.
Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes*, 4 (3), 257–277.
Hacker, D., & Sommers, N. I. (2014). *The Bedford handbook* (9th ed.). Boston: Bedford/St. Martins.
Halliday, M.A.K. (1994). *An introduction to functional grammar* (2nd ed.). London: Edward Arnold.
Hasselgård, H. (2009). Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 121–139). Amsterdam: John Benjamins.
Hasselgård, H. (2015). Lexicogrammatical features of adverbs in advanced learner English. *International Journal of Applied Linguistics*, 166 (1), 163–189.
Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, 4 (2), 237–260.
Hatzitheodorou, A-M., & Mattheoudakis, M. (2009). "It is more than true that television reproduces life": The effect of Greek rhetorical conventions on Greek learners' academic writing in English. In T. Tsangalidis (Ed.), *Selected papers from the 18th International Symposium on Theoretical and Applied Linguistics* (pp. 167–176). Thessaloniki: Monochromia.
Herriman, J. (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.
Herriman, J. (2013). The extraposition of clausal subjects in English and Swedish. In K. Aijmer & B. Altenberg (Eds.), *Advances in corpus-based contrastive linguistics: Studies in honour of Stig Johansson* (pp. 233–260). Amsterdam: John Benjamins.
Heuboeck, A., Holmes, J., & Nesi, H. (2008). *The BAWE corpus manual*. Available online from http://www.reading.ac.uk/internal/appling/bawe/BAWE.documentation.pdf, accessed in March, 2014.
Hewings, A., & Hewings, M. (2004). Impersonalizing stance: A study of anticipatory 'it' in student and published academic writing. In C. Coffin, A. Hewings &

K. O'Halloran (Eds.), *Applying English grammar: Functional and corpus approaches* (pp. 101–116). London: Hodder Arnold.
Hewings, M., & Hewings, A. (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, R. D., & Pullum, G. K. (2002*). The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
Hunston, S., & Thompson, G. (Eds.). (2000). *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press.
Hyland, K. (1996). Talking to the academy: Forms of hedging in science research articles. *Written Communication*, 13 (2), 251–281.
Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. London: Longman.
Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, 7 (2), 173–192.
Hyland, K. (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, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18 (1), 41–62.
Hyland, K., & Tse, P. (2005). Evaluative *that* constructions: Signalling stance in research abstracts. *Functions of Language*, 12 (1), 39–63.
Jespersen, O. (1927). *A modern English grammar on historical principles.* Vol. III. Heidelberg: Carl Winter.
Jespersen, O. (1937). *Analytic syntax*. Copenhagen: Levin & Munksgaard.
Kaatari, H. (forthcoming). Variation across two dimensions: Testing the Complexity Principle and the Uniform Information Density Principle on adjectival data. *English Language and Linguistics*, 20 (3), 533–558.
Kaltenböck, G. (1999). Which *it* is it? Some remarks on anticipatory *it*. *Vienna English Working Papers*, 8 (2), 48–71.
Kaltenböck, G. (2002). That's *it*? On the unanticipated 'controversy' over anticipatory *it*. A reply to Aimo Seppänen. *English Studies*, 83 (6), 541–550.
Kaltenböck, G. (2003). On the syntactic and semantic status of anticipatory *it*. *English Language and Linguistics*, 7 (2), 235–255.
Kaltenböck, G. (2005). *It*-extraposition in English: A functional view. *International Journal of Corpus Linguistics*, 10 (2), 119–159.
Labov, W. (1984). Intensity. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 43–70). Washington: Georgetown University Press.
Larsson, T. (2014). Introducing the Advanced Learner English Corpus (ALEC): A new learner corpus. Poster presented at the *2014 LOT Winter School*, VU University Amsterdam, Amsterdam, The Netherlands, 20 January, 2014.
Larsson, T. (2016). The introductory *it* pattern: Variability explored in learner and expert writing. *Journal of English for Academic Purposes*, 22, 64–79.
Larsson, T. (forthcoming). A syntactic analysis of the introductory *it* pattern in non-native-speaker and native-speaker student writing. In M. Mahlberg & V. Wiegand (Eds.), *Corpus linguistics, context and culture*. Berlin: De Gruyter Mouton.

Larsson, T. (under review a). A functional classification of the introductory *it* pattern: Investigating academic writing by non-native-speaker and native-speaker students.

Larsson, T. (under review b). *The importance of*, *it is important that* or *importantly*? The use of morphologically related stance markers in learner and expert writing.

Lee, D., & Chen, S. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, 18 (4), 149–165.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 8–29). London: Longman.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991* (pp. 105–122). Berlin: De Gruyter Mouton.

Lindgrén, S-A. (2015). *The BATMAT Corpus Version 1.1*. English Language and Literature, Åbo Akademi University, Åbo.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburgh: Edinburgh University Press.

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

Mak, K. T. (2005). The dynamics of collocation: A corpus-based study of the phraseology and pragmatics of the introductory *it*-construction. Ph.D. Thesis, The University of Texas at Austin.

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Basingstoke: Palgrave Macmillan.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Meyer, C. F. (2015). Corpus-based and corpus-driven approaches to linguistic analysis: One and the same? In I. Taavitsainen, M. Kytö, C. Claridge & J. Smith (Eds.), *Developments in English: Expanding electronic evidence* (pp. 14–28). Cambridge: Cambridge University Press.

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

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

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

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

Müller, S. (2005). *Discourse markers in native and non-native English discourse*. Amsterdam: John Benjamins.

Neff, J., Ballesteros, F., Dafouz, E., Martínez, F., & Rica, J-P. (2003). Formulating writer stance: A contrastive study of EFL learner corpora. In D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference. UCREL Technical Papers 16* (pp. 562–571). Lancaster University: UCREL.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.
Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes*, 21, 60–71.
Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130–149.
Paquot, M., Hasselgård, H., & Oksefjell Ebeling, S. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead*. [Corpora and language in use – Proceedings 1] (pp. 377–388). Louvain-la-Neuve: Presses universitaires de Louvain.
Peacock, M. (2011). A comparative study of *introductory it* in research articles across eight disciplines. *International Journal of Corpus Linguistics*, 16 (1), 72–100.
Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107–118). London: Longman.
Poos, D., & Simpson, R. (2002). Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In R. Reppen, S. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 3–23). Philadelphia: John Benjamins.
Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data analysis*. San Diego: Academic Press.
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.
R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Ramhöj, R. (2016). *On clausal subjects and extraposition in the history of English*. PhD dissertation: University of Gothenburg, 2016. Gothenburg.
Reilly, J., Zamora, A., & McGivern, R.F. (2005). Acquiring perspective in English: The development of stance. *Journal of Pragmatics*, 37 (2), 185–208.
Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, 7, 140–162.
Römer, U., & O'Donnell, M. B. (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.
Rosch, E. (2009). Categorization. In D. Sandra, J. Östman & J. Verschueren (Eds.), *Cognition and pragmatics* [Handbook of pragmatics highlights, Vol. 3] (pp. 41–52). Philadelphia: John Benjamins.
Rosenbaum, P. S. (1967). *The grammar of English predicate complement constructions*. Cambridge: MIT Press.
Rowley-Jolivet, E., & Carter-Thomas, S. (2005). Genre awareness and rhetorical appropriacy: Manipulation of information structure by NS and NNS scientists in the international conference setting. *English for Specific Purposes*, 24 (1), 41–64.
Scott, M. (2012). WordSmith Tools version 6. Liverpool, UK: Lexical Analysis Software.
Scott, M., & C. Tribble. (2006). *Textual patterns. Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Seppänen, A. (1999). Extraposition in English revisited. *Neuphilologische Mitteilungen*, 100 (1), 51–66.
Seppänen, A., Granath, S., & Herriman, J. (1995). On so-called 'formal' subjects/objects and 'real' subjects/objects. *Studia Neophilologica*, 67 (1), 11–19.
Seppänen, A., & Herriman, J. (2002). Extraposed subjects vs. postverbal complements: On the so-called obligatory extraposition. *Studia Neophilologica*, 74 (1), 30–59.
Shaw, P. (2004). Sentence openings in academic economics articles in English and Danish. *Nordic Journal of English studies*, 3 (2), 67–84.
Sinclair, J. (1991). *Corpus*, *concordance*, *collocation*. Oxford: Oxford University Press.
Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
Swales, J. M., & Burke, A. (2003). 'It's really fascinating work': Differences in evaluative adjectives across academic registers. In P. Leistyna & C. F. Meyer (Eds.), *Corpus analysis: Language structure and language use* (pp. 1–18). New York: Rodopi.
Swan, M. (2005). *Practical English usage*. (3rd ed.). Oxford: Oxford University Press.
Thompson, P. (2009). Shared disciplinary norms and individual traits in the writing of British undergraduates. In M. Gotti (Ed.), *Commonality and individuality in academic discourse* (pp. 53–82). Bern: Peter Lang.
Thompson, G. & Ye, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, 12 (4), 365–382.
Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.
UCREL Clustertool. Available online from http://corpora.lancs.ac.uk/clustertool/index.php, accessed in March 2015.
Zhang, G. (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

Table A1 provides an overview of the relevant abbreviations from the COBUILD grammars (Francis et al., 1996, 1998). Abbreviations that are only found in the second grammar (i.e. in Francis et al., 1998) are marked with an asterisk.

Table A1. List of abbreviations used in the COBUILD grammars with explanations added.

Abbreviation	**Explanation**
adj	an adjective phrase
ADJ	an adjective*
amount	a word or phrase indicating the amount of something
be V-ed	the lemma BE followed by a past participle
it	introductory *it*
-ing	a clause beginning with the *-ing* form of a verb
n	a noun phrase
N	a noun*
poss	a possessive determiner*
prep	a prepositional phrase
since	a finite clause beginning with *since**
that	a *that* clause
to-inf	a clause beginning with a *to*-infinitive form of a verb
V	a verb group
wh	a finite clause beginning with a *wh*-word
what/how	a clause beginning with *what* or *how**
when/if	a finite clause beginning with *when* or *if*