Complex anaphora with *this*: variation between three written argumentative genres

Peter Crompton

*American University of Sharjah*

**Abstract**

The concept of complex anaphora - ‘nominal expressions referring to propositionally structured referents’ (Consten *et al.*, 2007) - makes a useful distinction between a text-structuring function, one important to argumentative text, and the forms used to accomplish the function. Because complex anaphors often contain the demonstrative *this*, the contexts of all *this* tokens in three corpora of written argumentation – research articles, editorials, and student essays – were analysed in order to identify instances of complex anaphora. While the frequencies of pronoun use for complex anaphora were similar, the frequencies of determiner use varied, as did placement of anaphors within their host sentence, determiners appearing non-initially much more often than pronouns in all corpora, particularly editorials. Overall, there was greater variation between the patterns of use in research articles and editorials than between these and student essays.

**1. Introduction**

**1.1 Background**

The research described in this paper investigates the frequency of a particular kind of cohesion, illustrated in the following text extracts (suggested antecedent underlined, referring expression in italics):

---

1 The use of the term ‘antecedent’ is problematic: Consten *et al.* (2007) note that there may in some cases not be an antecedent for a complex anaphor but there will be what they call an ‘anchor’, in that the text may not contain an explicit proposition but will enable the hearer to infer a propositional referent. On the basis of similar evidence of ‘non-canonical’ anaphora Cornish (2010) argues a similar case for recognizing a more general category than antecedent, ‘antecedent trigger’. In this paper for economy, and because the term is so widely used by other scholars, I use ‘antecedent’.
Complex anaphora with this: variation between three written argumentative genres. Corpora 12(2).

(1) If we preserve it now and maintain this preservation, a whole ecosystem will be saved. (Student Essay)
(2) During the course of the next five weeks, in 32 countries, little will matter more than the World Cup. There will be some who argue that this shows our badly skewed priorities… (Editorial)

Such cohesion has been given many different names in the literature, including extended reference, reference to fact (Halliday & Hasan, 1976), situational anaphora (Fraurud, 1992), shell nouns (Schmid, 2000), abstract object anaphora (Asher, 1993), discourse deixis (Webber, 1991). The name used in this paper will be complex anaphora. Complex anaphors have been defined as ‘nominal expressions referring to propositionally structured referents (such as propositions, states, facts and events) while introducing them as unified entities into a text world model’ (Consten, Knees, & Schwarz-Friesel, 2007: 81). Thus, in (1) and (2) the full noun phrase this preservation and the pronoun this, respectively refer to the propositions contained in the clauses italicised and introduce these propositions as unified entities. By contrast, in the following extracts, the anaphors refer to non-propositionally structured referents:

(3) So the choice of the best model also depends on the error measurement, not only on data and horizon. Obviously, this error measurement depends on the ultimate purpose of the forecasting procedure. (Research Article)
(4) Mr Field’s appointment is particularly significant… This was the man whom Tony Blair entrusted with “thinking the unthinkable” … (Editorial)

To research text structure and in particular that of argumentative texts, I suggest that complex anaphora is a natural focus of enquiry. Since Halliday and Hasan (1976) identified conjunction as a form of cohesion, much research has been undertaken into the presence or absence in texts of markers of conjunctive or connective relations, particularly in comparisons of texts by proficient and novice writers (e.g., Sloan, 1984; Milton & Tsang, 1993; Granger & Tyson, 1996; Altenberg & Tapper, 1998; Bolton, Nelson, & Hung, 2003; Chen, 2006); complex anaphora may prove to be of equal or even greater importance. Studying the frequency of complex anaphora, however, is more difficult than that of conjunction because of the polysemy of the forms used to accomplish it; as illustrated in (3) and (4), this, for example, may also refer to non-complex entities. The absence of a one-to-one mapping of form and function complicates the process of identifying the forms which accomplish complex anaphora. Nevertheless, corpus-assisted manual analysis of the kind discussed in this paper reveals that complex anaphora is a high-frequency phenomenon in argumentative text.

If complex anaphora is important to the structure of argumentative texts, it is likely to be important to academic writing instruction. Professional academic genres such as the research article have long been recognised (e.g., Bazerman, 1988; Myers, 1989, Swales, 1990) as argumentative in their rhetorical purpose; this recognition informs instruction in the

---

2 This nomenclature is usefully summarised in Schwarz-Friesel and Consten (2011).
humbler pedagogic genres, as illustrated by the opening of Harvard University Writing Center’s ‘Overview of the Academic Essay’: ‘A clear sense of argument is essential to all forms of academic writing…’ (Duffin, 1997). The concept of ‘syntactic maturity’ (Hunt, 1970) common in much writing research associates the occurrence and frequency in a text of particular linguistic features, such as embedded clauses, with development in writing proficiency; the presence and frequency of complex anaphora might similarly serve to indicate development in writing proficiency, but at the linguistic level of discourse rather syntax. Academic writing instruction has for a long time drawn attention to the forms and functions of logical connectors. The more we know about the forms and functions of complex anaphora in academic writing, the better placed we will be bring these to the attention of developing academic writers.

What formal linguistic features can be used to identify complex anaphora? As will be shown in the review of relevant research below, to answer this question many researchers have sought to identify a set of open class lexical items, including general nouns (Halliday & Hasan, 1976), type 3 vocabulary (Winter, 1982), anaphoric nouns (Francis, 1986), labels (Francis, 1994) carrier nouns (Ivanic, 1991), shell nouns (Schmid, 2000), and signalling nouns (Flowerdew, 2006 & 2010). However, I argue here that complex anaphora is not in practice performed by open class lexical heads but by a set of closed class items which accompany them. Among this closed class set, a careful reading of previous research (reviewed below) suggests that the proximal demonstrative this, both as determiner and head, features prominently. Having made the case for the utility of this in identifying complex anaphora, the paper reports some further research to find out (i) how far this is actually used to perform complex anaphora in three corpora of argumentative texts relevant to academic writing, and (ii) how much variation there is between these corpora in the form of this (pronoun vs. determiner) and in its placement (initial vs. non-initial).

For convenience in the rest of the paper focus of the paper, complex anaphora will be referred to as CA and complex anaphora with ‘this’ as CAWT. For convenience - and not to imply any conceptual unity - all anaphora with this that is not complex will be designated as simple anaphora, with the corresponding abbreviations SA and SAWT.

1.2 Complex anaphors in nominal form

Although little research appears to have been carried out on the larger concept of CA, which embraces the use of both pronouns and lexical noun phrases, the use of the latter as complex anaphors, albeit using different terminology, has been investigated by various scholars. Francis (1994) used the term retrospective label for referring expressions similar to that illustrated in (1). According to Francis, labels ‘encapsulate or package a stretch of discourse’. The term encapsulation was coined by Sinclair who defined it as a process which ‘reclassifies a previous sentence by “demoting” it to an element of the structure of the new sentence’ (Sinclair, 1993: 8) and argues that encapsulation is effected by both ‘logical acts’, accomplished by conjunction (e.g., and, however) and ‘deictic acts’ (e.g. this, this subject).
In addition to encapsulation, CA can be seen as overarching the phenomenon described within systemic-functional grammar as *nominalisation* or *grammatical metaphor* (Halliday, 1994). In example (1), for instance, *this preservation* is a nominalization of the preceding clause, *if we preserve it*. In practice, systemic-functional analysts seem to limit nominalisation to lexically headed expressions, but arguably, insofar as a process which would be construed congruently as a clause is construed incongruently as a noun phrase, all CA, whether the anaphor is lexical noun phrase or a pronoun, is grammatical metaphor.

The ‘major criterion’ for Francis’s retrospective label is ‘that there is no single nominal group to which it refers: it is not a repetition or a “synonym” of any preceding element’ (1994: 85). However, in practice Francis only recognises as labels lexically-headed noun phrases - like *this preservation* in (1) - and not pronouns - like *this* in (2). Francis’s reason for not considering pronouns as labels is not fully articulated and is at first sight puzzling, particularly so in that her description of labels having ‘the ability to refer to “text as fact”’ is couched in terms borrowed from Halliday and Hasan’s (1976) account of *extended reference*, which Halliday and Hasan attribute solely to three pronouns - *it*, *this* and *that*. For Francis, however, an additional criterion for being a label, beyond the nature of the referent, is that it fulfil the function of ‘naming the clause or clauses it replaces’ (Francis, 1994: 85). By ‘naming’ Francis seems to intend that a label not only reiterates the meaning of the referent, but, in the choice of a lexical head, creates additional meaning, ideational and/or interpersonal: ‘the label indicates to the reader exactly how that stretch of discourse is to be interpreted’ (85), for example, a writer can present a complex referent as an event, problem, or fiasco and elaborate such meanings further with modifiers (e.g., tragic event).

From a corpus research perspective it is noteworthy that, although Francis mentions other reference forms forming parts of labels (*the*, *that*, *such*, *another*) and comparative forms (e.g., a similar argument, a more bombastic explanation), of the twenty seven examples of retrospective labels supplied in Francis (1994), all but six contain *this* as determiner. Three of these six occur in clausal predicates of subjects comprising pronominal *this*, for example:

(5) Anthony Burgess thinks hero worship is peculiar to the British. He explains it ....While *this* is an old-fashioned diagnosis..... (cited in Francis, 1994: 86, my use of bold)

In terms of CA, all uses of *this* to refer to complex entities are complex anaphors, regardless of whether their host clause contains nominal predicates. Francis does not give details of her discovery procedures for identifying labels but her examples suggest that examining tokens of *this* is likely to net most instances of label.

Overall, to relate the two concepts, we can conceptualise labels as only a subset of the class of complex anaphors. Labels do not include pronouns and there is a further semantic qualification for labels - that the head be ‘inherently unspecific’ (Francis: 89): Francis likens labels to ‘general nouns’ (Halliday & Hasan, 1976), ‘procedural vocabulary’ (Widdowson, 1983) and ‘all-purpose nouns’ (Peters, 1985). How ‘unspecific’ is to be operationalised is not clear; the head of the anaphor in (1) earlier in this paper, *preservation*, is abstract rather than

Concrete, but it does not appear in the lists featured in Francis (1994) and appears different in kind from the examples given in either her various categories of ‘metalinguistic’ label or the ‘general’ category (e.g., accident, achievement, action).

Two other theorised constructs which overlap to some extent with CAWT are shell nouns (Schmid, 2000; Hunston & Francis, 2000) and signalling nouns (Flowerdew, 2006 & 2010). As their names suggest, like labels, these constructs exclude pronouns. There are further qualifications which exclude some CAWT tokens from being categorised as shell or signalling nouns. A discovery procedure stipulated by both Schmid (2000) and Hunston and Francis (2000) for shell nouns requires that lexical heads be capable of operating cataphorically in structures such as postnominal clauses (e.g., the fact that something happened). Schmid further requires exponents of his shell noun concept to be capable of operating in complement clauses (e.g., the problem is that something happened). The head noun in example (1) of CAWT in this paper, preservation, does not meet these qualifications. Flowerdew’s signalling noun concept is also limited to ‘abstract nouns’ (2006: 348).

In order to clarify the difference between CA and these other constructs, it is perhaps worth noting that designation of a lexical form as a label, shell noun, or signalling noun can be no more than a specification of potential: such forms may appear in NPs which are not referring expressions and therefore not cohesive:

(6) Oh, and finally, might it be an idea for a political writer to do the interview? (BNC CH1)

(7) AND two thirds complain that underfunding is still a major problem. (BNC CH1)

Conventionally, it would appear that almost any noun, abstract or otherwise, has the potential to serve as part of a complex anaphor, as the following examples illustrate:

(8) The subjective inner world of animals was not denied and an evolutionary development of sensed-worlds was assumed because of the complexity indicated by the increasing sophistication of perceptual and communicational mechanisms, but only the behaviour could be charted and recorded in the actual passage of time. *This two-sidedness* in ethology harks back in fact to the seventeenth-century dualism of Descartes. (BNC AMG)

(9) THE US, Soviet Union, Britain, France and China have for many years had huge arsenals of strategic nuclear weapons in the form of ballistic missiles and manned bombers. These form today’s big stick: their owners hope that, by possessing the ability to kill tens of millions of people and destroy the fabric of modern industrial society, they will prevent an enemy from being stupid enough to risk all by waging all-out war. Beneath *this umbrella of deterrence* are tactical weapons. (BNC B7N)

---

3 Text examples in this paper beginning with the prefix ‘BNC’ are taken from the online version of the British National Corpus (Davies, 2008- ).

Within the overarching concept of cohesion, then, the category complex anaphor encompasses a broader set of forms than the noun-based categories label, shell noun, and signalling noun. The differences can be summarised thus:

**formally** complex anaphors are not confined to lexically headed noun phrases (i.e., they may be pronouns)

lexically headed complex anaphors are not confined to a finite set of heads capable of occupying certain syntactic slots

**semantically** complex anaphors are not confined by the concept of unspecificity.

The CA concept (Consten et al., 2007) offers a useful theoretical category for research on text structure. Related cohesive concepts like label, shell noun, and signalling noun have tended to conceptualise both the underlying abstraction (or complexity) and the cohesion of anaphors as deriving from lexis accompanying the anaphor (e.g., idea, argument). CA, by contrast, correctly in the view of this researcher, locates the abstraction in the referent and identifies the cohesion as deriving from the regular anaphoric function of closed set items (e.g., this, such), irrespective of whether the referent is simple or complex.

1.3 Demonstratives in complex anaphors

This section will review previous research on reference which sheds light on CA as performed by demonstratives. Among pronouns, demonstratives appear to be the most likely to be used for CA. As noted earlier, Halliday and Hasan’s (1976) concept of extended reference is similar to CA. Extended reference is restricted in their monograph to it and the demonstrative pronouns (1976: 66). In an early corpus study they analysed the 51 demonstrative pronouns in two chapters of Alice’s Adventures in Wonderland and found over half (31) to be performing extended reference. Halliday and Hasan describe extended reference as ‘one of the major cohesive devices of the English language’ and particularly characteristic of spoken English (1976: 67). Researchers on the referents of referring expressions within the Givenness Hierarchy framework have also identified a ‘strong association of demonstrative and non-nominal antecedents’ (Hedberg, Gundel, & Zacharski, 2007, citing research by Webber, 1991, and Hegarty, Gundel, & Borthern, 2002).

Gray’s (2010) study on this and these in a 480,000 word corpus of published research articles analysed the antecedents of both pronouns and determiners. Gray chose only to examine sentence-initial this and these tokens, probably for reasons discussed in §1.6, so it is not clear how generalizable her findings are to non-initial tokens. Gray found that 87% of the pronouns had clausal or larger than clausal antecedents and from the data she supplies it appears that 52% of noun phrases had such antecedents. In our terms, then, it would appear that in sentence-initial position both pronoun and determiner this are more likely to be performing CA than simple anaphora. Secondly, it would appear that initial pronoun this is even more likely to be performing CA than initial determiner this.
Botley’s (2006) study is the largest corpus-based study of CA as performed by demonstratives, both pronouns and determiners, in all sentence positions, so far conducted. Botley does not use the term CA but his ‘indirect anaphora’ (also ‘indirectly recoverable anaphora’) – the antecedent is non-nominal, the relationship between antecedent and anaphor is ‘not one of co-reference’ (2006: 24) - seems in practice to be a different term for the same phenomenon, although it includes cataphora. Botley identified all instances of demonstratives used for indirect anaphora (IA) (in total, 648) in three 100,000 word corpora representing different text types: news reports, fiction, and spoken argumentation (Canadian Hansard proceedings). Among Botley’s overall findings, the following are most relevant to the research described here:

- anaphora tokens ‘vastly’ outnumbered cataphora tokens
- determiners were more common than pronouns
- generally, proximal demonstratives were more common than distal demonstratives
- generally, singular demonstratives were more common than plural demonstratives


The studies outlined in this section all suggest that the demonstratives are often used to perform CA. As Botley’s study suggests that among the demonstratives, this is the most prototypical exponent of CA, the current study will focus on this.

1.4 Comparative frequencies of pronominal and nominal this

The comparative frequencies of the two forms of this, nominal and pronominal, to perform CA have seldom been a focus of study in their own right. However, previous research into a number of reference forms and cohesion-related concepts has revealed interesting patterns worth exploring further.

This is one of the commonest words in English, ranked 24th in the whole British National Corpus, 30th in the spoken component and 26th in the written (Leech et al., 2001). Biber, Johansson, Leech, Conrad and Finegan (1999) compared the frequency of this appearing in four broad text types in their research corpus – conversation, fiction, news, and academic. In the academic corpus this appeared much more frequently - 66% more frequently as a pronoun and 100% as a determiner - than in the other corpora.

There is a small literature on the use of demonstratives in written composition, usefully summarised in Swales (2005) and Wulff, Römer, and Swales (2012). This literature has been driven by pedagogical rather than descriptive concerns – what should novice writers be told about the existence of an apparently rhetorical choice between two demonstrative forms, pronoun and full noun phrase? Significantly for the present study, the composition literature frames this choice is terms of the determiner and not the noun: pronominal this is described as ‘unattended this’ and determiner this as ‘attended this’ (Geisler et al. 1985). In a syntactic analysis, a lexical noun would be conceptualised as the head or nucleus of the noun phrase and any determiner an optional complement; the ‘attended vs. unattended’ perspective, by contrast, reflects a semantic analysis, with the demonstrative form conceptualised as nucleus
and any ‘attending’ lexical noun as optional complement. For Geisler et al. (1985), the decision whether to use this unattended or attended is seen as involving a trade-off between two rhetorical ideals, namely avoiding redundancy and avoiding ambiguity. As documented by Gray and Cortes (2011), there is in fact a prescriptive tradition in academic style guides in favour of the second ideal: the APA style manual, for example, counsels writers to ‘eliminate ambiguity’ by replacing pronouns ‘when they refer to a previous sentence’ with full noun phrases such as ‘this test, that trial’ (American Psychological Association, 2000: 37).

Corpus-based studies, however, have noted that in practice published academic writers often employ unattended demonstratives (i.e., pronouns). Gray and Cortes (2011) found that in research articles (RAs) in Applied Linguistics and Engineering pronoun forms comprised nearly 20% of all tokens of this/these in all sentence positions. Römer and Wulff (2010), analyzing this in all sentence positions in the MICUSP corpus of written student essays (1.25 million words), found that 25% of tokens were pronouns. This proportion was ‘very stable’ across assumed level of author writing proficiency (based on undergraduate/graduate status and year of study) (2010: 118). They also found that apart from Philosophy there were ‘no major discipline-specific deviations from the general distribution’ of pronoun and determiner this (2010: 118). Boettger and Wulff (2014) looking at all tokens of this in a corpus of texts by students taking technical communication courses found pronoun forms comprised 33%.

Some studies confine their analyses to sentence-initial this. Swales (2005) looked at sentence-initial instances of this a corpus of academic RAs and found 36% of tokens were pronouns. He also found some interdisciplinary range (from 25% in dentistry to 56% in philosophy). Wulff et al. (2012) looking at sentence-initial this in MICUSP (student essays) found an even larger proportion of pronouns, 43%.

To summarise, the combined findings of Swales (2005), Römer and Wulff (2010), Gray and Cortes (2011), and Wulff et al. (2012) suggest, that in both published and student academic writing, between a fifth and a quarter of all this tokens are pronouns. If only sentence-initial tokens are considered this proportion shifts to more than a third. Student writers appear to use a greater proportion of pronouns than do published academic writers, both overall and sentence-initially. So far, however, little corpus research on academic text has analysed the referents of demonstratives, particularly not of pronouns. The next section reviews research which may shed light on CA using this in academic text.

1.5 Variation in frequency of CA by text type and author-proficiency

This section will review what previous research suggests about register and genre variation in the use of CAWT. Firstly, let us note that no corpus studies appear to have been conducted on CA in informal spoken discourse. Such research would be very useful in order for us to understand whether and how far CA differs between spoken and written argumentation.

Although Francis did not carry out formal variation studies, she notes that labels are ‘extremely common in the press and in all discourse of an argumentative nature’ (1994: 100) and Schmid observes that ‘shell nouns are more frequent in texts on abstract topics that are
written in a neutral or formal style and serve an expository or argumentative function’ (2000: 379). As we have seen, Botley’s study of Indirect Anaphora (IA) involved three different text types - news reports, fiction, and formal spoken argumentation. Botley found:

- there was a much higher proportion of IA in the argumentative corpus (Canadian Hansard) than the others
- statistically, differences in frequency of singular demonstratives between the three corpora were particularly significant (2006: 85).

Biber et al. (1999) did not analyse the referents of the demonstratives in their corpus, but bearing in mind the ‘strong association of demonstrative and non-nominal antecedents’ (Hedberg, Gundel, & Zacharski, 2007) cited earlier, in the absence of qualitative research on referents, the much greater frequency of this – as both pronoun and determiner – in academic texts gives ground for believing that more CAWT occurs in academic than in non-academic texts.

Within Petch Tyson’s (2000) study of demonstratives in corpora of student essays by native English speaker (NES) and non-native English speaker (NNES) authors of various L1s, she carried out a small-scale study of CA, which, following Fraurud (1992), she refers to as situation reference. Her findings suggested that the NNES students used frequencies of demonstrative pronouns to perform CA broadly similar to those of the NES students but used considerably fewer demonstratives determiners. Notably, the NES students used more this noun phrases than any other L1 group. Petch-Tyson concludes that higher levels of demonstrative noun phrases performing CA are associated with greater writing proficiency, the latter being impacted possibly by different L1 discourse patterns.

Overall, then, based on the research described, there seem to be grounds for believing that CAWT is particularly associated with formal texts performing argumentation, and that texts by NNES student writers may contain less CAWT than texts by NES student writers. Whether in general academic writing contains more CAWT than other kinds of formal argumentation or in particular expert academic writing contains more than student academic writing has not, to my knowledge, been researched. The current study will attempt to make these comparisons.

1.6 The placement of complex anaphors within their host sentences

Although providing no corpus evidence, Francis states categorically that labels occur in theme (1994: 88), that is, clause-initially. Yamasaki (2008: 80) notes that labels can appear in other positions but that placement in theme is ‘typical’. In parallel, much research on demonstratives in academic writing has analysed only sentence-initial tokens (Charles, 2003; Swales, 2005; Gray, 2010, Wulff et al. 2012): Swales admits that other positions are possible but argues that ‘subject position remains, as it were, the locus classicus of this device’ (2005: 2).

Boettger and Wulff (2014) is one of the few studies of *this* which considers tokens in all sentence positions. In a corpus of student essays the study found 52% and 34% to be in sentence-initial and medial positions respectively. Unfortunately - from the perspective of this study - the research did not address the distinction between complex and simple anaphora.

The reason researchers have assumed that complex anaphors will appear sentence-initially presumably derives from theories of information structure. Insofar as the entities to which CA refer must have been introduced earlier in the text, albeit not in nominal form, it would seem that anaphors cannot be entirely new information nor therefore appear as the focus of any clause. According to Halliday’s conceptualization of theme and rhyme, for example, in the unmarked case given information is presented in theme rather than rhyme and one of the motivations for grammatical metaphor is so that complex passages may be “‘packaged’ in nominal form as Theme’ (1994: 353). As noted earlier, some research on demonstratives has confined itself to sentence-initial instances of *this*, and this is possibly because of an assumption that such themes are the natural place to look for anaphora.

While intuition suggests that sentence initial anaphors are more likely to be performing CA than non-sentence-initial anaphors, it seems desirable to have research which confirms this intuition rather than assuming it. To do this, research needs to identify the referents of all complex anaphors, not only of those appearing sentence-initially. Accordingly the present study will analyse the referents of all instances of *this* in the data.

2. Method

2.1 Objectives

The research described here aimed to fill some of the gaps identified above. Firstly, it will identify and compare the frequency of all *this* tokens performing CAWT in three corpora of written argumentative texts, representing three differing genres and their writers. Among CAWT tokens it will further measure variation in use of pronoun *this* and determiner *this*. All CAWT determiner tokens will be examined, not only those occurring in phrases headed by items from closed sets (*labels*, *shell nouns*). The study will compare lexical variety in the heads of CAWT in which *this* is a determiner. Finally, the study will identify how much variation exists in the comparative placements of CAWT sentence-initially and non-initially.

To summarise, this research sets out to measure and characterise CAWT behaviour in terms of the following four parameters:

- frequencies of *this* as determiner and pronoun
- frequencies of CAWT pronoun (CAWTPRO) and CAWT determiner (CAWTDET)

---

4 These percentages are taken from the raw frequencies supplied by Boettger and Wulff (2014: 124).

- range of lexical heads used with CAWTDET
- frequencies of placement of CAWT within the host-sentence, initial vs. non-initial.

2.2 Data

The three different genres of argumentative text investigated are academic RAs, student essays, and newspaper editorials. These particular genres were selected because, as we have seen, previous research on *this*, labels, and shell nouns, has drawn on all three but no research has conducted a direct comparison of all three.

A comparison of CAWT in academic RAs and student essays seems appropriate in order to investigate whether or not, as is suggested in some of the EAP literature reviewed above (Petch-Tyson 2000; Flowerdew, 2006; Gray & Cortes, 2011), particular kinds of behaviour, such as higher frequencies of determiner *this*, are associated with greater proficiency in writing and/or in argumentation.

Comparing CAWT in academic RAs and editorials is similarly designed to show whether particular CAWT behaviours are a reflex of proficiency in writing/argumentation or features associated with particular genres. Editorials (leading articles from UK newspapers) have been selected as examples of high-proficiency but non-academic written argumentation. A comparison between CAWT in newspaper editorials and academic texts is considered to be of particular relevance to academic writing pedagogy: editorials and opinion pieces are often selected for inclusion in the textbooks used for academic writing instruction in generalist composition classes (e.g., Oshima & Hogue, 2006; Behrens & Rosen, 2010), presumably because they are formal in style but not discipline-specific in content, and therefore more accessible to students different degrees. In many respects this choice seems appropriate, but in the particular realm of CAWT this study aims to discover how closely the style of editorials mirrors that of advanced academic argumentation.

Details of the three corpora used are given in Table 1. Because of the need to analyse co-text, all texts in each corpus are whole texts.

<table>
<thead>
<tr>
<th>Short name</th>
<th>Variety</th>
<th>Genre</th>
<th>Authors</th>
<th>Words</th>
<th>Texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACAD</td>
<td>Expert academic argumentation</td>
<td>Research article (Economics)</td>
<td>Academics</td>
<td>100,584</td>
<td>10</td>
</tr>
<tr>
<td>EDIT</td>
<td>Expert non-academic argumentation</td>
<td>Editorial (UK broadsheet)</td>
<td>Journalists</td>
<td>99,870</td>
<td>182</td>
</tr>
<tr>
<td>STU</td>
<td>Non-expert academic argumentation</td>
<td>Academic essay</td>
<td>NNES undergraduate students</td>
<td>100,281</td>
<td>198</td>
</tr>
</tbody>
</table>

**Table 1: The corpora**

The ACAD corpus is a subcorpus from the Published Research Article Corpus (Cortes, 2007) and contains articles from *Applied Financial Economics, Advances in Theoretical Economics*, and *The Journal of Financial Economics*. Economics was chosen because firstly previous research had not been carried out on texts in this discipline. Secondly, if one were to characterize academic disciplines in terms of a hypothetical continuum of ‘hard’ to ‘soft’, such as that described in Hyland (2005), economics, containing numbers and formulas but also extended verbal argumentation, is assumed to be near the middle.

The EDIT corpus was compiled by the author from UK broadsheet newspaper (*Guardian, Independent, Telegraph, Times*) editorials appearing in summer 2010.

The STU corpus comprises short essays written by undergraduate ESL writers at the American University of Sharjah, (AUS), an American curriculum, English-medium university in the United Arab Emirates. The majority (87%) of the authors were L1 Arabic. The English proficiency level of the students varied considerably but all had to score a minimum of 500 in the TOEFL exam to enter the university. The essays were word-processed and submitted as regular class assignments for composition classes; they were often completed in the students’ own time and not produced under test conditions. Most of the essays (70%) collected (from 2007 to the present) were written in response to one of the prompts listed in the Appendix, the rest being written as part of other argumentative assignments. This corpus is still under construction and the data is designed to be comparable with that contained in the International Corpus of Learner English (ICLE), a corpus of student essays collected in European universities (Granger, Dagneaux, and Meunier, 2002) and the Louvain Corpus of Native English Speaking Student text (LOCNESS), a corpus of about 200,000 words (Granger and Tyson, 1996).

### 2.3 Procedure

All tokens of this in the data were identified using concordancing software *AntConc* (Anthony, 2012). The context of each token was then analysed to identify whether it was anaphoric or cataphoric, and if the former whether the referent was complex. Consten *et al.* explain that by ‘complex’ they mean: ‘the antecedent’ has to be a complex linguistic entity, which means that it consists of (at least) a clause’ and ‘the referent has to be a conceptually...
Crompton, P. (forthcoming). Complex anaphora with this: variation between three written argumentative
genres. Corpora 12(2).

complex item’ which means, in terms of Lyons’s (1977) classification not a ‘first order’
extent (persons or things) but a ‘second order’ (events or processes) or ‘third order’ (facts or
propositions) entity (Consten et al., 2007: 83).

Although deciding whether a potential token is in fact a complex anaphor may be time-
consuming, requiring careful reading of prior text, the process itself is relatively
unproblematic: in addition to the above positive criteria, the analyst must ascertain that the
candidate CA token is not making reference to a discourse referent which meets either of the
following conditions:

(a) the antecedent follows the token – analysed as cataphora

(b) the referent has already been referred to in a nominal form – analysed as simple
anaphora.

Here is an example of (b):

(10) Over and above the question of Mr Hayward’s individual gaffes, his responsibility
for the overall direction of BP’s public relations strategy with regard to the oil spill
needs to be taken into consideration. From a distance in time of more than six
weeks, this strategy must be pronounced a total failure. (EDIT)

A potential problem in the analysis of CA worth noting is that of determining the extent
of the referent; as Francis also points out in her discussion of the ‘fuzzy reference’ of labels
‘it is not always possible to decide where the initial boundary of its referent lies’ and
indeterminacy may sometimes be deliberate (1994: 88). For this study, the key analytical
determination was whether or not a referent was complex, (i.e. whether or not it comprised at
least one clause/proposition). The issue of the extent of the referent (i.e. whether it included
more than one clause/proposition) was not germane.

It is worth noting that some research (Petch-Tyson, 2000; Botley, 2006) has attempted
to go further and subcategorise particular instances of CA using more delicate typologies.
These may be based on semantic properties of the referent, such as Lyons’ (1977) above-
noted distinction between second- and third-order entities. Alternatively, Francis’s
classification of labels appears to be based on the semantics of the anaphor, for example as
general (e.g., action) or metalinguistic and within the latter sub-classification as
‘illocutionary’ (e.g., answer), ‘language-activity’ (e.g., diagnosis), or mental process (e.g.,
analysis). While such classifications are attractive, a problem with them is that they are
abstract and not watertight: Francis mentions ‘blurring and overlap’ in her categories (1994: 90) and Botley (2006) was unable to categorise 29% of the IA tokens found in his data. In
this research, therefore, semantic subcategorisation of each act of anaphora has been
eschewed.

5 There were only 13 instances of cataphoric this found in the data: for convenience these have been counted in
with SAWT tokens.
6 Example (9) illustrates a complex referent which extends over several clauses.
3. Results and Discussion

3.1 Frequencies of pronoun and determiner

Table 2 shows the frequencies of *this* overall, as pronoun, as determiner, performing simple and complex anaphora, by corpus, firstly as raw frequencies, then as normalized frequencies per 1,000 words. Before more detailed discussion, let us note that the normalised frequencies of *this* in this data (from 6.79 to 5.94) are broadly comparable with those for *this* in the British National Corpus Academic and Spoken Corpora (6.28 and 5.81 per thousand words respectively (Leech et al., 2001)).

<table>
<thead>
<tr>
<th>Form</th>
<th>Raw frequencies</th>
<th>Normalized frequencies (per 1,000 words)</th>
<th>LL for pairwise frequency comparisons (LL&gt;3.84 in bold)</th>
<th>Significance p value of LL statistic (p&lt;0.05 in bold)</th>
<th>Effect Size Log of Odds Ratio (only given where p value of LL &lt;0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACAD</td>
<td>EDIT</td>
<td>STU</td>
<td>ACAD</td>
<td>EDIT</td>
</tr>
<tr>
<td>this all</td>
<td>683</td>
<td>617</td>
<td>596</td>
<td>6.79</td>
<td>6.18</td>
</tr>
<tr>
<td>SA</td>
<td>289</td>
<td>301</td>
<td>264</td>
<td>2.87</td>
<td>3.01</td>
</tr>
<tr>
<td>CA</td>
<td>394</td>
<td>316</td>
<td>332</td>
<td>3.92</td>
<td>3.16</td>
</tr>
<tr>
<td>this pron.</td>
<td>184</td>
<td>243</td>
<td>207</td>
<td>1.83</td>
<td>2.43</td>
</tr>
<tr>
<td>SA</td>
<td>13</td>
<td>31</td>
<td>22</td>
<td>0.13</td>
<td>0.31</td>
</tr>
<tr>
<td>CA</td>
<td>171</td>
<td>212</td>
<td>185</td>
<td>1.70</td>
<td>2.12</td>
</tr>
<tr>
<td>this det.</td>
<td>499</td>
<td>374</td>
<td>389</td>
<td>4.96</td>
<td>3.74</td>
</tr>
<tr>
<td>SA</td>
<td>276</td>
<td>270</td>
<td>242</td>
<td>2.74</td>
<td>2.70</td>
</tr>
<tr>
<td>CA</td>
<td>223</td>
<td>104</td>
<td>147</td>
<td>2.22</td>
<td>1.04</td>
</tr>
</tbody>
</table>

**Table 2:** Frequencies of SAWT and CAWT, with significance and effect size of intercorpus variation

Table 2 also shows the statistical significance and effect size of differences between the raw frequencies based on pairwise comparisons of the different corpora. The statistic chosen to test for significance was log-likelihood, an alternative to Pearson’s $X^2$ (Rayson & Garside, 2000; Cressie & Read, 1989). Similarly to $X^2$, log-likelihood is calculated using the raw frequencies of items appearing in two or more corpora and the total wordcount for those corpora and makes no assumption of normal distribution in the data. It was chosen because it has been argued to be more reliable for data from large corpora (Dunning, 1993; Rayson & Garside, 2000). As with $X^2$, with one degree of freedom a log-likelihood (LL) value of 3.841 indicates a confidence level of 95% ($p<0.05$). In this paper each LL value from pairwise comparisons of item frequencies between the corpora is given in Tables 2 and 5, along with the corresponding $p$ value. In Figures 1-7, where any pairwise comparison between the frequencies of the same item in two different corpora has produced a significant LL value (i.e., $p<0.05$) that fact is noted in the figure legend. For such cases, Tables 2 and 5 also shows a measure of effect size: the statistic chosen for this was the natural log of the odds ratio, which measures the strength of association between two binary variables. In this case the two variables are (a) whether a word is in corpus A or corpus B and (b) whether or not a word is an instance of the given use of *this*. As Gries explains, the purpose of logging the odds ratio is to make it easier to compare effects: “the more a logged odds ratio deviates from 0 the stronger the effect” (2014, p.372).

Where particular frequency differences are significant this will be commented on the in following section. Overall, it is interesting to note here firstly that while significant variations in frequency are common in ACAD-EDIT and ACAD-STU comparisons, EDIT-STU variation is significant in only one case. Secondly, for those items which included the parameter of reference (simple or complex) all but one of the significant variations occur with frequencies relating to complex rather than simple anaphors.

From this point on, for clarity detailed finding concerning the four parameters identified in the Method section will be addressed as answers to specific research questions.

*RQ 1.*  *Is there inter-corpus variation in the frequencies of this as pronoun and determiner?*

To show inter-corpus variation more clearly, Figure 1 depicts the frequency data for pronoun and determiner frequency visually. Given the prescriptive tradition noted by Swales (2005) and Gray and Cortes (2011), which warns against ‘unattended *this*’, one might expect that the more formal in register a text is, the greater likelihood there would be of prescriptions being adhered to. Assuming RAs to be a more formal genre than editorials, one would therefore hypothesise that academic writers (ACAD) would use fewer *this* pronouns than editorial writers (EDIT). This hypothesis is confirmed: the ACAD corpus contains significantly fewer *this* pronouns than EDIT and the actual proportion of ACAD *this* tokens which are pronouns, 27% (184/617), which is in line with previous corpus research on academic writing (Gray & Cortes, 2011; Swales, 2015; Wulff *et al.*, 2012), proves to be smaller than for the other corpora, for EDIT 37% (229/617) and STU 35% (206/596).

One might also have hypothesised that, on the grounds that non-native English speaking students are likely to have a smaller lexical repertory, learner academic writers (STU) would resort to *this* pronouns more than either group of expert writers. In practice, this is not the case: roughly midway between ACAD and EDIT this, STU writers use of *this* pronouns is not significantly different from either of the other corpora.

Figure 1 also illustrates a higher frequency of determiner *this* in the ACAD corpus than the other corpora, a difference which is significant. As will be discussed further below, this distinctiveness seems particularly associated with CA.

![Significant intercorpus variation: pronoun: ACAD-EDIT; determiner: ACAD-EDIT, ACAD-STU](image)

**Figure 1**: Frequencies of *this* as pronoun and determiner

**Frequencies of simple and complex anaphora**

*RQ 2. What proportion of this-tokens perform CA?*

The answer to this question for each corpus appears to be ‘a little over half’ (illustrated in Figure 2). Overall, it is interesting to discover that in all three corpora, the use of *this* to perform CA is not exotic: *this* is more likely to be performing CA than SA.

Significant intercorpus variation:
complex: ACAD-EDIT, ACAD-STU

**Figure 2**: Frequencies of *this* performing simple and complex anaphora

**RQ 3.** *Do academic writers and editorialists use similar frequencies of CAWT?*

As we have seen, previous research by Botley (2006) suggested that CAWT might be distinctive of argumentative texts and previous research by Francis (1994) and Schmid (2000) argued that subsets of CAWTDET, namely labels and shell nouns, are distinctive of argumentative texts. Whether professional academic writers use more or less CAWT than editorialists has not previously been researched. Figure 2 shows that they do: the variation in frequency of CAWT between the ACAD and both the other corpora is indeed significant while that for SAWT is not. Breaking down frequencies of CAWT tokens by word class, however, we find a distinction between pronoun and determiner.

**RQ 4.** *How much variation is there in the frequencies of CAWTPRO?*

As Figure 3 shows, for all three writer groups, pronoun *this* is very much likelier to be performing CA than SA. While this pattern is common to all three corpora, the EDIT writers seem to make slightly more use of CAWTPRO, significantly more than the ACAD writers.

Significant intercorpus variation:
- simple: ACAD-EDIT
- complex: ACAD-EDIT

**Figure 3:** Frequency of *this* as pronoun performing simple and complex anaphora

**RQ 5. How much variation is there in the frequencies of CAWTDDET?**

Figure 4 depicts the variation between corpora in frequency of tokens of determiner *this* performing CA. While the minor frequency variation for simple anaphora is not significant, the more marked frequency variation in complex anaphora for all three pairwise comparisons is significant, the largest effect difference being that between ACAD and EDIT writers. Student writers appear to be at an approximate half way point between the other writers.

**Figure 4:** Frequency of *this* as determiner performing simple and complex anaphora

![Graph showing frequency of 'this' as determiner](image)

Significant intercorpus variation:
complex: ACAD-EDIT, ACAD-STU, EDIT-STU

**Figure 5:** Frequency of *this* as pronoun and determiner performing complex anaphora

![Graph showing frequency of 'this' as pronoun and determiner](image)

Significant intercorpus variation:
CAWTPRO: ACAD-EDIT
CAWTDET: ACAD-EDIT, ACAD-STU, EDIT-STU

**Figure 5** brings together the data for complex anaphora as performed by pronouns and determiners and illustrates further the intercorpus variation. Given previous research on NNES learner writer problems with labels and shell and signalling nouns (Petch-Tyson, 2000; Flowerdew, 2006) it is interesting to note that the NNES student writers in this study are not obviously underusing determiner *this* or overusing pronoun *this* by comparison with the NES authors of the EDIT corpus. To summarise, expert academic writers make greater use of lexical NPs with *this* as determiner than the other writer groups. This greater use

comes not, as we might have hypothesised, instead of pronoun *this* but in addition to it. From a pedagogic perspective, these findings might suggest that instructors should give some attention to making students aware of higher frequencies of CAWTDET as a distinctive feature of academic writing.

**Range of lexical heads used with CAWTDET**

*RQ 6. How much variation is there in the range of lexical heads of CAWTDET?*

Table 3 shows the frequencies of tokens and types of lexical heads of CAWTDET tokens and the derived type token ratios (TTR), and Figure 6 illustrates the variation. The ACAD corpus clearly has the greatest variety in terms of types, but also the smallest type-token ratio.

<table>
<thead>
<tr>
<th></th>
<th>ACAD</th>
<th>EDIT</th>
<th>STU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>223</td>
<td>104</td>
<td>147</td>
</tr>
<tr>
<td>Types</td>
<td>110</td>
<td>80</td>
<td>92</td>
</tr>
<tr>
<td>Type Token Ratio (%)</td>
<td>49%</td>
<td>77%</td>
<td>63%</td>
</tr>
</tbody>
</table>

**Table 3:** Types, tokens and type token ratio of heads in CAWTDET by corpus

![Figure 6: Types and tokens of heads in CAWTDET](image_url)

In the matter of lexical variation, the dissimilarities between the corpora in terms of disparities in field, or subject-matter, make it difficult to draw conclusions from these data. However, some features appear worthy of comment. The larger TTR and the greater proportion of types appearing less than three times in the EDIT corpus (shown in the final line of Table 4) could reflect the broader range of semantic fields covered in commentary on daily news and also greater stylistic variation in a genre which is more literary. From a pedagogic perspective, it is noteworthy that the type-token ratio for the STU corpus is not lower than both the NES corpora, but at a midway point between them. It is also surprising, in view of the much narrower semantic field of the ACAD texts, that none of the commonest
ACAD items appears to be content-specific, that is, relates to the fields of economics or finance. Instead, many – e.g., assumption, analysis, reason, strategy, hypothesis - reflect an epistemic concern with method and reasoning, while the two of the EDIT items, crisis and atmosphere, reflect political analysis. It would seem that composition instructors could help students to argue academically by drawing their attention, in academic texts read in class, to the epistemic emphasis of lexical heads accompanying CAWDET.

<table>
<thead>
<tr>
<th>Item</th>
<th>Freq.</th>
<th>Item</th>
<th>Freq.</th>
<th>Item</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>case</td>
<td>27</td>
<td>crisis</td>
<td>12</td>
<td>case</td>
<td>11</td>
</tr>
<tr>
<td>result</td>
<td>21</td>
<td>approach</td>
<td>4</td>
<td>way</td>
<td>10</td>
</tr>
<tr>
<td>assumption</td>
<td>8</td>
<td>case</td>
<td>4</td>
<td>issue</td>
<td>5</td>
</tr>
<tr>
<td>approach</td>
<td>6</td>
<td>atmosphere</td>
<td>3</td>
<td>situation</td>
<td>5</td>
</tr>
<tr>
<td>issue</td>
<td>6</td>
<td></td>
<td></td>
<td>change</td>
<td>4</td>
</tr>
<tr>
<td>sense</td>
<td>5</td>
<td></td>
<td></td>
<td>idea</td>
<td>4</td>
</tr>
<tr>
<td>analysis</td>
<td>4</td>
<td></td>
<td></td>
<td>matter</td>
<td>4</td>
</tr>
<tr>
<td>pattern</td>
<td>4</td>
<td></td>
<td></td>
<td>problem</td>
<td>4</td>
</tr>
<tr>
<td>reason</td>
<td>4</td>
<td></td>
<td></td>
<td>action</td>
<td>3</td>
</tr>
<tr>
<td>strategy</td>
<td>4</td>
<td></td>
<td></td>
<td>evil</td>
<td>3</td>
</tr>
<tr>
<td>fact</td>
<td>3</td>
<td></td>
<td></td>
<td>fact</td>
<td>3</td>
</tr>
<tr>
<td>hypothesis</td>
<td>3</td>
<td></td>
<td></td>
<td>type</td>
<td>3</td>
</tr>
<tr>
<td>information</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>possibility</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>respect</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>view</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>107</strong></td>
<td><strong>23</strong></td>
<td><strong>59</strong></td>
<td><strong>48</strong></td>
<td><strong>22</strong></td>
</tr>
<tr>
<td>% of tokens appearing ≥ 3</td>
<td><strong>48</strong></td>
<td><strong>22</strong></td>
<td><strong>40</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of tokens appearing &lt; 3</td>
<td><strong>52</strong></td>
<td><strong>78</strong></td>
<td><strong>60</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4:** Most frequent lexical heads (F=3 or above) appearing with CAWTDET (items common to two lists in bold)

**Frequency of placement (initial vs. non-initial) of CAWT within the host-sentence**

**RQ 7.** *Is there a tendency for CAWT to be placed initially?*

As we have seen, there has been a general tendency in the literature to assume that initial or thematic placement is the unmarked case for complexly anaphoric noun phrases (Francis, 1994; Charles, 2003; Yamasaki, 2008). Corpus-based research into comparative frequencies of determiner and pronoun (Swales 2005, Wulf et al. 2012) has also restricted its sample to tokens which are sentence-initial in the sense of *this* being the first word in the sentence. For this research, to investigate how much CA performed with *this* does actually appear sentence-initially and how much non-initially, the position of each CA token within its syntactic context was analysed. The placement of each token within its host independent clause was considered. Following the practice adopted in much systemic-functional research (reviewed in Crompton, 2002) the initial position, or theme, was interpreted as extending from the start of the clause to the main verb. Here are some examples:

**Initial CAWT**

(11) The more important side in *this asymmetric conflict*, however, is the Israeli. (EDIT)
(12) Is *this* what Mr Clegg means when he talks about the ‘new politics’? (EDIT)
(13) SEO firm shares under perform the market by nearly 21% on average and *this*, too, differs significantly from zero at the 1% level. (ACAD)
(14) *This* makes it hard for a person to live with him/herself. (STU)

**Non-initial CAWT**

(15) But there is obviously more to *this deal* than commercial logic alone. (EDIT)
(16) See Tirole (1987: 177) for a justification of *this assumption*. (ACAD)
(17) There is also a time component to *this*. (ACAD)
(18) But sadly, some of the women understand *this freedom* in a very different way. (STU)

If one were to assume that the principles of information structure, including the placement of anaphors within clauses, were common to proficient English users, one would hypothesise that there would be the least difference between the frequencies for the two expert writer corpora. Correspondingly, one would predict that inexpert student writers, writing in a second or foreign language with possibly different means of indicating information structure than those of English, might have control problems leading to smaller frequencies of sentence-initial placement.

Table 5 shows the actual frequencies of all CAWT initially and non-initially. As the literature would have led us to expect, all three writer groups are indeed much likelier to place CAWT tokens initially. The ACAD corpus writers’ tendency to do so is stronger than that of the other writer groups and the pairwise comparisons with other corpora show that this variation is significant.

<table>
<thead>
<tr>
<th>Form</th>
<th>Raw frequencies</th>
<th>Normalized frequencies (per 1,000 words)</th>
<th>LL for pairwise frequency comparisons (LL&gt;3.84 in bold)</th>
<th>Significance p value of LL statistic (p&lt;0.05 in bold)</th>
<th>Effect Size Log of Odds Ratio (only given where p value of LL &lt;0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACAD</td>
<td>EDIT</td>
<td>STU</td>
<td>ACAD</td>
<td>EDIT</td>
</tr>
<tr>
<td>initial</td>
<td>315</td>
<td>216</td>
<td>238</td>
<td>3.13</td>
<td>2.16</td>
</tr>
<tr>
<td>non-initial</td>
<td>79</td>
<td>100</td>
<td>94</td>
<td>0.79</td>
<td>1.00</td>
</tr>
<tr>
<td>initial PRO</td>
<td>157</td>
<td>171</td>
<td>150</td>
<td>1.56</td>
<td>1.71</td>
</tr>
<tr>
<td>non-initial PRO</td>
<td>14</td>
<td>41</td>
<td>35</td>
<td>0.14</td>
<td>0.41</td>
</tr>
<tr>
<td>initial DET</td>
<td>158</td>
<td>45</td>
<td>88</td>
<td>1.57</td>
<td>0.45</td>
</tr>
<tr>
<td>non-initial DET</td>
<td>65</td>
<td>59</td>
<td>59</td>
<td>0.65</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 5: Frequency of CAWT initially and non-initially with significance and effect size of inter-corpus variation
Figure 7 illustrates the variation in placement of CAWT forms between corpora. If we consider first CAWTPRO, in the left half of the chart, we observe a strong preference for initial position in all three corpora. However, we can also observe from the smaller frequency of non-initial CAWTPRO that the ACAD corpus’s preference for initial position is significantly stronger than that shown in the other corpora.

Significant intercorpus variation:
PRO non-initial: ACAD-EDIT, ACAD-STU
DET initial: ACAD-EDIT, ACAD-STU, EDIT-STU

**Figure 7: CAWTPRO and CAWTDET initially and non-initially**

In contrast, for CAWTDET the right half of the chart shows (a) much greater, and significant, variation in frequencies of initial position, and (b) higher frequencies of non-initial position for CAWTDET (frequencies almost uniform across the corpora). These properties together mean smaller preferences for initial position than is the case with CAWTPRO in the ACAD and STU corpora and an actual dispreference in the EDIT corpus.

Overall, then, the ACAD writers show a significantly stronger preference than the EDIT writers for initial position for CAWTPRO and than both the other writer groups for initial position for CAWTDET.

More generally, comparing right and left halves of Figure 7, one of the most striking aspects of the data is the much greater proportions of non-initial CAWTDET than of non-initial CAWTPRO: despite the broad inter-corpus variation, all the writer groups seem to feel considerably greater freedom to place CAWTDET non-initially than is the case with CAWTPRO. To my knowledge there is no previous research on this topic, so at this stage inferring possible causes is necessarily somewhat speculative. One plausible explanation would seem to be based on processing: let us suppose that a recency effect means that the referent of an initial *this* expression is more accessible. Conversely, the appearance of other referents in the initial segment of new sentence makes the referent of a non-initial *this*

expression less accessible. In this case, the writer’s further specification of the referent by means of a lexical head would serve to facilitate processing. This might explain why the writer of (19), for example, feels no need to specify the referent of this, while by contrast the writer of (20) feels the need to specify this as an approach:

(19) To simplify the exposition we limit our analysis to two-person, finite, multistage games with observable actions. This also allows us to use a notation.... (ACAD)

(20) One way to deal with this difficulty is to conduct market research and use the results in the design of the trading mechanism. There are a number of problems with this approach. (ACAD)

However, processing constraints cannot explain the original decision to place a complex anaphor non-initially, or why, as Figure 7 shows, EDIT writers uniquely disprefer initial position for complex anaphors with determiner this; this issue clearly requires further research. Authentic texts are of course extremely complex phenomena, with many factors potentially affecting writers’ all linguistic choices. Corpus-based studies such as this, by indicating larger patterns of behaviour, can only show what choices in practice are often made – information which can then be used to generate working hypotheses as to function and motivation for further research.

4. Conclusion

Let us summarise what the present study has shown about the use of this in argumentative text. Starting with comparative overall use of pronoun and determiner forms, in line with previous research on different data, a quarter to a third of instances of this in academic writing are pronouns. In line with the results of Gray’s (2010) research on initial demonstratives in RAs, expert professional writers, both academic and non-academic, are flouting academic style-guide writers’ proscription of using ‘unattended’ this.

Turning to comparative overall use of this for CA and SA, in all three corpora, this is used to perform CA more often than SA. However, there is a striking difference between pronoun and determiner use: in the great majority of instances this pronouns in all sentence positions are used for CA rather than SA. Combining these findings with those of Botley (2006), it would seem fair to say that the use of pronoun this primarily as a complex anaphor should be recognised as a distinctive feature of argumentation, both non-academic and academic.

By contrast, in all three corpora determiner this is used for SA more than for CA. However, unlike pronoun this, register differences between academic and non-academic argumentation are apparent. As we have seen, expert academic prose has significantly more determiner this tokens than the other corpora, but this greater frequency is in addition to rather than at the expense of pronoun this. Moreover, in expert academic prose the proportion of determiner this tokens with a complex referent is significantly higher. Overall then, expert academic writers seem more likely to perform CA using determiner than pronoun this. One

wonders whether an intuitive perception of this higher proportion lies behind the academic style-guide writers’ promotion of determiners – at the expense of pronouns – as the preferred form for this.

Regarding the particular lexis expert academics use to accompany determiner this – *labels* in Francis’s (1994) terms – while some of the lexical heads commonly used in academic argument, such as *case*, are also common in other argumentative genres, those with a more epistemic orientation may be distinctive to academic argument.

Turning to placement of complex anaphors, that is sentence-initially or elsewhere, we find the greatest variation between the practices of the different writer groups. The academic writers, both experts and students, are more likely than the editorialists to place complex anaphors, particularly lexically headed complex anaphors, sentence-initially. Editorialists are distinctive in marginally dispreferring sentence-initial position for lexically-headed complex anaphors. In view of the assumption commonly made in the literature, the fact that large proportions of complex anaphors in all three corpora do not occur sentence-initially is surprising. This finding lends support to the contention of Consten *et al.* ‘that the strict dichotomy of thematic versus rhematic entities in texts has to be abandoned. Complex anaphors…are means of textual continuity and progression at the same time’ (2007: 99).

Where complex anaphors appear sentence-initially, as is largely the case with pronoun this, illustrated in example (12) there is a case to be made they are being used in the service primarily of text continuity; in Hallidayan terms (1994), they appear to fulfil the prototypical thematic function of being the point of departure for the main clause. However, where complex anaphors appear non-initially, such as determiner this in example (13), it would seem that they are, as Consten *et al.* argue, a means of both continuity and progression. As well, then, as the choice of determiner versus pronoun for the form of complex anaphor, the writer’s choice of both position and function (e.g., subject, complement, object, adverbial) for the placement of complex anaphors looks like an issue worthy of further research.

In terms of variation between the corpora, the greatest variation for all the above parameters is not between either of the expert writer corpora and the student corpus but between the expert academic corpus and the editorial corpus; thus, in nearly all the figures in this paper, the column representing frequencies for the student corpus is at a midway point between those representing the expert academic writers and the editorialists. As the students are not expert writers and indeed in most cases writing in a second or foreign language without benefit of editing, this pattern is surprising. As suggested earlier, it may indicate that the variation is accounted for not by differences in linguistic proficiency, but genre differences, such as content (or field) and communicative purpose. The midway position of the student corpus frequencies possibly reflects the fact that student essays are something of a hybrid genre: their writers are essentially prompted to argue in an academic style about content, which like that of editorials, is not academic.

In terms of the specific pedagogic concern of selecting model argumentative texts which are appropriate for academic writing instruction classes, on the basis of frequency evidence alone, editorials seem to offer a broadly appropriate model of this to perform CA.

For both RAs and editorials, for example, the frequency of pronoun *this* for CA and SA and its positioning in sentences is similar. However, as we have seen, the difference between editorials and academic research writing becomes clearer in the use of determiner *this* for CA and in its positioning. Over time it would seem a good idea to expose students to extracts from more specifically academic texts, drawing attention to the high frequency of lexically-headed CA, the epistemic orientation of heads in lexically-headed CA, and the tendency to thematic placement of complex anaphors, especially lexically-headed CA. In particular, students who may be overwedded to the use of encapsulating logical connectors like *therefore* and *however* in their own writing could have their awareness raised of the range of nuanced alternatives offered by *this* as determiner.

To revisit research on student writers’ use of CA and demonstratives, Petch-Tyson’s (2000) findings from data collected in North America and Europe suggested that L1 could have a significant impact on how much and how students employ the different English demonstratives to accomplish CA. The findings for the Middle Eastern students represented in the corpus used here do not support that particular interpretation; however, further research on CA in the texts of ESL learners from other backgrounds would shed further light. More generally, our understanding of what is distinctive about CA in written Anglophone argumentation—and therefore possibly difficult for novice academic writers with different L1s—might be helped if we had more data on how CA is performed in the mother-tongues of different student populations. Another useful area of research would be on the performance of CA in other genres in English with which students may be more familiar, including formal and informal spoken genres.

This research has confined itself to CA as expounded by *this*. In order to further understand the role of CA in structuring text, it would be useful to identify and compare conditions and restrictions on the use of other forms, including the other demonstratives, used to perform CA.

References


Davies, M. 2004-. BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc/.


Gray, B. 2010. ‘On the use of demonstrative pronouns and determiners as cohesive devices: A focus on sentence-initial this/these in academic prose’, *Journal of English for Academic Purposes* 9, pp.167-183.

Gray, B. and V. Cortes. 2011. ‘Perception vs. evidence: An analysis of “this” and “these” in academic prose’, *English for Specific Purposes* 30, pp. 31-43.


**Appendix. STU essay prompt**

Write an essay of 500-1000 words on one of the following:

1. “Crime does not pay.” Discuss.

2. "Money is the root of all evil." Discuss.

3. “Feminists have done more harm than good to the cause of women.” Discuss.

4. “The prison system is outdated. No civilised society should punish its criminals: it should rehabilitate them.” Discuss

5. “Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.” Discuss.

6. “A person's financial reward should be commensurate with their contribution to the society they live in.” Discuss.

7. “Marx once said that religion was the opium of the masses. If he was alive at the start of the 21st century, he would have said ‘television’ instead of ‘religion’.” Discuss.

8. “In our modern world, dominated by science, technology, and industrialisation, there is no longer a place for dreaming or imagination.” Discuss.