SAME TUNE, DIFFERENT SONGS: BANALITY, CRITICAL INVENTIONS, AND COLLOCATIONS IN *LORD OF THE FLIES* CRITICISMS

by

DUSTIN ELIAS GRUE

B.A., University of Lethbridge, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(English Language)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2011

Abstract

What place does formulaic language have in literary criticism? On the one hand, as Douglas Biber (2006) suggests, repeating word combinations "are important for the production and comprehension of texts in the university" (p. 135). But on the other hand the repetition of stock or conventional phrases opens up academic writing to the charge of repetition, an act proscribed as 'banal.' So formulaic language is both good and bad: necessary but also necessary to avoid. And the study of literature might be especially subject to such folly, since the very epistemology of criticism *is* repetition – critics' reproductions (through quotation) of an author's writing, critics' echoes of one another, secondary texts, etc. By necessity, a chorus of voices critiquing the same texts motivates the creation of conventional language – but what is this necessity? When is it generative, and when is it banal?

Under the theoretical guidance of Relevance Theory, and using methods from corpus linguistics and the Digital Humanities, this work investigates formulaic language in a corpus comprised of literary criticism. Such criticism is 46 works on William Golding's (1954) novel, *Lord of the Flies*. I also sketch the history of the theoretical concept of 'collocation'—generally, the tendency for words to cluster around other words—and argue against the model of collocation that favours semantic conventionalization, where collocations are, essentially, coded with meaning. A main finding of this work is that collocations are often attributed to other speakers—real or fictional—and therefore their meaning is more pragmatically oriented than semantically conditioned.

Data analysis is performed through automated rendering of the corpus using custom scripts, and qualitative analysis – of both the output of such rendering, and distanced reading of the corpus. The centerpiece tool of this work is a text-tool I developed that produces a visualization of terms' collocates. This visualization is based on Howard White's (2007a, 2007b) work in bibliometrics, and graphs collocations on two axes corresponding to the central tenets of Relevance Theory. Other quantitative methods of investigation describe the discovery of a term saliency metric (Chapter 2) and extended distributions of terms around other terms (Chapter 4).

Table of Contents

List of Tables

List of Figures

[1] Reprinted from White (2007a) with permission. Please see note on page 66.

Acknowledgements

Many thanks to my supervisor and thesis committee, Dr. Janet Giltrow and Dr. Laurel Brinton, for all of their work seeing this project through to the end – their contributions have been immeasurable.

The Social Sciences and Humanities Research Council of Canada provided funding for this project.

Dedication

To Dr. Allison Balcetis.

1 Introduction

1.1 Banality, Repetition, Criticism

Reviewing a collection of criticisms on William Golding's novels,[2] Patrick Swinden (1987) ironically asks, "What is it about Golding that makes critical conversation about him so banal?" (p. 570). Obviously, Swinden did not receive the collection very well, expressing his complaint by calling the criticism 'common' and 'regular' – banal: the same comment iterated over and over. And as enticing as Swinden's comment is—Golding criticism, at first glance, *does* appear to be unceasingly replicated[3]—it bears some reflection: situated within the genre of 'the critical review' Swinden comments on the genre of literary criticism, implicitly indicating a quality of 'good' criticism as opposed to 'bad.' Simply put, good criticism must be 'different,' while bad criticism is 'the same.'[4] But ought criticism *be* 'different' in order to be productive? Does the same comment, iterated, not inform? And is *Lord of the Flies* criticism really 'the same'?

This reflection on genre about genre, or "situated language about situated language" (p. 190), Janet Giltrow (2002) terms "meta-genre": a term for analysis that considers not just the situatedness of the production of an utterance but the wider, frequently discrepant motivating conditions of the multiple 'situations' and 'contexts' in which the utterance is produced. As Giltrow finds, these conditions are often expressed

[2] *William Golding: Novels, 1954-67: Lord of the Flies; The Inheritors; Pincher Martin; Free Fall; The Spire; The Pyramid.* Edited by Norman Page, published 1985.
[3] In fact, there is no doubt *Lord of the Flies* (1954) criticism, to some extent, does suffer the bane of banality. As late as 2007, an article published in *Sino-US English Teaching* argues that *Lord of the Flies* expresses "the theme that evil is human nature" (Xiao-chun p. 61) – the same comment about *Lord of the Flies* that has been expressed in criticism for over fifty years.
[4] In the same volume in the same journal, *Notes and Queries*, in another review by a different author, a critic receives acclaim because "his meticulous scholarship eschews conjectures, destroys myths, restores reputations, and opens up new perspectives" (p. 511).

obliquely, as prescriptions and proscriptions governing how a genre ought to be used. But these 'rules' are ideal, and therefore often unrealizable. Consequently, meta-genres gloss over or occlude not only the multiform, rather than unitary, utterances which constitute a genre, but the inherent manner in which such utterances might seem to oppose the generic function they serve (what Bakhtin [1981], in "Discourse in the Novel," calls the stratification of language and internal dialogism of 'the word'). That is, meta-genres are systems of rules that are created and broken – necessarily broken, because this system seeks to act upon a situation alien to the actual situation immanent within the utterance.

And we do find that Swinden's criticism-as-difference rule cannot be maintained: not restricting his charge of banality to Golding's critics, Swinden also comments, essentially, that Golding himself plagiarized (replicated, reiterated) *Lord of the Flies* from Walter George's (1926) *Children of the Morning*. I wonder, though, if Swinden ever read *Children of the Morning*, because of his erroneous comments on the novels' supposed similarities, and also because he *himself* mimics remarks made by *Spectator* columnist Auberon Waugh in 1983 (to whom Swinden alludes, but makes no precise attribution). This is to say, *Swinden's charge of plagiarism is itself a plagiarism* or, less accusingly,[5] Swinden reiterates another critic's comments in a context he finds relevant. Thus, we see that repetition in criticism cannot simply be 'wrong,' but similar remarks might be repeated in potentially productive contexts: banalities become banal through perfunctory use, though iterated language finds productivity in its motivated repetition

[5] Because meta-genres do often express contradictory rules, it might be tempting to limit a meta-generic critique to a kind of disciplinary exposé. However, these observations are not made to highlight hypocrisy, and are certainly not conclusions in themselves, but *are* points of entry into critical discussion. Giltrow (2002) emphasizes the 'non-judgmental' nature of such an investigation.

and reception. But all of this ought to be less surprising than simply expected: we are engaged, after all, with literary criticism.

These questions of criticism, repetition, banality, and invention underlie the present investigation: a corpus study of the collocations found in *Lord of the Flies* literary criticism. A substantial portion of my thesis will be devoted to addressing just what collocation means and how it might be used as an analytic term. Without identifying myself as a 'collocation originalist,' I will argue that there has been a significant change in the way the term has been deployed—from John Firth, onwards—that sells short collocation's productive origins as a term of difference and uniqueness.

Literary criticism is an interesting area for such an investigation because, as much as I have questioned Swinden's suppositions—and would like to engage with the criticism-as-difference meta-genre—his question stands: just "*what is it about Golding* that makes critical conversation about him so banal?". Just as the arguments appear to be largely replicated, so too is the language critics use: seemingly poignant collocations are present within the discourse, formed from critics reiterating the same or similar phrases, with the result that this repetitive language is almost idiomatic. Consider the phrases "human nature," "fall of man," and others. And yet, while there is a certain degree of 'coalescence' of argumentation, the coalescence of language persists even when critics do make *different* arguments. To put it simply, critics use the same collocations to mean different things – *especially* when they disagree, they speak in *similar* ways. I am interested in formulaic language as points of divergence in meaning.

Divergence in meaning is an approach counter to that of current scholarship in collocation studies – though, as I have intimated, this need not be so. Chapter two will

present a short history of collocation, and propose that, first, it has drifted away from what J.R. Firth, the category's progenitor, originally intended or the possibilities he entertained. Second, historically, collocation has been considered in two irreconcilable ways: this I call the 'polar' distinction of collocation, where on the one pole collocation is considered statistically, and at the other pole it is considered semantically. What unifies these views, though, is that however 'collocation' is defined, there is a profound desire to establish how collocations function and therefore to bracket what they mean: collocations are assumed to have relatively stable, coded meanings and have therefore been subject to semantic conventionalization. I propose, however, that collocations not only defy semantic conventionalization but are pragmatic markers of context. We can therefore better understand collocation with a theory of communication, Sperber and Wilson's Relevance Theory, which suggests that the construction of communicative context and the establishing of Relevance within that context is the key to linguistic communication. In addition to presenting a history of collocation, in chapter 2 I will answer the first question—just how formulaic *is* this banal language—and present some intriguing findings in quantitative corpus analysis of phrases: in sum, it would appear that term *saliency* is strongly connected with both the frequency of these terms and how many different collocations incorporate these terms form (the degree to which one term associates or fraternizes with another). A term might be considered more salient when it is highly frequent and less commonly associating, and less salient when it is infrequent and commonly associating.

In chapter 3, I explicate more fully Relevance Theory, how this theory might explain collocation, and present a text tool I have developed from an integration between

Relevance Theory and Bibliometrics to visualize collocations in this corpus. A Relevance Theory approach to collocation offers a framework for understanding how collocations aid in the construction of context. My strong claim is this: some collocations are not used to convey meaning in the coded sense, but to establish a context for communication. In this way, the motive for the formal pairing of two (or more) terms into a collocation is to have the hearer recognize the act of the pairing itself as ostensive: this ostensive act makes manifest these words as the words of others, as echoed utterances, and therefore helps to establish meaning by inter-orienting the speaker's attitude amongst other, real or fictional, speakers' attitudes. In *LOTF* criticisms, we see this through the tool I present, and *experience* these echoes—potentially, as Swinden does—as resounding widely but diffusely: in other words, as banality.

In chapter four these tools and theories are put into action – I investigate the corpus for commonly occurring collocations, potentially banal collocations, and the collocates of negation. With the collocates of negation I explore, within the terms of Relevance Theory, what negation (e.g. "is not," "not just," etc.) activates for the reader's apperceptive background and frames as Relevant in the context of the body of criticisms.

At this point, I will offer some clarification and early speculation: invention, genres of criticism, banality, and collocations are all different things. I have chosen to investigate a body of work that supposedly defies the proscription against repetition (by way of banality) by looking at formulaicity in language, but this does not mean that banality *implies* formulaicity. Not necessarily so. My guess is that what unites this multi-faceted discourse of criticism under the banner of banality is a particularly *un*critical return to common contexts. These criticisms are not just banal because they

repeat one another, but reflect the novel—*Lord of the Flies*—and one another in a particularly specious manner. They are therefore not making the same arguments, but making *some* argument within the same 'space.' This space is what we might regard as context: critics do not make the same arguments, but do create the same context in which to make arguments. Banality is therefore not a substance, but a space – what is banal is by consequence a *formal* property. When we speak of banality we are much more speaking about the arrangement than the finding of meaning – language's 'form' as it relates to its 'meaning,' to frame this in terms of classical distinctions. Collocation is the result of an arrangement responding to its situations of use, and the solicitation of this use as meaningful.

1.2 A Note on My Corpus

My corpus is comprised of 46 English criticisms of *Lord of the Flies* published between 1960 and 2009, listed in the Modern Language Association International Bibliography. The corpus size is 200,000 words. A complete listing of the titles appears in Annex A. Although some of these works are available in digital format, I used a digital scanner and OCR (Optical Character Recognition) technology to convert print text to computer- readable files. Although OCR accuracy has been good (~97% accurate) I have had to correct many errors, and it should be acknowledged that some errors—however few, and even after a proofread—will remain in the final version of the corpus. To analyze this corpus, I used *AntConc* (Anthony 2011) and text-processing programs I developed in the course of this work. *AntConc* is a general concordancing tool, and I wrote a number of scripts in the *Python* (2.7.2) programming language – including the centerpiece text tool of this work, presented in chapter three.

2 Echoes and Collocation in *Lord of the Flies* Criticisms

2.1 Fiction on Criticism

I was the shadow of the waxwing slain
By the false azure in the window pane;
I was the smudge of ashen fluff – and I
Lived on, flew on, in the reflected sky. (L. 1-4)

--Vladimir Nabokov, *Pale Fire*

Vladimir Nabokov's (1962) *Pale Fire* is a seminal work in a genre espousing the fictional, narrative qualities of literary criticism. The work has two parts: the first is a 999 line poem entitled "Pale Fire," written by the fictional poet John Shade. The second is a criticism of "Pale Fire" by Shade's admirer, the fictional scholar Charles Kinbote (Shade's 'Boswell'). The text itself is presented as non-fiction in the typical form of the critical edition, edited by Kinbote. The critique of criticism as fictional and narrative is therefore delivered in part by telling a story about criticism: a subversion of form, the product of which is a 'third part' – a narrative fiction.

The idea of mirrors and echoes—repetition and replication, generally—abounds in this genre, and is reflected in the opening four lines of *Pale Fire*'s "Pale Fire": Shade declares himself the shadow of a dead bird, killed by striking a window it has mistaken for clear sky. Shade's shadow, his disembodied projection, persists in life in its reflection. Kinbote's criticism on these lines begins with a literal narration, then wanders—takes a 'flight of fancy'—to a story about birds from his fictional homeland. Whether Nabokov's attitude is that criticism—the reflection of text—is such a conduit

for this projection, harbouring an author's disembodied projection but motivating such flights of fancy, I cannot say: but Kinbote's criticism, of course, is also a reflection, both *on* the poem and positioning itself with*in* it. In any event, however, the basic idea of 'replication and criticism' is a salient pairing in the text.

In fact, there is a dual projection/reflection in the poem's first part: the one described above, oriented outward, and another with its original inward. The lines immediately following those reproduced above describe Shade delighting himself by standing inside his lighted house, in the dark of night, watching the inside of the house projected outside via a familiar optical illusion:

> And from the inside, too, I'd duplicate
> Myself, my lamp, an apple on a plate:
> Uncurtaining the night, I'd let dark glass
> Hang all the furniture above the grass, (L. 5-8)

These four lines are reflections of the first four and, this notion of reflection is furthered by a biblical allusion contained in these verses. The key phrase in this passage is "dark glass" (line 7), an obvious evocation of 1 Corinthians 13:12: "For now we see through a glass, darkly; but then face to face" (KJV). This evocation is obvious because in *Pale Fire*'s forward another, more explicit allusion is made to this same verse. In the fictional forward, Kinbote quotes Professor Hurley, who suggests that the surviving poem "Pale Fire" might be substantially shorter than what poet John Shade had originally intended: "None can say how long John Shade planned his poem to be, but it is not improbable that what he left represents only a small fraction of the composition he saw in a glass, darkly" (p. 2).

So, we can be reasonably assured that the recurrence of 'dark glass' is an intentional callback to 'glass, darkly' – an anaphoric reference to a semantically similar item, consequently establishing what Halliday and Hasan (1976) term "textual cohesion." [6] By creating this textual tie through the repetition of collocations, the biblical passage elicited in the first instance is imported into the second, compounding thematic reflections on optics. Now, all of this is to say: a discussion *about* reflections is augmented by a *textual* reflection. The semantics parallel the pragmatics. [7]

This is a truncated stylistic analysis, but I use this example to demonstrate the idea of collocation as textual reflection, or echoes, capitalizing on the lucky coincidence that such an example happens to be located in a genre concerned with such reflections: a genre of fiction demonstrating the fictionality of literary criticism. I say all of this to introduce the idea that formulaic language might be *productive* repetition, reflection, echoing. In the case of *Pale Fire*, this tightly constrained, internally resounding environment of echoes motivates a qualitatively precise path of investigation that leads to

[6]As Andrew Goatly (1994) points out, however, the second instance does not *directly* refer to the first as if it were unmediated by a mental representation of that 'thing': such an understanding of anaphora is a naive over-application of Hallidayan linguistics (p. 147). Goatly notes that Halliday and Hasan's term *endophora*, a referent established internally within a text, "is in many cases a misnomer" (p. 147), since a text can never be a closed system and is always constituted by the play between text and inter-text. This critique is obviously correct (a text's insides and outsides do not constitute a hard boundary), though I do think it is beneficial to differentiate between internally and externally resounding 'echoes,' as I will point out in the introduction to chapter three. And further, I will point out that my opinion is not that the second occurrence is a re-coding of the first—or at all a direct mimetic reflection—but gives evidence to the author's intent, called 'ostensiveness' in pragmatic theory, which ultimately does 'point to' an intertextual (in this example, a biblical) import. Additionally, in chapter three I will take up Goatly's theories in more detail, since, in the work referenced above, he is actually primarily concerned with Relevance Theory.

[7]I cannot help but mention here Jacques Derrida's (2007) "Psyche: Invention of the Other," which bears a strong family resemblance to the discussion here. In this work Derrida identifies utterances in Francis Ponge's poem "Fable" that are both 'constative' and 'performative,' and uses this conflation to support his critique of Speech Act theory. Our observations are relatively similar, although I am certainly not claiming *Pale Fire* underpins (or denies) some kind of linguistic truth.

productive criticism. As well, this is my first, formal refutation of the tacit claim that multiple authors saying the same thing is 'banal' and therefore 'wrong.'

Indeed, given the critical-edition form of *Pale Fire* the reader *is* led to believe that *multiple* authors' consciousnesses are at work, even when this is clearly not the case. In the 'dark glass' example above, two different authors (Professor Hurley, reported through Charles Kinbote; and John Shade) echo this term across *different* documents rather than across sentences *within* a document (intertextual versus endophoric reference, as noted in footnote 6). And although such repetition from two different consciousnesses might be disruptive in fiction, because of *Pale Fire*'s critical form—in which an intersubjective 'sharing of consciousness' between critic and author is expected—this resumptive use of 'glass, darkly' satisfies rather than disrupts the genre. Repetition is intersubjective and anticipatory: as I will argue in this chapter, collocation is also repetitive, intersubjective, and anticipatory.

I propose that the exciting qualities of collocation I have intimated above have disappeared in contemporary theorizations of formulaic language. J.R. Firth, who introduced the notion of collocation in 1951, was concerned with the magic of language: the idea that manipulating the environment (producing minute disruptions in the air, for example) could do something as amazing as share a consciousness. And although his description of collocation was fairly ambiguous, I think a certain productive quality existed in his formulation of the term that has progressively been stifled. What we are left with is a fairly banal conception of the concept: two or more terms occur together that might be assigned some kind of sense or meaning – collocation as a process of semantic conventionalization. I will argue, however, that this denies the full potential of

collocation, historically and pragmatically. Although *frequency* has become a key consideration in collocation, where the status of collocation is conferred as if it were an award based on commonness of occurrence, I will argue that this is *one* measure – but not the whole story. Frequency is important, but only insofar as it denotes *repetition* and, as Nabokov insinuates, reflection is life giving. We might do better justice to collocation by restoring its magic.

2.2 Collocation

This present study is of formulaic language in literary criticism – a corpus of literary criticism perceived to be formulaic. Patrick Swinden (1987) builds his charge against criticisms of William Golding's *Lord of the Flies* (*LOTF*) by calling them "banal" – repetitive and familiar. Such criticism therefore both contravenes *and* abides by meta-genres of composition: by establishing a pattern a certain rule of textual construction is established, and using pre-set patterns lends a sense of generic fluency, but too much repetition—and too many formulas—is frowned upon. In pedagogy, too, idioms and other formulaic phrases have become very popular since mastery over these phrases is considered one of the last stages of language mastery (Moon 1998; Hill 2000). On the other hand, however, an over-reliance on prefabricated expressions—even for users perfectly fluent in their national language—can mark a certain generic disfluency: comparing published academic writing from scholars with varying levels of experience (graduate and post-graduate), Ken Hyland (2008) finds that more experienced academics' writing actually contains fewer clusters than their less experienced counterparts. Viviana Cortes (2004) finds that post-secondary students do use lexical bundles (trigrams identified by their frequency), but infrequently and inexpertly. Alan Partington (1998),

quoting Ronald Carter (1987), also notes the downside to collocation: "Too much respect for normal collocation and grammar produces language which is 'too familiar and thus banal'" (p. 17). Formulas, therefore, are both the best and worst parts of writing.

The question, then, is under what conditions is repetition variously banal or productive? When is it good and when is it bad? This question will be taken up by looking at formulaic language—and specifically collocation—in a corpus of *LOTF* criticisms. In this chapter I will briefly gloss the history, theory, and philosophy of collocation and propose that the theory of the phenomenon itself has become banal – and further propose how it might be revitalized. More pointedly, I will argue that the popular conceptions of collocation are irreconcilable—with 'collocation' as statistical on one hand, and as phrases with intuitive meaning on the other—and that this dissimilarity reproduces the classical rhetorical distinction: that between form and content, or words and meaning. Collocation might be revitalized by reconciling this distinction.

The study of formulaic language, especially in academic writing, is not fresh terrain. Douglas Biber has written extensively on the subject, especially in *University Language: A Corpus-Based Study of Spoken and Written Registers* (2006) – as well have others contributing to (and essentially founding) the study of formulaic language. An abbreviated list: Firth (1951, 1957); Sinclair (2001, 1991); Kjellmer (1991); Herbst (1996); Moon (1998); Fernando (1996); Biber and Conrad (1999); Wray (2002); Wray and Perkins (2000); Bartsch (2004); Hyland (2000, 2008); Cortes (2004, 2008); Götz-Votteler and Herbst (Eds., 2007). Furthermore, in keeping with the quantity of theorists engaged in the subject, the number of terms developed to identify this (general) phenomenon is also extensive: Wray and Perkins (2000) identify forty-five different

terms denoting formulaic phrases (pp. 3-4), beyond 'collocation', from "Amalgams" to "Stable and familiar expressions with specialized subsenses" (p. 3). Indeed, the *diversity* of terms used to express a similar (or the same) phenomenon is revealing: "The *multifaceted nature* of *formulaic language* is evident from the variety of ways in which it has been characterized" (Wray and Perkins, 2000, p. 3, emphasis mine). It would seem that the dynamic potential of formulaic language is reflected even in its nomenclature.

Furthermore, from the discipline of language pedagogy, Jimmie Hill (2000) identifies the historical development of terms to describe formulas, and also suggests how this development influenced the use of these categories in language instruction. Hill writes that the range and extent of formulaicity was fleshed out progressively, and that "[i]t is only recently through the rise of corpus linguistics that the extent of the fixedness of much language has been more widely recognized" (p. 50). Furthermore, given the historical precedence of these older terms, new types of formulas have been given less attention:

> It seems sensible to continue using those terms and categories which language teachers have found useful in the past – idioms and phrasal verbs – while introducing the term *collocation* to name and categorise that language which has previously been ignored or undervalued. (p. 50)

To summarize these two perspectives, Wray and Perkins (2000) give us a synchronic view of the field, while Hill (2000) provides somewhat of a historical comment. What is evident from both is the idea of discovery and dissatisfaction: as the field expanded, the new-found terms were found to be inadequate and another category was proposed. Hill (2000) suggests that formulaicity was first detected as belonging to one of two types of

features, between the syntactic configuration of phrasal verbs and semantically-encoded idioms,[8] and Wray and Perkins (2000) identify just how many other terms now populate the field. My own proposal is that this polar view of collocation re-animates the false distinction between form and content – and that this top-down categorization is an overly ambitious attempt to inscribe formulas with semantic meaning, which consequently erases their pragmatic potential. Even at a rather high level, we see that the *experience* of formulaic language is salient and meaningful (connoting fluency and banality) and therefore reducing these formulas' meanings to fine-grained categorization denies these higher-level functions.

Far from being merely 'banal,' then, we might conceive of formulaic language as the product of banality *as* difference, embodying Mikhail Bakhtin's (1981) centripetal and centrifugal linguistic forces: "Alongside the centripetal forces, the centrifugal forces of language carry on their uninterrupted work; alongside verbal-ideological centralization and unification, the uninterrupted processes of decentralization and disunification go forward" (p. 272). Collocations are by necessity standardized forms, vouched for by social convention but also ready for exploitation, for reformulation. The acquisition of collocations is evidence of speakers gaining fluency in a national language, but also arises as an inventive principle *of* language: in new dialect formation, for example, as Edgar Schneider (2003) points out (p. 249). Thomas Herbst and Katrin Götz-Votteler (2007), moreover, begin their introduction to a special edition of *ZAA*, "Collocation and Creativity," with the claim that, "[t]he term *collocation* presents an almost prototypical

[8] Of course, even in this polar conception these categories can also be combined. For example, residents of New Jersey might say, "I'm going down the shore" to indicate that they are travelling to the beach from some more inland location. In this case, *go* collocates with *down*, and the elided preposition grants the phrase idiomatic meaning where this 'special meaning' would be absent in other constructions: "I'm going down *the road*," for example (which, in turn, can have its own idiomatic meaning).

example of the phenomenon of polysemy" (p. 211, emphasis in original). And although Herbst holds the polar view of collocation (as I will show later) it is indeed this fact of the phenomenon—polysemy, the inescapable *potential* for recombination—that underpins its variability.

So although the study of formulaic language in general—and collocations more specifically—may not be situated on fresh terrain, likely owing to its innate volatility neither has it become a developed 'field': just the opposite, as the site has been thoroughly trodden yet the scholarly turf remains disturbed rather than cultivated. Indeed, this muddied terrain persists despite collocation's hopeful origins.[9] In "Modes of Meaning" (1957 [1951b]), John Firth introduces the term 'collocation' as a level of analysis in his descriptive linguistics, a term denoting the expectancy that one word be located next to another: "One of the meanings of *ass* is its habitual collocation with an immediately preceding *you silly*" (p. 195, emphasis in original). And although Firth explicitly differentiates collocation from a word's contextual (p. 195) and lexical meanings (p. 196), this seems to be overpowered by his oft-quoted command: "You shall know a word by the company it keeps!" ("A Synopsis of Linguistic Theory" 1968 [1957] p. 179). So although it might seem that Firth relates word proximity to a 'semantic horizon of meaning,' where words in close proximity modify each other's meanings in the sense that they 'blend' or 'leech,' this is not the case and (I propose) a frequent misreading. Academics who are not directly concerned with linguistic investigations of language tend to latch on to this statement as a catchy 'quote,' justifying linguistically the

[9] Although here I imply that it was John Firth who coined the term—and it is fairly clear that he did popularize its linguistic use—Sabine Bartsch (2004) provides evidence that the term was actually in use, in similar contexts, before Firth (pp. 30-32). Further, Cortes (2004) identifies nineteenth and early twentieth century uses of the term.

motive for 'contextualization' (as if this motive was disputed). Firth's concern *is* with meaning, but meaning through form: "Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words" (1957 [1951b] p. 196). His objective is actually to relate a word's placement within a text (next to other text) to its use, and therefore to its meaning-as-placement.

I find Firth's original formulation, however vague, convincing and productive in its focus on *form*, *expectation*, and *use* because these directly relate to a central theme of this study: banality. Etymologically linked with 'banal,' in addition to 'common' and 'trite,' are also the concepts of history, class, and power: for 'banal', the *Oxford English Dictionary* (2010) notes a more archaic meaning, "[o]f or belonging to compulsory feudal service"; "a kind of feudal service, whereby the tenants of a certain district are obliged to carry their corn to be ground at a certain mill, and to be baked at a certain oven for the benefit of the lord." We might think of collocation as a type of feudal service, in that words are 'expected' to serve a certain function by appearing at a certain place at a certain time. This is how Firth separates context, meaning, and form: at the level of collocation, part of a word's *meaning* (because of its compelled function) is actually determined by its behaviour around other words. It really seems to be akin to a 'lexical behaviourism'[10] – we might know a word by observing how it acts.[11]

[10] In a particularly odd passage from "A Synopsis of Linguistic Theory," Firth describes the agency of words: "Though Wittgenstein was dealing with another problem, he also recognizes the plain face-value, the physiognomy of words. They look at us!" (p. 179.)

[11] John Firth is indeed a behaviourist, generally, rejecting the duality between mind and body: "As we know so little about mind and as our study is essentially social, I shall cease to respect the duality of mind and body, thought and word, and be satisfied with the whole man (sic), thinking and acting as a whole, in association with his (sic) fellows" ("The Technique of Semantics", 1951[1957a], p. 19).

And while I like the idea of focusing an investigation of formulaicity on the behavioural traits surrounding collocations, in addition to examining just 'what the constituents of collocation do' I would like to consider *why* these words do what they do, how they are pressed into service, the wide variety of functions this service might entail, and the difference between the potentialities and realities of these functions. These peasant-words are, essentially, co-opted constructions, put under various pressures and pushed into acting in a manner—and therefore having a certain meaning—dislocated from their histories. As Bakhtin notes, "The word lives, as it were, on the boundary between its own context and another, alien, context" (1981, p. 284). So while we must look at the behaviour of a word, we must also question the wide range of social and linguistic forces—the *genre* of criticism, and the *discipline* of criticism—exploiting this productivity. This is the crux of my argument for collocation, and therefore banality, as *difference*: as a phenomenon, formulaic language arises to accommodate wide ranges of situations and not a reduction of meaning as has been previously posited. As a particular feature, formulaic language is a refined site that glosses over multiple alien utterances, a meeting place of variously inflected words and meanings, and the remains of a confrontation between discordant text. Like meta-genre, and akin to feudal service, my hypothesis is that collocations are marks of potential *rebellion*.

Finally, I am modeling this discussion of difference and productivity on rhetorical genre theory (Miller 1984), which recognizes that genre, rather than serving as a restrictive template for communication that passively provides templates for expression, responds to and motivates a rich production of utterances (Giltrow and Stein 2009): as Bakhtin writes, "stratification is accomplished first of all by the specific organisms called

genres" (Bakhtin, 1981, p. 188). Like collocations, we can only see genres as a formal repetition—as Giltrow (2009) claims, just one instance does not make a genre—but this does not mean that meaning is reduced to the most frequent realization, or to either their constituent elements or idiomatic meanings. I contend that collocations, as textual echoes, have a 'third part' to their meaning: they are traces of their socially motivated production, evidence of common practices of authors and expectations of readers.

But all of this, Firth's formulation, one that provides for an investigation of sameness *and* difference, has been developed in the last half-century into a conception of collocation that erases this hybridized nature. This has been accomplished partly through a shift in emphasizing the *frequency* of collocation as prior to its function. The "collocation" entry on *Wikipedia* (2010) reflects this new, often-used definition: "Within the area of corpus linguistics, collocation defines a sequence of words or terms that co-occur *more often than would be expected by chance*" (emphasis mine). But language is not just 'chance,' consequently pure statistics reveals very little in terms of collocation, and there seems to be no reason why the greater instances of collocation are now privileged as opposed to the fewer. It appears that an increasing desire for disciplinary objectivity signaled this shift. As Vivian Cortes (2008) notes, "most of the attention on [formulaic language] has shifted towards the analysis of recurring word combinations identified empirically rather than intuitively, as in the case of lexical bundles" (p. 43). Antonia Martínez (2010) states this prejudice bluntly: "We will regard collocations as the statistically significant co-occurrence of words within a short span in a text" (p. 763). Instead of "silly asses" we now speak of statistics.

But what is *linguistic* 'statistical significance'? Regarding such statistical significances—and specifically in regards to Mutual Information—Sabine Bartsch (2004) writes: "Unfortunately, the assumption of a random distribution, i.e. of a completely independent distribution of words in a language sample, is a mere methodological convenience, a myth that does not reflect faithfully the reality of linguistic structure" (p. 100).

Firth's most helpful methodological suggestion for disambiguation, directed away from a focus on frequency, is this: "Statements of meaning at the collocational level may be made for the *pivotal* or *key words* of any restricted language being studied" (1957 [1968], p. 180, emphasis mine) – and this, in order to narrow the field of study. Although Firth does use the word 'habitual' with collocation (habitual is a collocate of collocation), I think that its sense is better expressed as 'expectant' rather than 'frequent': 'habitual' as it relates to a particular disposition or attraction. (For example, I can 'expect' red wine with a meal of red meat, and not just because this is, statistically, the culinary combination of choice.) Lexical items, like foods, can just taste good together, and their repeated arrangement does not *denote*, but *is a result* of the satisfaction of this expectation. This type of analysis, the criticism of collocations' frequencies, does have its critics: Wray and Perkins (2000) agree with limiting the priority afforded to frequency, citing Tina Hickey (1993) and Peter Howarth (1998), suggesting that "it may be premature to judge frequency as a *defining* feature of formulaicity. It has yet to be established that commonness of occurrence is more than a circumstantial associate" (p. 7, emphasis in original).

To this point, I have sketched the character of collocation and formulaic language in current scholarship, and their historical origins. I propose that the dominant, polar point of view frames formulaic language as sets of phenomena bound to two disparate poles: formulas as statistical on the one hand, and idiomatic on the other, with the rest of the field populated by categories of varying functions. But the problem with this is that this top-down categorization defies Firth's original formulation (if not in its nature, then in its effect), and denies the productive nature of formulas because a top-down characterization obviates an analysis of the socially-situated and 'on the ground use' of such phrases: collocations, as top-down categories, are assigned single meanings or a finite set of meanings. What I have yet to address is *why* this shift in meaning occurred, and how this fits into a broader history of such linguistic dichotomies. I suggest that corpus linguistics motivated this shift to frequency-based, top down categorizations, and that the result maintains the classical rhetorical division between innovation and disposition.

Indeed, often this polar characterization is quite rigidly defined, where collocations are conceived as belonging to *either* one side *or* the other: Thomas Herbst (2007), for example, distinguishes between collocations along lines of quantitative and qualitative reasoning, types he terms 'sandy beaches' and 'guilty conscience'[12]:

> One type, the *sandy beaches*-type, refers to specificity in statistical terms of co-occurrence in the language, or in a corpus, -- where statistical significance is not necessarily determined in terms of absolute frequency of co-occurrence but

[12]In an earlier work by Herbst, entitled "What are collocations: Sandy beaches or false teeth?", this distinction is typified by the collocations "sandy beaches" and "false teeth." Herbst argues that collocation ought to be understood in the restrictive sense, where "false teeth" counts as a collocation but "sandy beaches" does not since "false teeth" contains irreplaceable units but "sandy beaches" is based on statistical frequency.

> calculated on the basis of some sort of measure of mutual expectancy. In the second type, the *guilty conscience*-type, the combination is significant because it is established or institutionalized, to use a word common in word formation, and somehow unpredictable on the grounds of the meaning of the words. (p. 211)

That is, a distinction is made between statistics and meaning, where in the second case a collocation is identified as such by something along the lines of idiomaticity in the manner of "semantic opacity" (Sinclair 1991). While "beaches" tends to occur with "sandy" (Herbst's corpus for this is a collection of European travel brochures), a "conscience" need not be guilty—there may be no statistical justification for this combination—and further, the meaning of a "guilty conscience" might not be deduced from simply the combination of adjective and noun. Further, Herbst assigns these two categories, "quantitative" and "qualitative" (statistical and idiomatic) collocations, to different disciplines: corpus linguistics and foreign language linguistics, respectively. This typification is common. In addition to Hill (2000), referred to at the beginning of the chapter, in *Patterns and Meanings* (1998) Alan Partington sketches a three-part division: "textual" (Sinclair 1991), "statistical" (Hoey 1991), and "psychological" / "associative" (Leech 1974). We have seen the last two types before, representative of statistical and institutional collocations, while the first—'textual collocation'—refers not to 'meaning' but to the orthographic linearity that constructs phrases. Textual collocation, as Partington writes, "is a consequence of the linearity of language, or, conversely, if we view text as a process rather than product, it is the principal method, together with syntax, with which this linearity is constructed" (p. 15). Nadja Nesselhauf

(2005) reiterates this polar distinction: "Among the many diverse uses of [collocation], two main views can be identified" (p. 11).

So, collocations are classically categorized as either subjective or objective, and objective collocations belong to the domain of corpus linguistics. But corpus linguistics, as Charles Meyer (2002) points out, is not so much a discipline as a method – and therefore it would stand to reason that the *technology* of corpus linguistics is a motivating factor in the shaping of a particular notion of collocation and, further, it implies that such theorizations are capable of being shaped by this type of technological / methodological interaction in the first place. The rise of computer-searchable collections seems to have motivated the statistical vein of this division, and the collocation algorithms in most concordancing software maintain this by displaying a corpus' collocations, most usually, according to their frequency of appearance.

Firth's original formulation, in its vagueness, did not lean one way or another between these 'poles,' and therefore this originary theory seems unlikely to have motivated these disparate veins. In fact, what is truly amazing is that Firth is invoked in association with *either* side of the polar conception: Partington (1998) claims the *subjective* category is akin to Firth's 'expectancies' (p. 16), while Nesselhauf (2005) also ascribes the development of the *objective* category to Firth: "The frequency-based approach goes back to J.R. Firth and has been developed further in particular by M.A.K. Halliday and J. Sinclair" (p. 12).[13] This mixed-up origin story also finds its way into reference manuals. For example, the *Routledge Dictionary of Language and Linguistics* defines 'collocation' as a "[t]erm introduced by J.R.Firth in his semantic theory to

[13]Sinclair and Halliday were students of Firth, though I do think that their conceptions of language diverged considerably.

designate characteristic word combinations which have developed an idiomatic semantic relation based on their frequent co-occurrence" (1998). Here we find a grab-bag of terms pulling from either side of the polar distinction: 'idiomatic,' 'semantic,' 'frequent.'

Finally, if we maintain the distinction of collocation as between statistical (or even simply orthographic) and semantic categories, we deny what binds these categories together: collocation's inventiveness, its productiveness. However vague Firth's formulation of collocation was he does suggest collocation as a generator for—rather than a container of—meaning. Through combination and iteration collocation is productive in the way it yokes terms together and maintains these groupings as novel, subject to re-use and re-formulation. An exploration of this productiveness, however, is attenuated by maintaining such a polar distinction. And further, such productiveness is meaningful for listeners and speakers. As Ken Hyland (2008) points out, "while clusters are simply statistical regularities of language use for the analyst, they actually reflect a lived reality for users" (p. 44). This sobering thought refocuses our attention on language-use, language-users, and language-listeners. The concern we ought to investigate, then, is one of *salience*: as I have argued, the preoccupation with collocation and frequency is not necessarily productive, but this does not mean that frequency means nothing. In what ways does frequency *mean*? How can we use corpora to capture clusters and the 'lived reality' they reflect? Of what significance are collocations to *salience*? We now turn to a corpus of language perceived as formulaic—the *LOTF* corpus—to answer some of these questions.

2.3 Fun with Words and Numbers

So, how formulaic is *LOTF* criticism? What measures do we have to represent this? Of course, the *sense* language engenders is separate from words' explicit realization (or statistical distribution), as anyone who has performed any type of corpus search whatsoever (even a keyword search, online, for example) knows very well: this project is motivated by Swinden's (1987) experience that *LOTF* criticism is 'banal,' and although I, too, sense a repetition and near-idiomaticity in reading the criticisms this does not mean it will necessarily be measurable, especially in terms of raw frequency (by way of ordered frequency lists, etc). Indeed, as I have been arguing this should not be the case. However, from reading in the corpus I do know that certain word clusters are present, collocations such as "human nature" and "inherent evil," that *do* contribute to the sense—and therefore banality—of the criticisms.

So, to get the most basic sense of the word clusters in the corpus, I present in Table 2.1 an ordered list of the 30 most frequent 2 to 5-grams:

Rank	n-gram	Frequency
1	of the	2733
2	in the	976
3	and the	678
4	to the	557
5	on the	482
6	Lord of	468
7	Lord of the	459
8	the Flies	441
9	of the Flies	437
10	Lord of the Flies	419
11	from the	380
12	it is	365
13	is the	340
14	of a	334
15	to be	334
16	that the	329
17	by the	323
18	the island	317
19	the boys	315
20	is a	305
21	of his	301
22	for the	299
23	with the	298
24	he is	275
25	as a	253
26	at the	251
27	the novel	246
28	the beast	226
29	in a	212
30	as the	207

Table 2.1: Ordered list of the 30 most frequent 2-5 grams

These results, at first glance, are hardly the dramatic picture of repetition. As tends to be the case, the top grams are dominated by function words: “of the,” “in the,” “and the,” etc. The size of the grams, too is quite small: only three of the top 30 are larger than a bigram, and they are, predictably, elements of the title: “lord of the,” “of the flies,” and “lord of the flies.” (The next trigram appears in ordinal place 54 with 110 instances, “on the island.”) However, this is not to say that intuitions about the corpus are wrong, and

poignant collocations are not also frequent: “human nature” does occur, quite high on the list, at place 123 with 68 instances. Further, other clusters, such as “the beast” and “the world,” though not immediately obvious clusters, do connote a reflection on some sort of essence of humanity. Collocates of “evil” are much more dispersed. “inherent evil” only occurs 7 times, but “evil” itself is a highly productive—though dispersed—term, occurring 301 times and in 110 different multigrams.

The above chart indicates fairly common phrases occurring, and does not exactly reveal the banality of the corpus. But another very basic measure of formulaicity is to compare single term frequency with phrase frequency. The following table, Table 2.2, is an ordered list of the 30 most frequent 1 to 5-grams:

Rank	n-gram	Frequency
1	the	14811
2	of	8590
3	and	6078
4	to	4515
5	a	4118
6	is	4109
7	in	3350
8	of the	2733
9	that	2381
10	his	1902
11	as	1674
12	The	1527
13	he	1506
14	it	1347
15	with	1226
16	for	1188
17	by	1110
18	on	1108
19	are	994
20	in the	976
21	not	959
22	from	914
23	be	877
24	Ralph	853
25	an	835
26	which	827
27	has	800
28	Golding	756
29	was	750
30	at	719

Table 2.2: Ordered list of the 30 most frequent 1-5 grams

Here, again, we find nothing obviously, or quantitatively, formulaic: of the top 50 items only two are multigrams while the rest are single terms.

However, these results may be representative of the method rather than of the data. Matthew O'Donnell (2011) points out that the ordered frequency list is inherently flawed, to the point that—in some sense—it skews representations of frequencies. These representations are skewed because, essentially, the ordered frequency list counts terms

multiple times. In Table 2.1, for example, there are 459 instances of "lord of the" and 415 of "lord of the flies." And while this is an accurate representation of these phrases' frequencies, it is rather *insensitive* to clusters: if "lord of the" is always a constituent element of "lord of the flies," then why is "lord of the" counted separately? In this case, "lord of the" is in essence counted twice. Furthermore, if we add up the purported number of the occurrences of "the" in this list, including "the" in any multi-gram, we arrive at a number far surpassing its actual occurrence in the corpus. In essence, the ordered frequency list is a contextless representation of data in a very deep sense, because the magnitude of any one item only holds true if it is evaluated in absence of all other items. So, if we are interested in clusters (and we *are* interested in clusters), we should only count these larger chunks and not their constituent elements. The problem is greater with highly productive terms, such as function words, where "the" and "of" occur with extremely high frequencies but, of course, also occur in a great many chunks.

To fix this problem, O'Donnell (2011) proposes the "adjusted frequency list," an ordered list that counts only the 'largest chunks' and leaves *uncounted* those chunks' constituent elements. For example, in his corpus from the *BNCBaby Demographic* O'Donnell shows that while the pronoun "I" is the most commonly occurring item in an unadjusted, ordered list of n-grams, this term frequently constitutes the phrase "I don't know." Adjusting for the frequent collocation of terms, "I don't know" actually becomes the top rated item while the frequency of "I" is reduced significantly. (Of course, so too would be reduced the frequency of the other single items, as well as the bi-grams: "don't" and "know", as well as "I don't" and "don't know.") Using this method, a certain threshold frequency must be reached for a multi-gram to reach 'chunk' status – that is, to

qualify as a 'top-level' chunk and subsume its constituent elements. This method has two functions: first, to suppress unigrams that primarily—or exclusively—appear in larger chunks; and, by consequence, to therefore 'elevate' chunks in terms of their prominence. An adjusted frequency list applied to a corpus with many recurrent phrases will therefore elevate these phrases over single terms.

Applying this method to his corpus, O'Donnell's results are dramatic: the top ten items in a 1 to 5-gram combined *unadjusted* frequency list are all unigrams: mostly pronouns, articles, and discourse markers. However, in the *adjusted* list five of the top ten results are multi-grams (interestingly, all trigrams). Phrases include "I don't know," "do you want," and "I don't think." O'Donnell's results are impressive, and suggest not only a simple way of approaching a formulaic data set but also that his set of data *is* highly formulaic. Indeed, the premise of his investigation is founded on O'Keefe et al.'s (2007) contention that "many chunks are as frequent or more frequent than the single-word items which appear in the core vocabulary" (p. 46). O'Donnell's results bear this out.

So, given the success of O'Donnell's investigation, I set out to apply this method to my corpus. To do so, I developed a suitable algorithm and programmed a piece of software with the open source programming language *Python* (2.7.2). While conceptually simple, the algorithm itself is somewhat complex and although I cannot give space to an explanation here I can be contacted for further discussion and for the source code.

However, the results I achieved were less than compelling. With my data, the difference between an adjusted and unadjusted list is minimal. Below (Table 2.3) is an

unadjusted list on the left, and next to it on the right is an adjusted list (1 to 5-grams, minimum frequency of 3) of the top 30 terms in my *LOTF* corpus:

Raw (unadjusted) Frequency List		
Rank	**n-gram**	**Frequency**
1	the	16463
2	of	8637
3	and	6263
4	to	4584
5	a	4276
6	is	4114
7	in	3671
8	of the	2807
9	that	2410
10	his	2018
11	he	1883
12	as	1796
13	it	1631
14	for	1287
15	with	1262
16	on	1180
17	by	1141
18	in the	1097
19	but	1035
20	are	997
21	not	986
22	from	953
23	this	895
24	be	875
25	an	845
26	which	834
27	ralph	825
28	at	810
29	has	803
30	they	794

Adjusted Frequency List		
Rank	**n-gram**	**Frequency**
1	and	1500
2	the	894
3	of	540
4	of the	481
5	in	424
6	a	418
7	in the	379
8	or	378
9	is	372
10	to	369
11	and the	347
12	his	343
13	that	337
14	to the	275
15	he	263
16	by	254
17	as	245
18	with	244
19	but	225
20	for	221
21	are	217
22	by the	211
23	this	207
24	which	192
25	of a	186
26	it	180
27	with the	178
28	they	174
29	ralph	174
30	on the	166

Table 2.3: Unadjusted and adjusted term frequencies

Even with this method, the data hardly denote the picture of formulaicity. In fact, I applied this method to multiple collections of text, including a million word corpus of Canadian provincial superior court decisions (Giltrow 2008) and—for a sample of

spoken discourse—a collection of transcripts from a television show of political criticism (97 episodes of *Glenn Beck*), but I did not observe near the elevation of chunks—and concomitant suppression of single terms—as did O'Donnell (2011).

In fact, in terms of the *LOTF* corpus not only did this method fail to show its formulaicity, it tended to *hide it*. As you might recall, in the unadjusted list the intuitively commonly occurring phrase "human nature" had a relatively high frequency, with 71 instances. However, in the adjusted list, "human nature" is pushed all the way down to place 2336 with only 9 instances. This is because the phrase "human nature" itself occurs within many other common phrases—and only 9 of those instances are discrete—and therefore the frequencies of it and related chunks are essentially rent apart and dispersed around the bottom of the list. Their formulaicity renders certain terms invisible.

Now, this is neither to criticize O'Donnell's method nor to find fault with his claims: the adjusted frequency list was never claimed to be a revolutionary method for displaying data. However, it is both disappointing that such a method is not applicable in corpora experienced as formulaic, and interesting that it should actually suppress the formulaicity confirmed by meaner means. Consequently, another metric was required, one that captures the very idea of formulaicity: the degree to which words form chunks in the corpus.

2.4 A New Hope

The motivation for generating another metric came from my disappointment with the adjusted frequency list. As I have noted, not only does it not sufficiently elevate chunks in my corpus, it effectively hides them by suppressing and 'dispersing'

potentially key terms. The term 'human,' which is a top-100 term (384 instances), has only 39 discrete instances and the phrases it forms have such low frequencies they are dispersed around the bottom of the list. The term is almost *too* productive: too commonly occurring and widely associated, to the point that its adjusted frequency drops too low to be considered 'salient.' So, I developed a metric to rank single word terms based on their proclivity to form chunks (called 'collocability'). What I was after here was a measure that would capture those highly formulaic phrases based on the highly productive term "human" – essentially, a measure of banality as a *diffuse repetition*.

Interestingly, however, ranking terms just by how many chunks they form essentially reproduces a standard frequency list – the principle here is that a *frequently occurring* term also tends to be *widely associating*. That is, frequency is correlated with collocability. The following, Table 2.4, is a list of the top 30 terms, ranked by how many chunks the terms comprise and with their respective frequencies:

n-gram	# chunks	Frequency
of	2666	8637
to	1348	4584
his	414	2018
as	398	1796
he	385	1883
the	353	16463
with	331	1262
on	322	1180
from	267	953
is	257	4114
by	254	1141
and	235	6263
was	223	754
an	203	845
has	199	803
a	193	4276
at	180	810
they	177	794
that	151	2410
this	142	895
have	140	567
we	139	696
their	129	615
no	123	434
but	123	1035
for	114	1287
i	106	462
out	99	341
like	98	414
what	95	457

Table 2.4: 30 most frequent terms, ranked by chunks and frequencies

The correlation between frequency and collocability is not perfect but, as Figure 2.1 shows, it is very close. Figure 2.1 plots the top 30 terms (based on their collocability) with the scale of term frequency on the left and the number of chunks the terms form on the right. Given this tight correlation, there seems to be no benefit to calculating a metric based purely on collocability – term frequency must be taken into account, too.

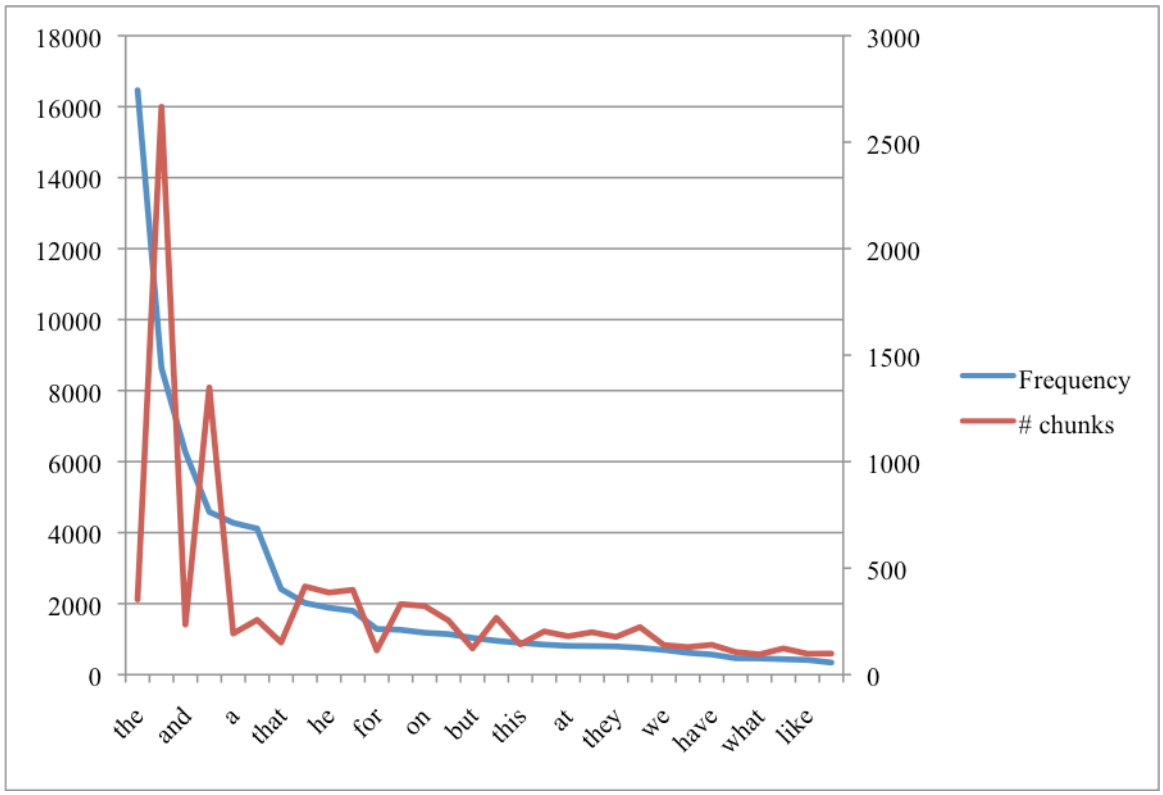

Figure 2.1: Collocability correlated with term frequency, 15 terms

After some experimentation, I did find an algorithm that produces an interesting result. This algorithm takes the logged frequency of a term and divides this value by the number of chunks in which the term appears. Thus, both collocability—expressed as the number of chunks a term forms—and term frequencies are taken into account. The formula, then, is *log(f)* / #chunks, which can be conceptualized as the number of chunks per occurrence of a term. Importantly, the number of chunks in which a term is found is determined using the algorithm from the adjusted frequency list: that is, only the 'top-level' chunks are considered discrete units in this count:

> E.g.: "Lord of the Flies" is a 4-gram. If we were determining the number of chunks in which 'lord' occurs, we would *not* count "lord of" or "lord of the" (assuming that these chunks do not occur without "flies"). That is, only top-level,

discrete chunks are counted. This is, therefore, an offshoot from and a significant benefit of the adjusted frequency list.

This metric works because the resulting value will be highest when the frequency of a term is high and the number of chunks in which the term appears is very low. This is to say, the terms at the top of the list are commonly occurring but narrowly associating, while the terms at the bottom are freely associating.

The results of this method are quite dramatic, because the list it produces separates what I call 'salient keywords' at the top from function words at the bottom. This separation is so clear-cut because the 'top' and 'bottom' sets of terms, combined, actually resemble the top terms produced from a standard frequency list – it is almost as if a stopword list was used instead of the natural patterning of lexical frequencies and collocability:

Saliency Adjusted List	
Bottom of List	
Rank	**n-gram**
3531	and
3532	has
3533	an
3534	is
3535	was
3536	by
3537	the
3538	from
3539	on
3540	with
3541	he
3542	as
3543	his
3544	to
3545	of
Top of List	
1	jack
2	him
3	head
4	itself
5	fable
6	roger
7	thing
8	forest
9	boys'
10	savagery
11	god
12	indeed
13	course
14	hunting
15	heart

Standard Frequency List	
Rank	**n-gram**
1	the
2	of
3	and
4	to
5	a
6	is
7	in
8	that
9	his
10	he
11	as
12	it
13	for
14	with
15	on
--	--
16	by
17	but
18	are
19	not
20	from
21	this
22	be
23	an
24	which
25	ralph
26	at
27	has
28	they
29	was
30	we

Table 2.5: Saliency adjusted list and frequency list

This separation between function words at the top and 'salient' terms at the bottom occurs because although the top and bottom sets of terms both occur very frequently, the bottom, functional set occurs in many more chunks than the top set. Term

frequency is taken into account, but modified with a mathematical function such that a linear, rather than logarithmic, progression is approximated. Essentially, this means that "Ralph" certainly does occur less than "the," but not so much less. Very briefly, there are two simple methods for effecting this transformation (for converting a logarithmic curve into a linear progression): the first is taking the *log* value of each term's frequency. This 'flattens' the curve. The second, simpler method is to simply assign each ranked term the value of its ordinal position in the list. Mathematics aside, a practical demonstration of working with frequency distributions: the following charts plot the frequencies of the first 500 words in the corpus using three different methods. The first chart plots frequency (raw count), the second plots the logged frequency, while the third plots the ranked position of the word in the list:

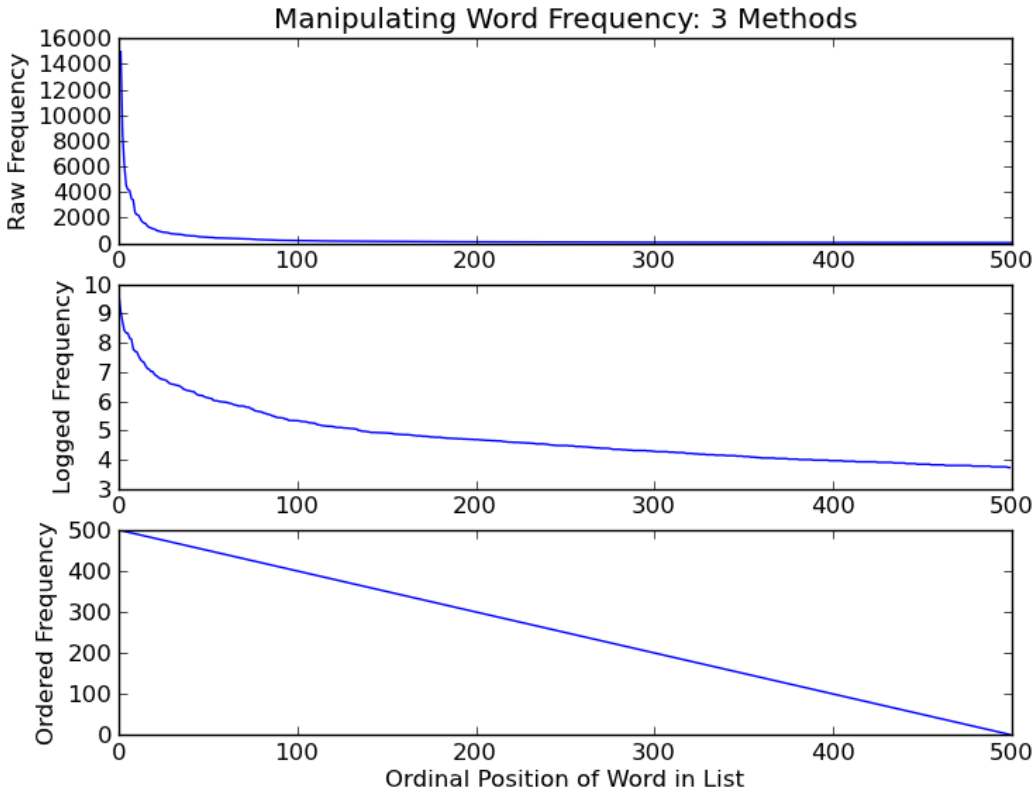

Figure 2.2: Manipulating word frequency distributions: 3 methods

The extreme curve produced by the raw frequencies of terms (first plot) is what we should expect from language data – this distribution accords with George Zipf's (1965) observation that a plot of frequencies from a sample of natural language will form a log-type distribution. Logging the frequencies (as represented in the second plot), then, does essentially two things: 1) it 'smoothes out' the curve, and 2) it reduces the magnitude of term frequencies. This has the effect of lessening the difference of a term's frequency with respect to its 'neighbour,' especially for the top ranked terms. (For example, in a list based on raw frequency, the 1st term occurs nearly ten times more than the tenth ranked term, though in an ordered list with logged frequencies the 1st term occurs only ~1.1 times the 10th.)

So, this metric is reasonably practical, and suggests a general principle: **the saliency of a term in a corpus is related to its collocability and frequency, where a term's salience is highest when it is frequent and narrowly associating and lowest when it is infrequent and commonly associating**. (Though, again, we do find that the lowest values are actually assigned to the *most* commonly occurring terms, where they also appear in a large number of repeating chunks.)[14]

Collocations are keyed to term salience, or, at least, bear a significant relationship. This is a subjective judgment, but these subjective judgments are what we are after: investigating the multiple ways frequency *means*, and how collocations constitute texts, we find that *objective* measures can approximate *subjective* experience. The results of this metric—as relatively coarse as they are—are immediate, and clear: rendering of a corpus with this metric separates function words from lexical items, and orders the lexical items in a manner that approximates subjective experiences of salience. The salient terms brought to the top, as well as the banal terms co-existing with the function words at the bottom, are just intuitively obvious enough—they just 'sound right' enough—to demonstrate the power collocations exhibit in natural language. To me, this also suggests we are still just scratching the surface of the possibilities that investigations into formulaic language offer. The next chapter will elaborate on these ideas of collocation and salience, and introduce Relevance Theory and bibliometrics as ways to

[14] Here, I must emphasize that the frequency I am using for the algorithm is the raw count for the number of times the term appears in the corpus, and not the number of times the term appears in a 'chunk.' Consequently, a term with a very high overall frequency, and which only appears once as part of one chunk, would still be ranked very high. (For example, if Piggy's occurs 1000 times, and only in the chunk Piggy's glasses three times, it would still be ranked very high.) Interestingly, though, this overall frequency ranking, taking individual instances along with their larger chunked counterparts, seems to be important to the metric. Using the raw count only from when the terms appear in chunks produces a result that is, again, heavily correlated with raw frequency and does not result in the dramatic separation observed above.

further develop the approximation of subjective experience using objective measures and, therefore, to resolve the polar conception of collocation.

3 Relevance Theory and Collocation

3.1 Widely, Weakly Resounding

> Generally and scientifically, echo has two coextensive histories: the mythological one and the scientific one.[46] Each provides a slightly different perspective on the inherent meaning of recurrence, especially when that repetition is imperfect.
>
> --Mark Danielewski, *House of Leaves*

Mark Danielewski's (2000) *House of Leaves* follows from—echoes, reiterates—Nabokov's (1962) *Pale Fire*, both works seemingly inflected through the Boswellian literary strain: a story about a story told through the obsession of a critic over the object criticized. Structurally, these two works are obviously related. While *Pale Fire* is a story about a made-up poem, Danielewski's novel adds another layer: *House of Leaves* is a story about a story about a made-up movie. To summarize the plot, Johnny Truant (who struggles to 'absent' himself from the nested narratives) discovers the academic work of a deceased neighbour, Zampano. Zampano's work is a fragmented, unfinished criticism of the movie *The Navidson Record*, which Truant realizes Zampano has entirely fabricated, along with his critical methods, scholarly observations, and academic-styled citations.

The sentences quoted above are from the Zampano-work and preface an exegetical passage on the made-up-movie's representations of space. Though formally they are academic (complete with a footnote, whose referent is fabricated), functionally they are fiction: by way of this generic inversion the non-scholarliness of the prose is made salient and the reader therefore has little expectation to learn about echoes, in a scientific or historical manner, but is motivated instead to expect a non-proximate context

to make 'echo' meaningful. That is, the purposeful display of form[15] gives the reader reason to extend the meaning of echo—though that meaning may not be immediately known—and re-orient other parts of the novel to make echo make sense. The 'meaning' therefore does include the proposed etymological history but also extends beyond it. At the very least, the mere mention of echo makes salient (recalls, echoes) certain parts of the story.

Of course, my own motive in invoking *Pale Fire* (1962) and *House of Leaves* (2000)—fiction about criticism and repetition—is not to offer criticisms of a genre but to make salient 'some other' part of my work: my work on collocation and linguistic echoes. There is a lot to dislike about *House of Leaves*, but I do appreciate one idea about echoes: an echo is not just the replication of a sound and the degree of likeness between the original sound and its double. An echo has a third part: by its reflection, it denotes the boundedness of a certain environment. We call this echolocation: the time it takes a soundwave to return to its source gives information on the space in which it was produced.

My point here is that an echo occupies a space, and if something occupies a space it has a shape – in other words, it has a form. A sound released returns, but it returns with additional information – it reveals the shape of its container, in a process known technically as 'echolocation.' This chapter will develop a technology to sketch the shape of collocations' echoes within their container, my corpus of *Lord of the Flies* criticisms. For this, we will need a form of echolocation, a way to see how utterances are generated,

[15] The footnote in the epigraph, for example, is a very formal way of denoting this generic inversion. Kenneth Burke (1923), on the subject of such inversions, in a short, obscure paper reviewing Gertrude Stein's work ("Engineering with Words"), suggests the term "fallacy of subtraction" to describe writing in which its content does not 'live up' to its form: the "full potentialities of [the artist's] medium", as Burke says, are not exploited. The artist "is getting an art by subtraction; he [sic] is violating his genre" (p. 410).

returned, and the space between them – a way to see the unsaid 'in the dark.' Relevance Theory, coupled with a theory from information science, will provide such a method. The goals in this chapter will be to propose a way forward in this investigation: an investigation of the internal variegation of collocations in *Lord of the Flies* criticisms—word associations echoing within the corpus—and what these collocations suggest about the shape of the criticism as a whole.

But I am not yet finished with *House of Leaves*. It's a weird book: the made-up movie, the object of Zampano's criticism, is a documentary-style film of a modern-day American family who moves into a paranormal house – a house with a door that leads into its impossible, labyrinthine insides. An expedition inside the house reveals megalithic chambers and moving hallways, which Zampano describes at length in his criticism (and for which he apologizes, because such description is 'un-academic'). These large, unlit rooms motivate the explorers' need for echolocation – and, of course, Zampano's need to discuss echoes in his critical work. The idea of 'echoes' becomes highly relevant in a poignant event in the movie, the lowest 'level' narrative, when the protagonist reaches the lowest level of the house.

The protagonist of the movie, David Navidson, after descending impossible distances into the bowels of the house, finds himself suspended in an unlit, no-gravity, expansive chamber. With nothing else to do, and an ingenious way of providing light, he decides to read. Navidson combines his last belongings—a book of matches and a book proper—and sets fire to each page of the novel to provide enough light, but rarely enough time, to read the next:

> Here then is one end: a final act of reading, a final act of consumption. And as the fire rapidly devours the paper, Navidson's eyes frantically sweep down over the text, keeping just ahead of the necessary immolation, until as he reaches the last few words, flames lick around his hands, ash peels off into the surrounding emptiness, and then as the fire retreats, dimming, its light suddenly spent, the book is gone leaving behind but invisible traces already dismantled in the dark. (p. 356)

I mentioned my ambivalence towards this novel, which is due partly to its endless, unfocused self-referentiality. Using terminology from Relevance Theory, which I will explain in this chapter, the "poetic effects" here are too varied and too weak. Almost any conclusion can be drawn, but with no evidence of *ostensiveness* – such conclusions' validity is only determined institutionally from without (through the institutionalization of criticism). That is, the critical possibilities from the passage above are numerous – so numerous and yet so determined that the echoes generated are diffuse and their texture dull. Any number of readings is possible:

1) In *Pale Fire* John Shade burns his rejected notes (disposing of them in a "pale fire"), while in *House of Leaves* Navidson burns text to give it life. But while there is a nice parallelism in this genre-bound thematic antinomy of 'birth and death,' in *both* works the text is consumed after being passed over by the critic's eye.
2) A parable of deconstruction: since Navidson is suspended in a limitless expanse, deprived of echoes, deprived of the ability for an utterance to be repeated (deprived of context), the "immolation" is "necessary" not to illuminate the text

but because the text itself is impossible: "already dismantled in the dark." In "Signature, Event, Context" (1988), Derrida's refutation of Speech Act Theory, Derrida claims that an utterance must be able to be repeated (echoed) since it is always a citation – a performative iteration. With no echoes, however, there can be no performance. Further, intertext in this echoless room is mutually destructive: remember that this scene involves *two* books (the novel as well as a match book), and their contact results in combustion. Finally, the contradiction inherent to this act is reinforced by the book Navidson reads: inexplicably, *House of Leaves* itself.

3) But the above reading does not address Navidson's subjectivity, which is realized by his constitution and mastery over the text. Navidson's final act before being consumed by the house is his consumption of the book, and it is only when the "*its light* [is] suddenly spent" that "the book is gone." And when the book disappears, so too does Navidson: it is therefore the absence of light, or losing the ability to make text intelligible (and not the fire), that obviates book and body (corpus and corpse).

4) Through its logical contradictions *House of Leaves* posits criticism as a primarily critic-centric aesthetic experience. Since *The Navidson Record* is a movie, descriptions are largely visual—and in many ways *believable*—but also impossible *to* observe: it is questionable, for example, how Zampano sees "invisible traces […] in the dark" in the movie. These impossibilities imply that the essence of criticism is not text, but 'the self' articulated through text – the

projection of an embodied 'sixth sense': the literary sense (ability) / sensibility.

I have delayed mentioning that Zampano, the movie's critic, is blind.

All of these criticisms are possible. Any can be inferred. There is evidence for each – but I advocate for none of them. What I want to emphasize here is that they *are* possible, but possible only in a limited way. These are all *types* of critical utterances: things others might say. Their validity comes from outside the text, in that they are socially or institutionally—rather than textually—determined: in short, these are echoes, and though their sound is diffuse it is also predictable.[16]

I have taken this detour to illustrate my first claim in this introduction: that echoes have a 'third part,' and that that part can tell us something about the size and shape of a body's boundaries – a corpus' form, for instance. *House of Leaves* is unbounded, and as such we can inflect it not so much with criticism, but with *voices of criticality*: the echoes of what can and must be said. So the criticism motivated by this passage is both very specific—determined institutionally—and very broad: one can select from a wide range of things to say about the novel, but each of those 'things' is limited. In sum: the field is rich, but superficially so. I delimited the above criticisms numerically only for readability, and this ordered list of exegeses belies the whirlwind of critical possibilities that was positively consuming—and therefore debilitating—to write.[17]

However, I would like to seriously take up one last idea from *House of Leaves* – the most basic, almost-literal narration of the above passage. To reiterate, Navidson

[16] In this case, illuminating (echolocating) a *social* and *institutional* rather than textual form.
[17] Which motivated my delayed revelation of Zampano's blindness, for example. The fictional critic's blindness is not an incidental fact, of course, and muddies up the rest of the interpretations. But since, again, there is an overwhelming (consuming) number of possibilities of criticisms it was necessary to delay to render intelligible. This is, therefore, a narrative rendering of an embodied experience: the proposal that criticism is *primarily* narrative is the point of the genre I ape in this thesis.

partakes in "a final act of reading, a final act of consumption" as he burns his book – and by consequence burns his fingers. At their most basic level, texts in this genre espouse the phenomenological danger of consumptive criticism: a defining feature is the critic-protagonist's utter consumption by the object under study, resulting in a tragic personal narrative that has really nothing to do with the text. But *House of Leaves* poignantly demonstrates the mutual consumption of critic and text. In this case, Navidson 'is consumed' but he also consumes his book – and in so doing, he burns himself.

So let's not do it. Let's not consume the text we criticize. All of the works I cite warn against it, but none suggest an alternative: consequently, these works criticize—if not the futility—the potentially destructive *banality* of criticism. But we have a method and a theory, corpus reading and Relevance Theory, and therefore a way to criticize text but not consume it. These can be used to explore the internal variegations of its collocations, but not descend into its bowels. Through corpus reading there is no reason to burn one's fingers.

3.2 Relevance Theory

As I described in my introduction, in 1987 Patrick Swinden panned a collection of *Lord of the Flies* criticisms, calling them 'banal.' His charge was one of 'sameness,' or repetition, in a body of texts – in general, a negative reaction to a perceived use of language. In the last chapter, on collocations, I claimed that although repetition is *one* measure of collocation—and an important one—collocations are sites of linguistic difference. Though there tends to be a fixation on frequency and semantic conventionalization in current collocation studies—a tendency to afford *one* chunk with *one* meaning (its most common)—I propose that this is not just incorrect, but detrimental

to the idea of collocation as a productive linguistic category. Further, in refiguring this distinction we might shed some light on the character of banality. In this chapter I will propose a way to test these hypotheses. Using Dan Sperber and Deirdre Wilson's (1986/1995) "Relevance Theory," with an information science adaptation proposed by Howard White (2007a, 2007b), I will show how we might gauge the relevance effects of the internal, relevance-oriented variegation of collocations, and also the relevance effects of collocations themselves in my corpus of *Lord of the Flies* (*LOTF)* criticisms. The next and final chapter will be devoted to testing these hypotheses.

Relevance Theory (Sperber and Wilson 1986/1995) is a context-sensitive, efficiency oriented pragmatic theory of communication. Following from Paul Grice's Conversational Maxims, Relevance Theory supposes that "the expectations of relevance raised by an utterance are precise enough, and predictable enough, to guide the hearer towards the speaker's meaning" (Sperber and Wilson, 2005a, p. 607). Although Relevance Theory has been around for a long time now (first formulated in the early 1980s), and has generated a very large amount of scholarship, it has been put to surprisingly little use – and no use in terms of corpus investigations. The amount of scholarship generated from this theory really is incredible: to appreciate just how much has been written, consult Francisco Yus's "Online Bibliographic Relevance Theory Service" (2011), a frequently-updated website that indexes published works on Relevance Theory. Yus's author index is nearly 500 pages long (in its native font), contains over 85,000 words, and *lists over 2800 items*. To put this in perspective, this collection is well over 50 times the size of the corpus for this study. So although I can't

hope to summarize the literature, I can start to summarize the theory – and try to stake some survey markers in the field.

As a theory of optimal efficiency, Relevance Theory (RT) proposes that speakers communicate, and listeners perceive, evidence as to what should be understood as optimally relevant in a communicative situation. For example, regarding *House of Leaves* when I claimed that the pseudo-academic form motivates readers to search for a more distal context to make sense of 'echo' in the novel, this (unconscious) search for meaning is guided by a search for Relevance, and the process is set in motion by the speaker—in this case an author—making salient aspects of a shared cognitive environment. In Relevance Theory terminology, the concept of Relevance is considered a measure of efficiency, an expectation of maximum gain for minimum input. Sperber and Wilson (1986/1995), in the introduction to their seminal *Relevance: Communication and Cognition*, describe human cognitive and communicative capacities as "geared to achieving the greatest possible cognitive effect for the smallest possible processing effort" (p. vii). Relevance Theory defines cognitive effect as "a worthwhile difference to the individual's representation of the world: a true conclusion, for example" (p. 608). Relevance, in RT, therefore enjoys a very specialized sense,[18] and applies to practical examples (dialogues, etc.) and literature alike – though I have departed from standard glosses of the theory in selecting, as my first example, the former rather than the latter.

Typical explanations of Sperber and Wilson's theory have tended to claim it as an escape from the 'code theory' of communication, in which an utterance is encoded with

[18] And certainly not a lay sense, as has sometimes been assumed. Sperber and Wilson are quite clear about this: "there is no reason to think that a proper semantic analysis of the English word 'relevance' would also characterise a concept of scientific psychology" (*Relevance*, p. 119).

meaning, communicated, then decoded in a parallel process. This typical explanation claims that Relevance Theory advocates for an inferential (pragmatic) over coding (semantic) theory of language. However, although Sperber and Wilson do reject the code theory as a totalizing explanation *of* communication, they hardly escape from it as a theory. The technical aspects of communication, they claim, are many: using the analogy of transportation they show how it would be inconceivable to use one technology of transport as a theory to explain all others – locomotion as a general theory that explains bicycling, for example (1986/1995 pp. 2-3). Sperber and Wilson therefore do not deny encoding/decoding as a productive process in communication, since parts of their theory rely on it, but maintain that it is insufficient to explain communication in general. (Incidentally, in this work I do not claim that collocations *cannot* be semantically conventionalized, just that this is an insufficient, secondary, and restrictive explanation for their meanings.)

Likewise, however, it would be incorrect to claim that RT is not at least somewhat dismissive of a coded mode of language, and favourable to inferential modes. So although a code model can work with inference models this is not to say that it enjoys equal 'privilege' in their theory. Further, different parts of their theory use different modes: generally (and coarsely), implicatures are inferred while explicatures are decoded. The majority of their work is devoted to expounding upon Grice's theories of inference. RT claims, essentially, that explicit communication is both *enriched* and *restricted* by inferential information: stimuli that 'point towards'—or give evidence for—the speaker's intended meaning. Since linguistic meaning is underdetermined in language—an utterance cannot *explicitly* convey all of the information required to make

it meaningful—implicatures are required to supplement and refine the utterance, where implicature are unstated, but deducible, assumptions that are spontaneously generated within the context of the situation.[19] Sperber and Wilson define implicatures as: "a contextual assumption or implication which a speaker, intending her utterance to be manifestly relevant, manifestly intended to make manifest to the hearer" (*Relevance*, p. 194). These implicatures generate new ideas, or 'assumptions relevant in a context,' and thus augment the semantically underdetermined language such that meaning can be generated. The product is an efficient positive-feedback generation of cognitive effects, a ratcheting-type action.

Related to the implicature is the explicature. To explain it simply, and avoid the pitfalls in making an explicit/implicit distinction, I will define an explicature as a statement that generates fewer implicatures: the refinement of an implicature, an implicature in its logical form. But while explicature might sound quite dull (deadened by what might be called non-poetic, denotative language) they are actually quite vibrant because they, too, 'enrich' the utterance and generate contextual effects. Robyn Carston (2002) explains the distinction as dependent on the degree to which meaning is encoded versus implied by the utterance, where the meaning of an explicature is encoded while an implicature's meaning must be inferred: "the conceptual content of an implicature is supplied wholly by pragmatic inference while the conceptual content of an explicature is an amalgam of decoded linguistic meaning and pragmatically inferred meaning" (p. 134).

So, what we can say about Relevance Theory, in terms of other theories, is that RT theorizes a *model* of communication whereas others (communication by way of

[19] 'Implicature' is actually a fairly broad category and can be further subdivided into categories of implicated premises and implicated conclusions based on a deductive logic. However, I will not consider these subcategories.

inference, or code) theorize *modes*. This division subsumes technologies of communication (modes) under a fundamental principle (that can be modeled). Relevance Theory, therefore, is totalizing: it can accommodate communicative modes under the principal of Relevance. Thus, if coded communication works it is guided by Relevance (and, indeed, is motivated in the first place by Relevance). But, if efficiency is in fact the guiding principle of communication and cognition, a purely coded mode (or any other single modality) would not make sense: having to assign one code and one key to every single utterance would be very inefficient, even for commonly used phrases (such as collocations). Efficiency, then, is the underlying principle of RT.[20]

The principle of efficiency is supported by RT's novel notion of context, where context is not pre-established and fixed but dynamic and extendable. In this way, context can be expanded to generate the maximum assumptions for minimal processing. Francisco Yus (1998) summarizes this point well, referencing a work by Sperber and Wilson that predates their comprehensive—1986— proposal of RT:

> S&W (1982a) reject the picture of context [as] a monolithic entity that is accessible to interlocutors beforehand during interaction. Instead, they propose a much more dynamic view of context as a construct that has to be established and

[20] And also a key point of contention. Kent Bach (2006), for example, points out that a problem with RT is that there is no viable way to measure efficiency: "The most obvious problem is that of how to quantify and to measure degrees of cognitive effects and degrees of processing effort. The formulations I've seen of relevance-theoretic concepts and principles are too vague to be of much help in this regard" (p. 7). This is a fair critique, but I think the plausibility of this quantification depends on where it takes place: in the mind, which likely would be impossible, or system-side – a predictive type of measure. But further, it might not even be necessary to quantify Relevance as a precise measure: my method, for example, evaluates cognitive effects and processing effort on the system side, and as a *relative* value. In this way, collocates are not given a definite value but estimated as gradients in relation to one another. As Howard White (2007a) writes, "S&W's relevance is not a matter of *yes* or *no* but of *more* or *less*" (p. 538, emphasis in original).

> developed in the course of interaction in order to select the correct interpretation (p. 307)

Rather than proposing that speakers seek, and listeners receive, a message's meaning in terms of an established context, interlocutors seek the appropriate context to make an utterance meaningful and it is this search, resulting in 'interpretation,' that produces contextual (cognitive) effects. So, on the one hand RT's 'context' *is* monolithic since this one term encompasses many determinative variables, including such abstract concepts as 'history' and 'culture.' But, on the other hand, Sperber and Wilson's conception of context is quite radical because this monolith is continually changing its shape, extending and retracting, in response to the communicative environment.

However, this concept of context has been a particular point of contention in RT scholarship, with criticisms generally falling along two lines: questioning *what* context encompasses, and *how* context is constructed. For example, Andrew Goatly (1994), in his frequently-cited "Register and the Redemption of Relevance Theory" proposes, essentially, that Sperber and Wilson's conception of context is too narrow because it does not include aspects of sociality: Goatly argues that Grice's theories and RT "are flawed through their failure to consider cultural and social context" (p. 139). Goatly argues that genre and register, specifically, cannot be divested of their communicative significance. On this contention, I completely agree: genre cannot be discounted in communication. However, I do disagree with Goatly's argument that RT does not accommodate 'the social.' A quick response to this criticism would be that RT does include 'the cultural' and 'the social' just fine – these are included in 'the context.' But, to be fair, the problem

is more complex than the widely encompassing structure I proposed previously, and does illustrate one of the central concepts of Relevance Theory: ostensive action.

Essentially, Goatly (1994) claims that Sperber and Wilson deny social-historical variables in their theory. Using examples from *Relevance* (1986/1995), Goatly suggests that features such as phonology and genre work powerfully—and often primarily—as a means for interpreting an utterance. For example, the phonologic manner by which an utterance is produced (with rising / falling tones), or the form in which an utterance is produced (in terms of genre[21]), can be sufficient for an interlocutor to retrieve meaning. However, by suggesting that RT does not account for these things, Goatly seems to suggest that such social elements are independent from language: Goatly refers to such elements as 'meta-linguistic' and not properly linguistic. My understanding of RT, however, is 1): no hard line between meta-linguistic and linguistic acts is drawn; and 2) the key to the creation of context lies not in such coded categories but in ostensiveness: with 'ostensive behaviour' referring to intentional actions that point (in an abstract sense) to a feature (or idea / concept) in the shared cognitive environment. Sperber and Wilson's (1986/1995) book is full of ostensive action: ostensive sniffing, ostensive leaning, ostensive yawning – where, in each case, it is the ostensive act that makes salient a feature in the shared potential cognitive environment to guide interlocutors to select the correct context to interpret an utterance. And if one can lean / sniff / yawn ostensively then it seems reasonable that phonology can be an ostensive act as well.[22] Genres (as I

[21] Andrew Goatly seems to take a semantic, 'coded' approach to genre: a certain form, in a certain situation, equates to a certain genre. Further, this genre can be inscribed with meaning.

[22] Which raises the question: what range of speech features *can* be ostensive? For example, an accent obviously communicates a large amount of information—in terms of identity—but cannot properly be called ostensive because the speaker cannot (always, or fully) intend for their accent to have such a function. To my knowledge, *this* question remains unanswered by RT, and is a currently relevant topic for sociolinguists.

conceive of them) are dynamic entities, since they might arise independently through regularized social interaction (Miller 1984), but as they are defined by Goatly—as coded social / communicative interactions—they retain a certain ostensiveness. Consequently, RT accommodates social, historical, experiential, elements just fine: they enter the conversation just as everything else does – this is to say, *linguistically*. To summarize, I would say that it is not the theory of context that is too restrictive but Goatly's conception of what counts as ostensive pragmatic interaction.

And it is this concept of ostensive action and mutual manifestness that leads to the second common criticism of RT's 'context': *how* context is selected, and *what* is mutually manifest. RT replaces mutual knowledge—the knowledge parties know *is known* between them—with 'manifest assumptions': assumptions in a potential cognitive environment that, essentially, become highlighted to provide the ostensive stimulus for communication. Critics, however, claim that this distinction Sperber and Wilson propose is minimal. Yus (1998) sums up this criticism in a paraphrase: "how do speakers distinguish the information they really *know* from that they really *share*?" (p. 310, emphasis in original). Essentially, these criticisms claim 'mutual knowledge' is no different from Sperber and Wilson's 'cognitive environments.'

However, I think that this criticism does not fully embrace Sperber and Wilson's innovation: the complaints are aimed at a knowledge beyond a reasonable doubt, while Sperber and Wilson's improvement on establishing context essentially *lowers* 'the burden of proof.' Mey and Talbot (1988), also quoted in Yus, write: "Cognitive environment is in principle not distinguishable from mutual knowledge, as long as it is supposed to have some such intersubjective 'reality'" (p. 250). But 'reality' does not necessarily enter the

equation. Shared cognitive environments can be fictional – such as in literary fiction. Studies in motivated visual perception (such as the famed 'gestalt switch'), as well, show that people can be motivated to perceive 'real' objects in the 'real' world – that do not *really* exist. Thus, this criticism does seem to hold on too tightly to the old theory of mutual knowledge, and confuse it with Sperber and Wilson's reformulation.

This reformulation is one of Relevance Theory's most important contributions: not just showing that context is flexible and dynamic, but also how it 'updates' theories concerning intersubjective awareness *of* this context. Although RT does employ a type of 'mutual knowledge' (generally, knowledge shared between people in an environment), Sperber and Wilson replace this term with 'mutually manifest assumptions' in shared 'cognitive environments.' The problem with making 'mutual knowledge' work as a pragmatic term, Sperber and Wilson contend, is that in addition to interlocutors holding the same assumptions—sharing mutual knowledge—they must *also* hold the assumption that this knowledge is shared. Moreover, since they affirm Grice's claim that 'at least part of the meaning of the utterance must be recovered from the intention of the speaker that that meaning be recovered' (*Relevance* p. 53), for mutual knowledge to work it must always be accompanied by the additional knowledge that this knowledge is mutual. All of this leads to a problem of recursion—the n + 1 dilemma—because then the mutuality of the mutual knowledge must also be known, and so on.

All of this is replaced, as I have noted, with a lower standard of proof. Instead of agreeing upon shared knowledge (what interlocutors 'really' share), RT states that implicatures are generated when something is 'made manifest' (or 'more manifest') in a cognitive environment. And here, 'cognitive environment' maintains a dialectical

relationship with its manifestness: from the moment an utterance is produced, it generates and draws upon resources from the speaker's current environment, encyclopaedic knowledge, previous dialogues, and intentions. This environment is never guaranteed, but contingent. (Indeed, if it was guaranteed, a coded theory of communication could be a viable model.) At this point, to illustrate concepts let's consider an example:

> From some time ago, I recall a sign in a fitness facility that read: "PLEASE KEEP WEIGHTS AWAY FROM MIRRORS. MIRRORS ARE EXPENSIVE AND DIFFICULT TO REPLACE."

The sign's most basic meaning is clear: 'don't break the facility's mirrors,' even though this is not stated explicitly. This may be derived from the following implicatures (or a coarse abstraction of the following, omitting obvious entailments):

a) People who lift weights engage in activities that are strenuous.

b) Some strenuous activities are vigorous.

c) Vigorous activities can result in accidental contact with nearby objects.

d) Vigorous activities can break nearby mirrors.

e) Breaking mirrors would incur costs for the centre.

f) It would be mutually detrimental to incur costs for the centre.

g) It would be detrimental to risk breaking the facility's mirrors.

The message, therefore, is a proscription on a *type* of activity (presumably reckless exercise) that might damage the facility, with the sign's author intimating a kind of risk management. Again, the most basic meaning generated here is the imperative, "DON'T BREAK THE MIRRORS", drawing on the reader's "encyclopaedic knowledge" of mirrors as fragile objects and weightlifting as a strenuous activity. The

implicatures enriching this statement construct a context of risk management (pointing, again, toward types of physical activities, and a consideration of the benefit of that activity versus the monetary expense it could reasonably incur). Moreover, these assumptions hold true in a context where it is, specifically, a human activity that might break the glass. The "WEIGHTS," inanimate objects incapable of damage on their own, function metonymically to draw attention to activities *using* weights rather than the weights themselves. Therefore, eliciting the notion of weight-use constructs a context with a high probability of being mutually manifest to users of the facility – a context that includes a sense of shared responsibility and physical activity. (Consider another advisory on 'dangerous objects,' a syntactically identical sign that motivates entirely different implicatures: at a zoo, for example, where a sign warns parents to "PLEASE KEEP CHILDREN AWAY FROM ENCLOSURE EDGE".[23]) Relevance Theory predicts that all of the assumptions I have just made explicit (in addition to many other assumptions, more weakly implicated) are processed unconsciously, giving rise to a range of contextual effects from this rather simple sign.

Furthermore, this example also shows how Mutual Knowledge is an unusable term. In this case, for Mutual Knowledge to be a functional concept the sign's author would have to know that the facility-user knows (or can imagine) these proscribed exercises, *and* the further knowledge that this is the intended meaning of the sign. The reader would have to know this, and know that the writer intended this, etc. So while Mutual Knowledge necessitates a never ending cycle, in Sperber and Wilson's formulation the line is drawn much sooner: in RT, interpretation ends at the *first* available conclusion. The first available conclusion that meets the expectation of

[23] "Children are expensive and difficult to replace."

relevance is the most Relevant – the first is the most efficient: the most gain for least effort. While we can certainly investigate weaker implications of the sign's meaning, we must also respect that these implications might not be recovered in the context of their generation. (These weaker implications, as I referred to them regarding *House of Leaves*, are termed 'poetic effects.')

Moreover, there is a difference between *implicature* and *inference*: where implicature is an inference guided by ostensification. As examples, I would say that the criticisms I presented in the introduction are *inferred* rather than *implicated*, hence their breadth but limited depth: there is evidence for these interpretations, but this evidence comes almost exclusively from outside the novel and therefore lacks the guiding 'consciousness' of the author. This also highlights the difference between *ostension* and *intention*: Danielewski very well could have intended for all of these conclusions to be drawn—he could have intended for this entire discussion to take place—but there is little evidence of this. The breadth is too wide, the potential for inferred criticisms are too numerous, for too little 'cognitive effect.'

Thus, in Sperber and Wilson's formulation, the primary focus of consciousness is not towards the environment (or 'context') but to ostensiveness, the evidence for interpretation: Relevance Theory is, therefore, a *primarily* intersubjective theory of communication in that the central mechanism is estimating others' consciousnesses. In this way, the answer to the question, 'How do we know what other people are paying attention to?' is: we don't. But we do have evidence of others' attention, if only weak, [24] and this weak evidence generates assumptions based on optimal relevance. The main

[24] The evidence is weak, and likely 'projected' from the self. That is, a speaker estimates another's consciousness at least partly based on his or her own.

activity in communication is not evaluating something in terms of the environment, but an estimation of another's perception of the environment.

Finally, this construction of context can 'expand' or 'retract' in a movement guided by efficiency. Returning for a final time to the fitness facility example, as I mentioned, the proscription is not regarding an object's proximity to the mirrors, but an activity motivated by that object. In this case, then, the context is *narrowed* to one we might broadly describe as 'activities that might take place in a fitness facility, but should not.' In addition, since the sign makes salient a *type* of activity, the context is also *expanded* beyond activities that are limited to weights: the intent of the sign clearly proscribes any unsafe-for-mirrors activity. Many other, weaker implicatures, might also reasonably be generated—pertaining to broad spectrums of communication like rhetoric, ideology, etc.—but are likely not generated in its context of use unless they are made manifest by an additional stimulus. For example, eliciting the implicature of 'mutual economic loss' and 'communal altruism,' one gym user could chastise another's possibly destructive activities by saying: "Hey, I don't want my fees to go up!."

The concepts explained above explicate most of Sperber and Wilson's theory, or, at least, a subset of the concepts relevant to my work on collocation. As a concluding note, a funny thing about RT literature is that it has become so expansive it seems necessary to select such a subset of terms, and this trend of subset selection persists in its expansive literature. As I mentioned, since Sperber and Wilson proposed this theory nearly twenty-five years ago, it has had massive uptake in, and has been refined through a variety of disciplines: from translation theory (Mateo 2009; Dooley 2008; Gutt 1990) to poetics (Pilkington 2000; Uchida 1998), and—of course—pragmatics (Carston 2002;

Blakemore and Carston 2005; Sperber and Wilson 2005a, 2005b, 2002). And although each discipline tends to extract and privilege a certain theoretical 'thread,' two things unite these sometimes disparate works: 1) across the board, little progress has been made in developing a viable *method* for using relevance-theoretic principles in corpus investigations[25]; and 2) these works do tend to promote one key idea: that RT is a theory of communicative efficiency, however articulated, and this theoretical unity has produced a formalism – the formula Relevance = Positive Cognitive Effects / Processing Effort. My work coincides with both of these trends, in that I seek to propose an application using this formalism.

3.3 Concluding Remarks on Relevance Theory

So that is an overview. There are some problems with Relevance Theory, which I have glossed, and though I have played counter to the criticisms I do not pretend to have solved them. There will likely remain a number of unsolved problems. Relevance Theory *is* imperfect, but I think its greatest problem is *not* its logical flaws but its lack of use. Very simply, RT has not really been put to work. I propose that so much scholarship (~2800 publications) has been generated, and so many problems have been identified, partly because of its lack of application: an idle theory makes theorists grow restless. So, if the greatest problem with RT is its lack of use, we might fix it by simply using it.

There are likely many reasons for RT's lack of use, not the least of which is the totalizing nature of its argument. Indeed, if it is a good description of communication and cognition it would necessarily be hard to observe since examples of it would be

[25] Not that we would expect such an application to be developed in the area of translation theory, but even there such explicit, precise methodologies tend to be undervalued.

everywhere. (It would be banal to say 'this or that' is an example of Relevance, since—as RT states—our cognitive faculties presume Relevance.) And in cases where RT has been used—in poetics, for example (Pilkington 2000; Uchida 1998)—its application has been functional rather than predictive: RT explains not only which implicature, from a number of possible implicatures, might be selected in interpreting an utterance, but also *how* this is possible.[26]

I think that the central problem with applying Relevance Theory is that its *terms* are difficult to apply.[27] And RT does not suffer from a lack of terminology – indeed, working with this theory and keeping all of its terms straight can be quite burdensome. (Of course, it does not help that many terms changed meaning between the first and second edition of Sperber and Wilson's [1986/1995] book: 'positive cognitive effects' is the new term for 'contextual effects,' for example.) I would even suggest that the aporia Goatly (1994) notes regarding Relevance might be resolved by a more careful fitment of its threads – all of RT's 'parts' must be finely aligned to properly mesh together. To summarize, if RT's inapplicability lies in its terms, and we want to make RT work, what we need is a way to map RT terminology onto an already functional method – a method of predicting the Relevance effect between two variables.

[26] On this note, in humanities research I have long thought that the term 'how' is often used in place of 'that,' and this effect is particularly pernicious in applications of theory. For example, a work might claim to be 'a Foucauldian examination of *how* power is articulated through social formations' and reduce to a series of claims *that* this phenomenon is observable along with observations regarding that phenomenon. Of course a phenomenon fitting a theory exists: this is what motivates theory in the first place. Allen Thiher (1997), in *The Power of Tautology*, makes related (though polemical, and sometimes specious) claims regarding literary theory and tautological reasoning. If what I am identifying does in fact exist, I also speculate that it was the 'functionalisms' of the middle of the century that supports this semantic phenomenon.

[27] Here, I am recalling a conversation with Janet Giltrow about Michael Volek's work on Relevance Theory and terms. Although the sense of the conversation was different, this did get me thinking about RT and its terminology.

3.4 Relevance Theory and Information Science

The best application of Relevance Theory, I think, does precisely this. In 2007, Howard White published two articles in which he reframes RT's central variables in terms of information science. His research makes a case for RT's application in information retrieval systems (in other words, library catalogues and the like), and—in fact—explains that these systems and the theories underlying them have been modeled with relevance in mind, and therefore its terms are directly mappable onto Relevance Theory's. White applies his formulation to a library catalogue search, where he shows the predicted Relevance of a user entered search term (the title of a book, or an author, etc.) to other works, authors, and genres.

White describes how the variables in RT are equivalent to the variables in a common bibliometric term weighting formula, the tf * idf formula, for gauging the relevance of documents to a query. In this formula, *tf* stands for 'term frequency,' while *idf* is 'inverse document frequency.' Term frequency refers to the raw count of tokens in a collection of documents, while document frequency is the number of documents in which a term appears. (In terms of document frequency, the raw count of terms does not matter. It is simply a binary count – whether the term appears at all, no matter how frequent.)

In information retrieval systems, the idea behind the tf*idf weighting formula is that these two variables can be used to gauge the relevance of documents in a collection to a seed term (a user-provided search term, for example). A document will be highly ranked if: 1) the seed term occurs in that document many times (its frequency is high), and 2) the seed term occurs in few documents in the collection as a whole (its document

frequency is low). Because *idf* is an inverse measure, a low document frequency will produce a high *inverse* document frequency, and therefore a high tf*idf score (because the two terms are multiplied together). White's work is innovative in two ways.

Howard White's first innovation is mapping these bibliometric variables directly onto Relevance Theory terminology, the central variables of Relevance expressed in the formula Relevance = Cognitive Effects / Processing Effort.[28] White claims that *tf* is equivalent to positive cognitive effects and *idf* is equivalent to processing effort. The assumption behind this translation is that a recurring term has the potential to produce greater cognitive effects—it essentially has more significance in the data set—but if a term occurs across many documents it becomes "semantically unfocused" (Roberts qtd. in White), essentially harder to relate it to the seed term in a precisely defined sense: "The logic of idf is that the more frequently a term appears across a collection of documents, the less 'semantic focus' it has and the less good it is at differentiating them" (White, 2007a, p. 540).

Of course, this type of statistical weighting is not new in information science, and, as White points out, algorithms based on the concept have been in use for quite some time. However, it is precisely *because* of the ubiquity of this algorithm that White's finding is important: information scientists are particularly attuned to the idea of Relevance, so the development of certain aspects of Relevance Theory in another field seems to be a corroboration of the theory. Indeed, information science has been able to apply this derivation for many decades (White mentions that the Bradford distribution, a Relevance sensitive metric, originated in S.C. Bradford's work from the 1950s).

[28] Where, again, Relevance will be greatest when cognitive effects are very high and processing effort is very low.

White's second innovation is leaving *tf* and *idf* unmultiplied, and plotting these values along two axes in what he calls a 'pennant diagram.' 'Predicted cognitive effects,' or *tf*, is plotted along the x-axis while 'predicted ease of processing,' *idf*, is plotted on the y-axis. The resulting plot, the 'pennant,' is roughly triangular in shape and diverges from a central point: the seed term. The pennant shape results from two statistical measures: raw frequency (predicted processing effort) and document frequency (predicted ease of processing). Figure 3.1 is a reproduction of one of White's diagrams. This diagram plots works co-cited with Herman Melville's (1851) *Moby Dick* (*Moby Dick* is the seed term), where each 'dot' represents a work. Works located along the x-axis are ranked according to how often they are cited with *Moby Dick*. White's description of the y-axis value is fuzzy, though it appears to be determined by the number of times the works are cited overall: "The corresponding document frequencies—the *df*s—are the counts of these terms in the collection whether they are [cited] with the seed term or not" (p. 540).

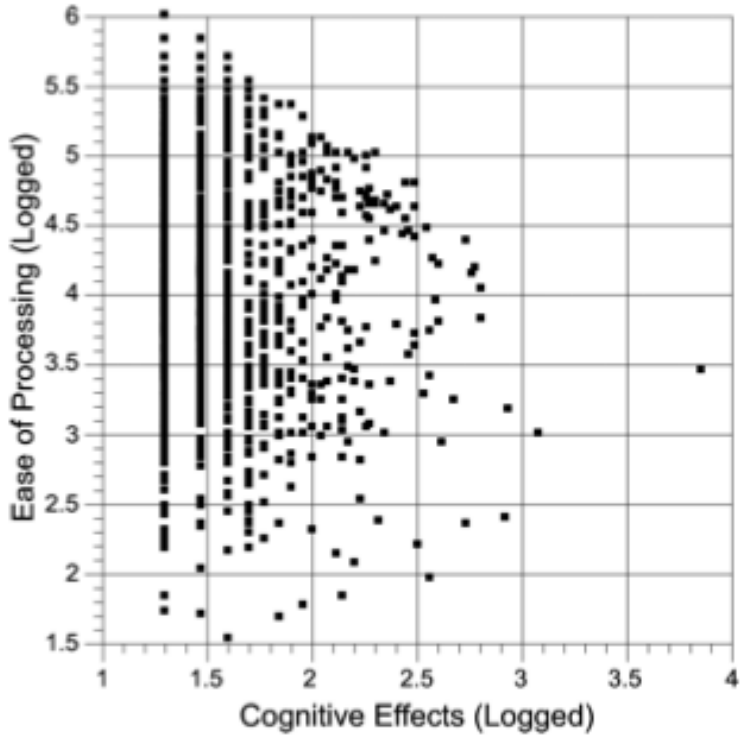

Figure 3.1: Howard White's (2007a) example pennant diagram[29]

Although conceptually simple, this second innovation—leaving the terms unmultiplied and plotting the independent values, term frequencies and document frequencies, on two axes—is quite powerful because this multi-dimensional approach allows for multiple ways of approaching the data. First, this plot dislocates relevance from simple frequency. By plotting values along two axes Relevance is not just determined by one variable. (Indeed, this resonates well with an argument against conventionalizing collocation's meaning, since both resist *consolidation*: consolidation of a collocation's *meaning* and consolidation of cognitive processes.) Second, it allows for an investigation of variegation – a relative rather than quantitative measure of a term's

[29] This figure is reproduced, with permission, under license from John Wiley and Sons: license number 2739460917253.
White, H. D. (2007a). Combining Bibliometrics, Information Retrieval, and Relevance Theory, Part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology, 58*(4): 536-559. FIG. 1. Pennant diagram for Moby Dick studies (p. 541).

collocates in the corpus. (In my upcoming examples you will note that that the axes are labeled numerically – this is an aesthetic decision to aid in comparing the figures and does not mean the numbers mean anything on their own.) Finally, this plot is predictive because it approximates these variables on the 'system-side' of a human-system interaction.

The attraction of this application is that it may be adapted for use with collocations. White's pennant diagrams plot works co-cited with other works, while collocations might be considered words co-cited with other words. My adaptation of this translation takes a seed term as its input and plots the term's collocates, which results in a similarly shaped pennant configuration. This plot shows the variety of ways one term is relevant to another—expressed as two variables—as well as the overall *shape* of the field. If collocations are a kind of echo, and these theories comprise a type of SONAR, then these charts are sonograms.

3.5 A Further Translation: From Bibliometrics to Collocations

To implement this method I developed a computer program with the open-source programming language *Python* (v. 2.7.2). *Python* is a versatile programming language with unique data types, making it particularly amenable for working with words.[30]

[30] Unfortunately, although *Python* is a great tool for the study of English a great irony accompanies its use. Historically, a large hurdle for the Digital Humanities has been one of access. Both access to information and access to text-tools have suffered due to the isolated nature of those involved in the discipline (there is no 'proper place' for Digital Humanists in academic institutions), and a poor or reluctant distribution of the discipline's products. (For example, in some instances I have expended a significant amount of time simply reproducing others' work in my own programs to test conclusions.) And although one of *Python*'s features is its relative ease of use, this ease comes at a price: *Python* programs cannot easily be shared between computers. The reason is technical: *Python* programs are scripts interpreted in a 'shell' at runtime and do not compile into executables [.EXE] or disk images [.DMG]. Consequently, they are not 'mobile' between computers, unless that computer has an identical or similar installation of *Python* to the program's source machine. All of this would be comical if wasn't such a shame: the programming language adopted by a group hindered by insularity and poor distribution adopts a tool that encourages isolationism and restricts the ability to share.

Further, an automated method was important to this Relevance-sensitive technique, since it affords iterative experimentation, as I will illustrate below.

Adapting White's application to work with collocation was not a straightforward endeavor. Even though only two variables are used, and one of these—cognitive effects—is just plotted using raw frequency, it was not immediately apparent what measure would approximate the processing effort involved in relating one term in a collocation to another. My first attempts tried to use *idf* in a straightforward, 'by the book' manner: a count of the number of documents in which the collocation appears. This failed to produce a useful distribution, however, since the document frequency was too correlated with raw frequency. That is, the more frequent the collocation the greater number of documents in which it appeared.[31] The following, Figure 3.2, is a plot of the collocates of "human," derived from our prototypical collocation "human nature." Raw frequency (logged) of each collocation is plotted along the x-axis, while document frequency is plotted along the y-axis:

[31] This is noteworthy, in itself, for two reasons: first, it is not necessary that this should be the case; and second, although there is a close correlation there is still a degree of variation. Unfortunately, these questions and implications cannot be taken up here.

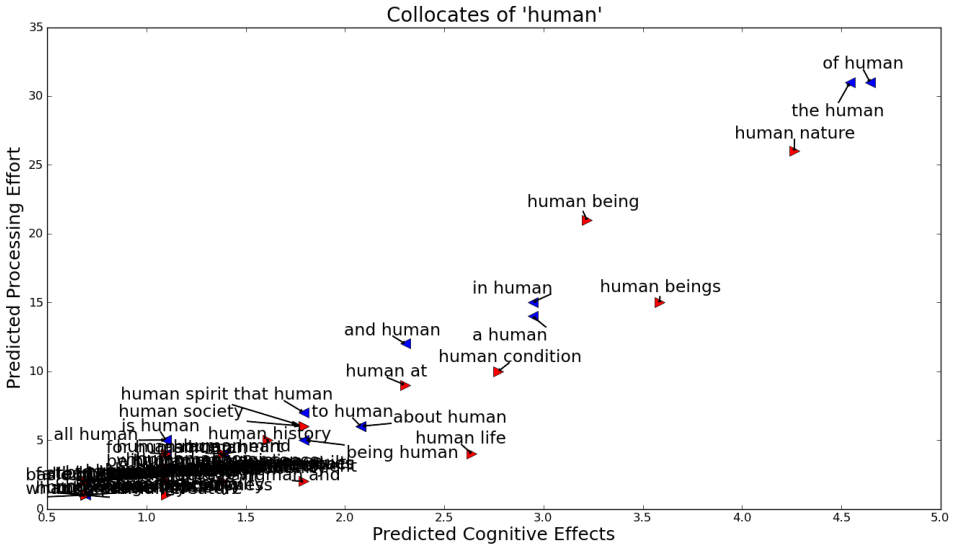

Figure 3.2: A first attempt at plotting the Relevance of collocations

All of the collocates of "human" that occur in the corpus at least twice are depicted in this chart by a triangle and a label. 1-R collocates are represented by red triangles (**red** for **right**), pointing right, and 1-L collocates as blue triangles pointing left. From this chart, we see that the most frequent collocates of "human" are both 1-L, 'the' and 'of,' but as the collocates become less frequent (as we move from right to left) the document frequencies also decrease in a linear, highly correlated regression (the slight curve is induced because the frequency count is logged). This was not a desired result, given that 'predicted ease of processing' is essentially saying the same thing as 'predicted cognitive effects' – both change together, while a pennant diagram displays the multi-faceted ways in which terms can be relevant.

So after giving it some thought, and evaluating this visualization, I realized that there are two problems with this approach. First, there is no reason to assume document frequency, in terms of collocation, is related to 'semantic specificity' or processing effort. In this case, not only does document frequency maintain a direct correlation with

frequency, but document frequency means something altogether different than 'semantic specificity'): a collocation appearing in a large or small number of documents *might* also indicate some kind of semantic specificity, but more likely it is an indicator of some other feature such as authorial style. Second, calculating the document frequency of a collocation does not seem to capture its internal relevance—the relevance of one term to another—but the relevance of the collocation to the corpus as a whole. Presumably, these problems arose from a naïve transliteration of 'co-citation' of works to 'co-location' of terms: there is no reason to assume they will work in the same way.

My solution to these problems was inspired by work I presented in the second chapter: my metric of term saliency, which is largely dependent on the number of different chunks in which a term appears. Further, I preserve the idea of a 'top-level' chunk (a multi-gram that is not a constituent element of any other multi-gram), where only these top-level chunks are counted. Essentially, this metric of term saliency is based on a term's varying proclivity to associate in groups – a term's 'focus' as measured in the chunks in which it is found. This metric produces a distribution that separates highly salient terms, terms that occur frequently but in few chunks, from highly productive terms such as function words: words that associate—still very frequently—but in far more chunks. Although this metric could not be 'mapped' straight across onto the method I developed I did experiment with it,[32] and it did get me thinking about the constituent elements of chunks.

Consequently, my solution for estimating processing effort is the following metric: the frequency of the collocation divided by the total number of the 'top level'

[32] The most basic reason for this is that the metric I proposed in chapter two is better suited to individual terms than to multi-word units.

collocations[33] in which the seed term's collocate is a part. The value will therefore be highest when the frequency of the collocation is very high ("human nature" appears many times in the corpus) and the collocate forms few collocations ("nature" combines with few other words). The idea behind this measure is that it identifies how variable a collocate is—how ready it is to 'hookup' with another term—relative to its collocation. For example, the following chart shows the top five collocates of "human," and how many *other* collocations those collocates form:

	Collocate	**Frequency**	**Number of Chunks**
1	***of*** human	104	2666
2	***the*** human	94	4575
3	human ***nature***	71	79
4	human ***beings***	36	10
5	human ***being***	25	29

Table 3.1: Top five collocates of "human"

So although "of human" has the highest frequency, "of" also combines with 2666 other chunks. "human nature," on the other hand, also has a relatively high frequency but "nature" combines with far fewer chunks. I propose that this captures the idea of semantic specificity, and therefore processing effort: "nature" is more *spcecific* than "of" to "human." "nature" occurs with "human" fewer times than does "of," but also occurs in fewer chunks overall. "beings" is even more specific, occurring in far fewer chunks and only minimally less frequent. The metric I propose, then, results from frequency and collocability of a collocation. Using this metric to gauge 'ease of processing,' I produced the following chart[34] (figure 3.4). All collocations are plotted that recur at least twice (i.e. raw frequency > 1) in the corpus. Raw frequency (logged) of each collocation is

[33] From chapter two: what I call 'top level' collocations are collocations that are not constituent elements of other collocations.

[34] An extra library for *Python*, called *Matplotlib*, allows for the programmatic generation of these charts. *Matplotlib* extends the abilities of *Python* to add better graph-plotting functionality.

plotted along the x-axis, while the values resulting from the metric explained above (also logged) are plotted along the y-axis:

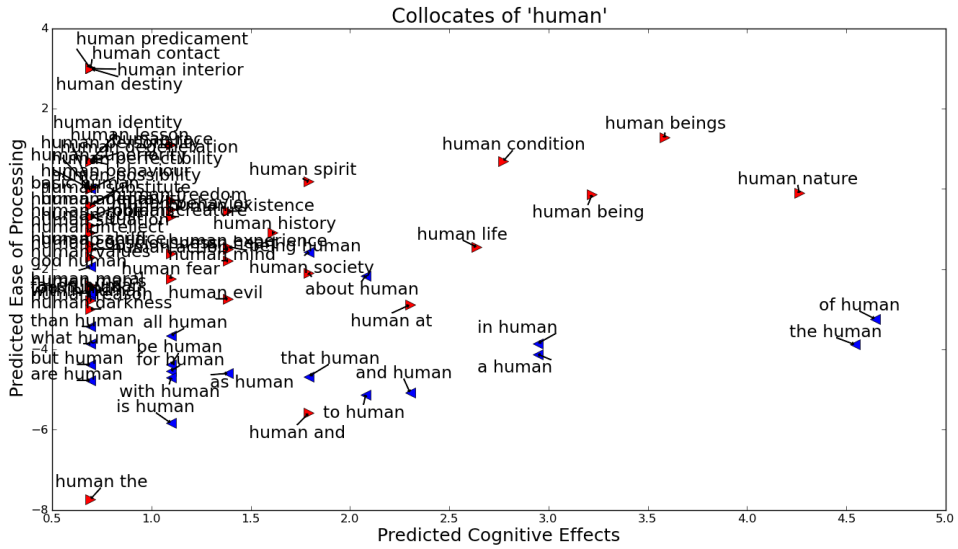

Figure 3.3: Relevance plot of "human"

Based on the plot's shape, and the terms' 'uncoupling' of frequencies (raw count and 'chunk count') and distribution along two axes, this was a success.

From this chart a pattern is immediately apparent. The collocates with high 'y'-values, the collocates that are 'easier to process,' are nouns with abstract values (e.g. "destiny", "existence", "spirit"). The lower y-scoring terms, the collocates that associate more widely, tend to be function words: articles, conjunctions, copulatives, etc. (This is not universally, true, of course, and a more detailed analysis will be the focus of the next chapter.) The terms along the horizontal, x-axis, tend to proceed from abstract to concrete from left to right: as the predicted cognitive effects increase, the collocates become more concrete. For example, consider the increasing 'concreteness' of "human condition" to "human being" to "human beings." "human the," perhaps unsurprisingly, scores lowest in terms of both processing effort and cognitive effects: for these

visualizations I have decided to ignore punctuation, hence why this collocation is present in the first place.[35]

To interpret this chart in terms of Relevance, the most relevant collocate to the seed term will appear in the top right corner of the chart: it has the highest cognitive effect and least processing effort. (Or, the collocation has the highest frequency, and the seed term's collocate has a high frequency and associates with the fewest other collocations.) In this chart, the most internally-relevant collocation is "human nature" – which, I think, is intuitively correct, and gives a much more accurate picture of Relevance than the first chart above.

In White's work, to assist interpretation he proposes dividing the diagram into different sectors, representing categories of relevance effects and co-citations. These categories are arbitrary, he explains, and index relevance in terms of "broad, qualitative gradations within the diagram" (p. 542). The three sectors White proposes are created by drawing lines from the most relevant term, the point of the pennant, outward. Figure 3.4 is the same as figure 3.3, but with these sectors demarcated.

[35] I have checked that no instances of "human the" appear in the corpus unbroken by punctuation. Of course, the combination of "human" and "the" unbroken by punctuation is not impossible, just unlikely.

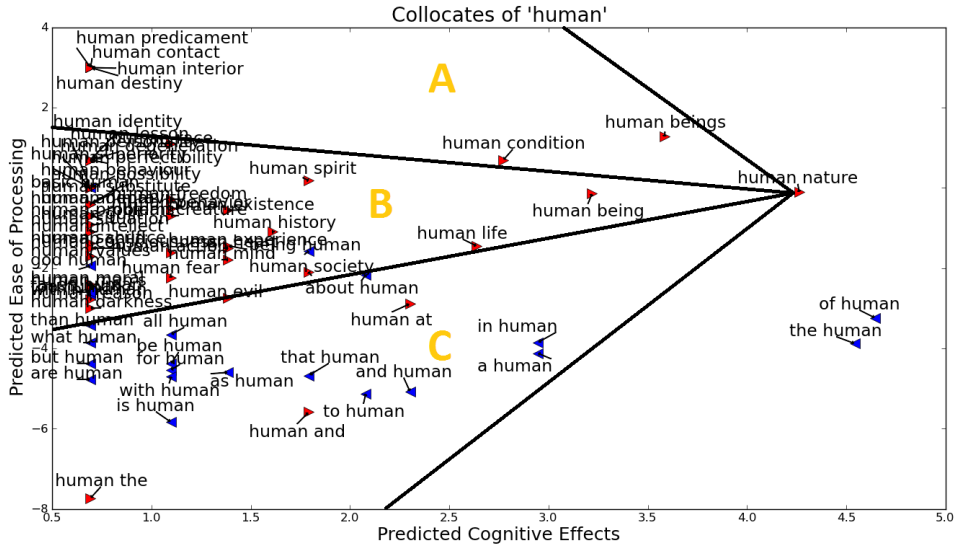

Figure 3.4: Relevance plot of "human" with sectors drawn

These sectors denote qualitative differences in the relevance collocates have within their collocations. The terms in section A, generally, have high relevance to "human" within the context of *LOTF* criticisms. Their relationship is readily apparent, relatively specific (they portray a 'cohesive' quality of human-ness), and they require little additional information or specialist knowledge to see their connection to the seed term. In section B, however, although the collocations contain traces of the qualities in A, they are more general, and the 'tone' is more variegated. Here we find "human fear," "human darkness," and "human evil," but also "human spirit" and "human freedom." While the collocations in A more adequately describe *LOTF* criticisms, the collocations in B could form in nearly any criticism of literature. Alan Partington (1998) refers to these types of collocations, collocations particular to a genre, as "*normal*" (p. 27, emphasis in original). The collocations in C, however, could form in nearly any kind of writing whatsoever. The terms here are generally function words, and almost exclusively 1-L collocates. There seems to be, therefore, an equivalence between sectors and the

genres of writing in which the collocations they denote would appear: and since I use *genre* here as a social/rhetorical term, these categories seem to coincide with spheres of *social interaction* (A: interaction with a specific text; B: a genre of criticism, interacting with no specific text; C: the shape of a language, a recognition of grammatical rules). Furthermore, each category demands varying degrees of specialization to make sense of the collocations in their context. As White writes:

> [I]t is easiest to see the relevance to [the seed term] of [terms] in sector A, and increasingly difficult to see the relevance of [terms] as we move through sectors B and C. The latter [terms] are by no means irrelevant to [the seed term]; they simply require more expertise, imagination, or effort to connect to it. (p. 543)

To distinguish the categories, White terms the A, B, C, sectors 'subordinate,' 'coordinate,' and 'superordinate,' respectively.

But these claims, especially regarding the C category, should give us pause. With respect to the 'superordinate,' C category, the idea that "the" is especially hard to relate to "human" might seem quite odd: it's really the most banal pairing in the world, determiner and noun head. It might be equally hard to believe that the relevance between these two terms requires any specialist knowledge at all, since the relationship is obvious and syntactic: "human," in many cases, essentially *needs* "the." But recall that this is not some abstract measure of association, or relatedness, but of processing effort: the ease with which the stimulus—the collocation—can guide interpretation to a viable conclusion. The question, 'How is "human" relevant to "destiny"?' can be answered in a number of ways on a range of levels: semantically, rhetorically, ideologically, etc. This question, though, is harder to answer with "the." Indeed if, as J.R. Firth (1957 [1951b])

suggests, one of a word's meanings really is its association with its collocate (p. 195), the meaning in this case is not immediately clear. Indeed, this is why finding this meaning is the job of linguists and other scholars in disciplines with high degrees of specialized knowledge. So, in this sense, a certain amount of 'imagination'—and a large amount of expertise—*is* required to relate "the" to "human." Furthermore, to re-invoke Relevance Theory, the *context* elicited by the collocations in C is very broad. I alluded to this above when I mentioned that these collocations, at first glance, appear to be genre-less. In this region we are not sure if we are making arguments about language in general, or more specifically about the language of the corpus. (And even making this distinction requires a substantial amount of specialized knowledge.)

However, this is not to say that the relevance predicted by these categories holds true in all contexts, or the context generated is always the same. This might be illustrated through a constructed conversation. White describes his proposal as akin to a library resource "recommender system," and—to bridge information science and linguistics—explains how it is like a conversation a user has with a human-designed system. Using the mainstay interlocutors of linguistic pragmatics, the fictional Mary and Peter, White describes conversations that might occur and the processing involved. In the following dialogues, Peter requests from Mary reading materials relevant to *Moby Dick* (the seed term). In this first example, Mary names materials that take little processing to relate to the text (A sector items):

> Peter: What should I read to follow up on *Moby Dick*? Mary: Oh, I don't know. For criticism, maybe *Studies in Classic American Literature* by D. H. Lawrence.

> Or *Love and Death in the American Novel* by Leslie Fiedler. If you want to stick with Melville, try *White Jacket* or *Mardi*. (p. 539)

These are quite specific, and are contrasted with another, more unfocused (but still relevant) response (C sector):

> Peter: What should I read to follow up on *Moby Dick*? Mary: Oh, I don't know. I have a whole run of *American Literature* over there. Why don't you go through that and look for articles? (p. 539)

The questions eliciting these dialogues are prefaced with the contextualizing phrase "to follow up on *Moby Dick*, " and therefore educe a context of academic research. (Implicatures are generated to create this context and derive the appropriate explicatures which generate this meaning.) Consequently, the distributions of terms, the 'recommendations' offered by the system, are not divorced of this context. Certainly not: in these dialogues, fictional Mary has a deep knowledge of Herman Melville studies and answers according to this knowledge. Her responses, therefore, echo knowledge from her experience: plotting these echoes, mapping their inter-orientations, reveals the shape of the field.

Finally, therefore, this shape can be 'flipped' given the right context established by the interlocutor. Melville-expert Mary replies using—what is known in pragmatics as—"encyclopaedic knowledge" [36] of the field. But if Mary is not an expert, and still chooses to reply with suggestions for Peter, she will generate another context and use a different set of assumptions. In terms of collocations, I imagine Peter making inquiries about one term's relevance to another: "Peter: what term might accompany "human" in

[36] Encyclopaedic knowledge is, essentially, assumptions one holds in memory: "encyclopaedic entries are sets of assumptions: that is, representations with logical forms" (*Relevance*, p. 92).

LOTF criticisms?". This question is fairly wide open, but Mary, assuming her expertise in these criticisms, should reply with "nature," and then the rest of the category A collocates. Again, however, this depends on the context established, and the encyclopaedic knowledge accessed. If Peter failed to propose the context of LOTF criticisms, or Mary had no expertise in the area, the reply might actually begin with the terms from section C: in this new context, it might seem more relevant to state first that "be" collocates with "human." A speaker of English could consult an encyclopaedic (or in this case 'dictionary') knowledge (to use RT terminology) of the term, which in this case would be syntactic rather than semantic. Thus, the terms' order of processing ease might be reversed by, essentially, selecting a different set of assumptions to populate the field.

3.6 Conclusion

To be clear, the measures and visualization I propose in this chapter are not meant to be precise, numerical methods for calculating internal collocational Relevance (called collocational strength, or association, elsewhere). Even if I were, there are many such other calculations with which it would compete (beginning with Mutual Information, proposed in the 1950s). Hans-Jörg Schmid (2007), for example, glosses several of these other calculations in "Non-compositionality and Emergent Meaning of Lexico Grammatical Chunks." But, on the other hand, compared with these calculations mine has the benefit of having a large amount of support: Relevance Theory offers a totalizing account of communication (hence the lengthy explanation), comes bundled with a large amount of scholarship, and opens up a multitude of further investigations (which we will explore in the final chapter). Furthermore, unlike other association tests, like Mutual

Information,[37] which rely on frequencies of single terms, the metrics I have developed use chunks to evaluate chunks: each chunk is assumed to carry with it a set of assumptions, those assumptions being the internal Relevance of its constituent parts, where these assumptions hold true in a context the collocation constructs. There are three benefits to my proposal. This method is: 1) a relativistic measure, which allows an investigation into the variegated manner in which collocates form around seed terms; 2) a visual method, using multiple variables of Relevance, making more apparent how collocations inter-orient themselves to one another in a manner not possible in other representations (such as the 'ordered list'). In fact, this visual method seems to be more conducive to Firth's collocational method as multi-dynamic investigations into the prolific variegation, and not just analyzing single word pairs: "an element of [collocations'] meaning is indicated when their habitual *word accompaniments* are shown" (qtd. in Herbst 1996, emphasis mine); 3) a visualization with an overall shape that allows us to see that 'third-part' of the echo – additional information made visible with technology.

In this chapter I have elaborated on the rationale, and sketched an approach and method, for the investigation. Using *House of Leaves* (2000), I tried to demonstrate weakly resounding echoes—*banal* criticisms—weakly *inferred* from the text because they are determined by forces alien to it, as well as advocating for a type of non-consumptive critical practice. In my summary of Relevance Theory I highlighted its criticisms and the huge amount of scholarship it has generated. My point is that this theory has the potential to illuminate communicative echoes, but its use has been

[37] Mutual Information is a probabilistic measure that uses raw frequency to calculate the probability of two terms co-occurring in a corpus.

hampered by a kind of negative feedback: its lack of use seems to perpetuate its lack of use, which is replaced with theoretical rather than practical musings.

My solution to RT's problems is—to put it crudely—to gloss over them: to put this theory to work, first, and critique the details later. To put this theory to work I have taken up Howard White's (2007a, 2007b) adaptation, and translated this co-citation method for use with collocation. I think that this method is promising because it is fairly unique—among the many other tests of collocational strength—and relativistic rather than empirical. Consequently, it should serve well to investigate the internal variegation of collocational forms. The last chapter will put this method to use. Here, I will test my anti-semantic conventionalization hypothesis and see what echolocation reveals about collocation. I hypothesize that word-pairings can be ostensive, rather than passively determined, acts. Therefore, like an echo these pairings contain a 'third part': the sense or meaning of each term, but also this term reflected as a *type*. In my corpus, these collocations are not representative of criticisms but of *types* of critical utterances: things others might say. And like my banal, faux-criticisms of *House of Leaves*, these echoes' sounds are prolific but diffuse.

4 SONAR Search

4.1 Monophonophobia

I faced it across the bin. The badger looked up and uttered the only really "strangled cry" I have ever experienced outside of fiction. This cry was the beginning of a high sound expressed in the funnies as *glug* or *gulp*.

--William Golding, *The Paper Men*

Each chapter of this thesis has begun with an example of a piece of literature blending the distinction between art and criticism – works that portray criticism as art, but 'art' in the banal sense: as 'artful,' artificial creations replete with pallid critical echoes. This genre of fiction represents criticism as dissociated from the art criticized, standing beside the art as its own unrelated artifice. But though criticism 'stands beside' its art, this is not to say it does not subjugate: the subjugated body, the corpus under reflection, is not the body of art but the body of the critic. The implication of this is a condemnation of criticism and an exposé of its misplaced focus: rather than the critic gaining mastery over the text, through criticism the text masters the critic.

William Golding himself wrote into this genre, publishing *The Paper Men* in 1984 – shortly after receiving the 1983 Nobel Prize in Literature. *The Paper Men* is a particularly acerbic reflection on literary criticism, aiming its critique precisely at the author / critic interaction and at repetition / subordination. The commentary is therefore aimed at a more 'social' rather than textual level – at least more so than the works referred to in chapters two and three. Rick Tucker, Professor of English, tenaciously pursues novelist Wilfrid Barclay, attempting to convince Barclay to sign a document making him his official biographer. Tucker is consumed by Barclay's work, consumed

by the desire to consume Barclay's papers, and does eventually gain access to Barclay's body (of work) by placing him within his debt: Tucker saves Barclay's life, and is rewarded with the promise of biography. But this promise proves false: Barclay reneges on the deal upon finding that the dire circumstances around Barclay's near-death and Tucker's heroism were inflated.[38] Spurned, and otherwise unable to gain mastery over Barclay, Tucker shoots and kills Barclay at the novel's end, rendering Barclay a corpse and his work a corpus: the murder literally completes Barclay's collection, since *The Paper Men* is Barclay's memoir and Tucker perfects it by terminating the work mid-sentence.[39]

The Paper Men relates to this investigation of echo, inference, and criticism as both the most pointed condemnation and the most reticent yet with regards to the literary critical problem of repetition. Barclay is a successful but self-consciously uncreative novelist, producing banal works but—as in the epigraph—assigning anything that displeases him to what might be called 'literary kitsch': 'expectation' is something to be derided, so long as it also might be controlled. Fiction is a site of repetition, and

[38] This demands contextualization and an interesting historical note. The specifics of Barclay's near-death in the novel involve Barclay and Tucker going for a hike in the mountains and Barclay slipping and sliding down a steep cliff, stopping himself by grabbing onto poorly-rooted plant life which threatens to give-way. Tucker rescues him by offering his hand, boosting him up and away from his demise. However, it is later revealed that Barclay was in no danger at all: had he fallen, he would have dropped onto a meadow, just under his feet, and not off the mountain. Though it is likely that Tucker knew this, Barclay did not: heavy fog enveloped the area and prevented him from seeing the ground. From a historical, text-reception perspective this is very interesting since the popular story of Golding's incredible success with *Lord of the Flies* is that critics, rather than the general public, were responsible for its initial, massive uptake. As R.C. Townsend (1964) reports, upon its publication *Lord of the Flies* received "an interestingly mixed reception" (p. 153) but was picked up by students and critics of the Ivy League American universities. It was the critics, therefore, who 'picked up' Golding's text and gave his career the initial 'boost' needed for his eventual monumental success. In the context of *The Paper Men* and Golding's general ill-view of critics, then, it is hard to escape from the obvious implication: Golding claims that he didn't need help. The critics' help was manifest but false – not only was it unnecessary it was unwarranted and unjust. He could have 'done it on his own.'

[39] The terminal sentences are almost embarrassing to report: "Now he is leaning against a tree and peering at me through some instrument or other. How the devil did Rick L. Tucker manage to get hold of a gu " (p. 191).

therefore deserves scorn, but is also a site over which Barclay is master: his common refrain (and expression of dark amusement), "ha! etc.", bespeaks a dependence on replication that he can only produce, ironically, elliptically. Tucker's attempts to fictionalize Barclay through biography are therefore doomed for failure, as this would expose Barclay to the indecencies of fiction of which he himself is responsible. To put all of this simply, Barclay is infatuated with and terrified by *repetition*.

So it may be ironic that Golding should concern himself with such repetition when criticism of his work, *Lord of the Flies* (1954), is labeled 'banal.' Or, perhaps *The Paper Men* (1983) is an indignant reply to such criticism, in as much as those criticisms implicate Golding as the critical works' implicit progenitor: recall that Patrick Swinden (1987) asks, "What is it *about Golding* that makes critical conversation about him so banal?" (p. 570, emphasis mine). Either way, *The Paper Men* criticizes criticism, as the other literary examples have, but also remains somewhat ignorant on the topic of repetition, suggesting that there might be something more, as yet unexplored features, to this critical feature.

This final chapter will explore these features. Using the theory and tool developed and explained in the last chapter, in this chapter I will present my findings on collocation in *Lord of the Flies* criticisms and argue against the semantic conventionalization—the semanticization—of collocations' meanings and standardization of their functions. Beginning with our prototypical collocation, 'human nature,' I will show that this phrase almost always appears on the horizon of others' speech and thoughts – as such, it might be called an 'echoic utterance' in Sperber and Wilson's Relevance Theory, as it is a "second-degree interpretation" (*Relevance*, p. 238),

an interpretation of another's thought. The most basic evidence of this is that "human nature" is frequently attributed to William Golding himself, as his interpretation of his own novel: "[27 – MFF]: Golding has summed up the theme of Lord of the Flies as follows: The theme is an attempt to trace the defects of society back to the defects of **human nature**." I believe that this provides evidence that the act of collocation—of pairing two words together—is ostensive, giving rise to various implicatures through this ostensiveness. This ostensiveness and the resulting implicatures is the 'third part' of the echo (echoic utterance) to which I have been referring: the ostensive combination of terms makes manifest the *constellation* of criticisms of which the *one instance* is a part and therefore the collocation's meaning—the implicatures to be derived—stem from this formal pairing since the resulting echolocation motivates the reader to look for the author's attitude to such a constellation. As I argued in chapter two, part of a collocation's meaning has to do with form as function, or a collocation deriving meaning from placement. This meaning-from-placement is how this category of collocation achieves Relevance: by making form ostensive, hearers will interpret a collocation not just in its context of use but also as others have used it. Admittedly, not all paired terms might necessarily have this communicative purpose, but perhaps this is the distinction of different kinds of collocation sought by the theorists discussed.

Second, I will investigate terms collocating with the four main character names of *LOTF*: Jack, Ralph, Piggy, and Simon. A cross-collocation comparison between these names' collocates reveals the relationships critics assign to them. These differ from the text of the novel itself, as might be expected. Furthermore, these patterns have implications for genre research though, in this vein, due to space, I can only speculate: an

analysis, however, reveals collocations that cannot be called ostensive, implicature-generating phrases since they seem to arise unaware and non-ostensive. Proper noun collocations seem to follow a process akin to that described by rhetorical genre theory: "typified rhetorical actions based in recurrent situations" (Miller, 1984, p. 159) generates forms we know as genres, and therefore genres are not dependent on the replication of some literary (or other) 'rule' of production (i.e. texts are not modeled after genres, as frameworks) but situations and actions motivate productions recognized as patterns. But at this point the story becomes murky: if collocations in criticism arise from critics responding to recurrent situation, what exactly is this situation? The art, the institution? Furthermore, I recognize these particular collocations in *LOTF* criticism as features of *a* genre (the public school story), a genre in which I believe *Lord of the Flies* participates, and therefore these collocations might be the products of generic activity. Two hypotheses might be made, here, both of which are reasonably far-flung. The first is that genre is really just the product of critics, as Ralph Cohen (1986) asserts: genre is a category of classification established by the professional participation of those responsible for criticizing literature, and the act of that criticism results in classifications made, maintained, and reconstructed. In this light, what I recognize as a feature concomitant with a genre—the collocation of 'proper noun+*and*', and the 'trio' of characters established—is not an experience with the literature but with the literature's reception: an experience with the genre of criticism rather than a genre of fiction. The other hypothesis is that critics really do participate in and extrapolate from art, and therefore my recognition of the genre is as contemporary with the other critics,

participating in the same situation and performing a typified rhetorical action – as in rhetorical genre theory, as defined above.

Finally, I will revisit the very notion of collocation by questioning just what we can learn from investigating co-occurrence of words both immediately collocating (as has been the near-exclusive focus of this study) and having an extended proximal relationship across spans of text. Because my argument on the "human nature" collocation and semantic conventionalization is based on this collocation's close relationship to another feature—though pragmatic, rather than textual—I reflect on the linearity of language, the necessity of such spatial relationships, and what this means for methodologies concerned, like mine, with word proximity.

4.2 'Human Nature'

As a reminder of the top-ranked collocations in the corpus, Table 4.1 is a simple ordered list of the top 30 bi-grams:

Rank	Count	Bigram	Rank	Count	Bigram	Rank	Count	Bigram
1	2733	of the	**11**	334	of a	**21**	275	he is
2	976	in the	**12**	334	to be	**22**	253	as a
3	678	and the	**13**	329	that the	**23**	251	at the
4	557	to the	**14**	323	by the	**24**	246	the novel
5	482	on the	**15**	317	the island	**25**	226	the beast
6	468	Lord of	**16**	315	the boys	**26**	212	in a
7	441	the Flies	**17**	305	is a	**27**	207	as the
8	380	from the	**18**	301	of his	**28**	205	is not
9	365	it is	**19**	299	for the	**29**	188	It is
10	340	is the	**20**	298	with the	**30**	145	there is

Table 4.1: Ordered list of the top 30 bi-grams in the corpus

As I noted in chapter two, most of these are expected. Of the top thirty: 13 are preposition + determiner; four are noun phrases (and another two components of the novel's title); 8 contain copulative *be*; one contains a conjunction ("and the"); and one is preposition + possessive pronoun ("of his"). These bigrams are expected because they

are components of typical literary criticism: the large number of prepositions, for example, bespeaks the literary critical concern with situating an argument textually or socially (e.g. '*in the* novel', '*as a* piece of art', etc.). On the other hand, I think the copulative *be* + negative particle ("is not", rank 28 with 205 instances) is interesting, not just a feature of criticism but of *LOTF* criticism, and it will be taken up in this section. However, of concern right now will be the bigram "human nature," occurring 71 times in the corpus with a frequency rank of 109.

Recalling chapter two and the 'saliency metric' presented therein, in which terms are ordered according to their frequency and collocability (their prolificness and semantic specificity, also the key ingredient to my collocation visualization chart), "human" was ranked alongside the function words – also a banality indicator, I suggested that "human" was a significant term for the investigation of collocation and banality because of its tendency to associate with other terms, and its frequency of occurrence. "human nature" is an intuitively obviously important collocation in the corpus, and one represented as the most relevant in the following visualization:

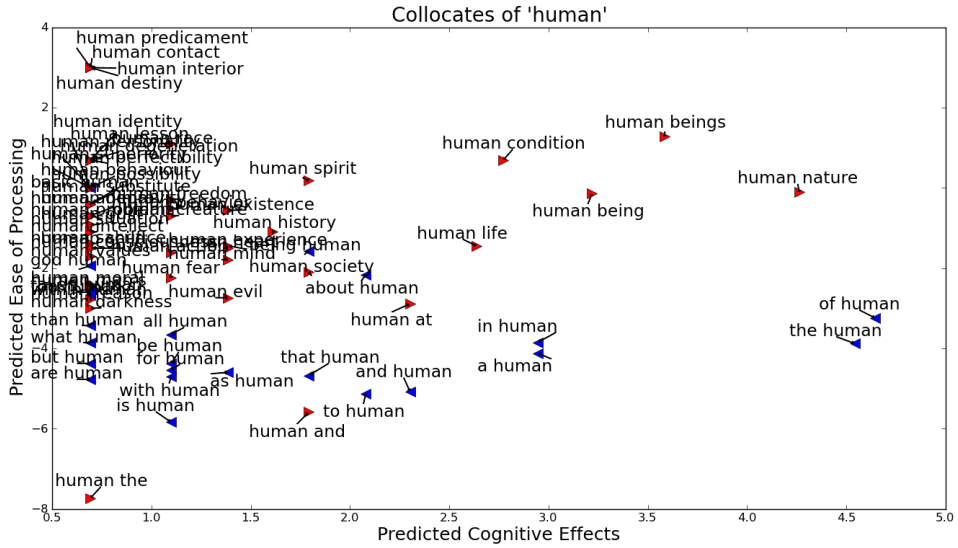

Figure 4.1: Relevance plot of "human"

To review, this chart derives from Howard White's (2007a, 2007b) work on bibliometrics and Relevance Theory, and ranks a seed term's collocates (in this case, collocates of "human") in terms of their frequency along the x-axis and their propensity to form in other 'top-level' collocations, relative to their collocates' frequencies, along the y-axis (so, in a collocation—"human predicament," for example—located high on the y-axis, the collocation "human predicament" occurs frequently but "predicament" occurs alongside relatively few other words than "human"). Further, these two axes correspond to the main variables in Relevance Theory, cognitive effects and processing effort, where the x-axis (term frequency) indicates cognitive effects and the y-axis (collocability) indicates ease of processing. Though two other collocations ("of human," "to human") occur more frequently than "human nature," "nature" and "human" are more tightly bound because "nature" occurs in fewer other collocations.

Before proceeding, I do want to contextualize this phrase's frequency in historical terms. Looking at data from the *Corpus of Historical American English*, plotted in

Figure 4.2, we see that the use of "human nature" tapers off into the twentieth century, and nearly flatlines right around the time *Lord of the Flies* is published in 1954.

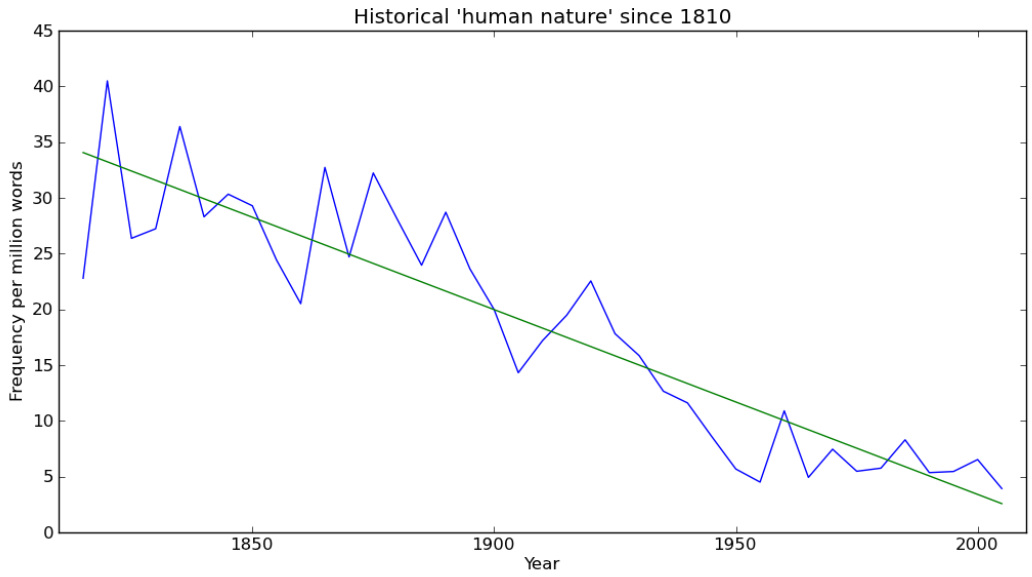

Figure 4.2: Historical "human nature" since 1810

I cannot stake too much of an argument on this, but this trend makes this investigation all the more curious since so many instances of "human nature" are found in *Lord of the Flies* criticisms when the phrase seems to have fallen out of the historical lexicon. And without dwelling on this point for two long, I did note a decline in the frequency of "human nature" across *all* forms of writing (academic, literature, news, etc.) and therefore frequencies in one area did not increase or decrease to occlude frequency changes in the others. One final note, a loose speculation: since the frequency of "human nature" is at least higher in the nineteenth century, I do wonder if maybe we find it so often in Golding criticism because these critics were also well-read, and therefore encountered it, in older 19th century literature. This is a hypothesis to be tested quantitatively, and one I have yet to see: Clifford Siskin, though, in *The Work of Writing*

(1998) claims that Romanticists replicate the language and forms of the Romantics in their academic work.

But back to 'human nature' in the corpus. First, 'human nature' is very often explicitly inter-textual. In 44 of the 71 instances, this phrase occurs either within or just on the edge of another's speech. In several cases, Golding's own words are reproduced – these instances are in the form of a well-known quotation in which he relates his intended theme:

> [10 – AsFab]: Golding explained that the novel illustrates that **human nature** is the source of evil, a view shaped by cruelties he had seen during World War II.

Some further examples:

> [27 – MFF]: Golding has summed up the theme of Lord of the Flies as follows: The theme is an attempt to trace the defects of society back to the defects of **human nature**.
>
> [37 - PaO]: The "trite, obvious and familiar" moral lesson of Golding's novel is that we are capable of the most heinous cruelties in the service of our pride. The "beastie" appears to the reader in a variety of guises: as a "snake-thing," "beast from water," "beast from air," and, finally, as an aspect of **human nature**.
>
> [35 – NotB]: In such works we find a tendency to present **human nature** at an extreme: in More's utopian fantasy and in Aldous Huxley's Island we see **human nature** and society at their best.

In these and many other cases, the critics explicitly arrange this collocation in and among others' voices. If we might identify a common *function* of 'human nature,' in this corpus it seems to work indexically and analogically – to point to a metonymic

representation of other criticisms and to construct the point being made. (But also used *obliquely* to construct the critic's own argument, and *explicitly* to construct another's.) In 35 – NotB, the critic is not making a comment on "human nature" directly, or even a representation of "human nature," but by citing this phrase establishes a history of related theses. In this way, the collocation is metonymic because by broaching such a phrase one must also have an opinion on it, and this opinion is put to further use by the reporting critic.

And even if there is a common *function*, we cannot point to a common meaning. In the first excerpt, the collocation implies a causal structure (human nature as a generative and deterministic device), in the second the psyche (a cloaked mental image), and in the third a typified narrative experience (a certain kind of work uses "human nature" in a certain kind of way). In the second selection, especially, the critic frenetically bounces back and forth between the singular and the general. The quotation marks frame the critique as both 'what some might say' and as identifiable features of the text. Finally, in 37 – PaO, in moving from constructed fictional things ("beastie," "snake-thing," "beast from water")—that are also disembodied projections, to the non-fictional but embodied "human nature"—the critic vacillates between dispositions and discourses. However, in spite of the diversity of uses of "human nature," the critics' works have mostly the same objective: to prove Golding right, and humanity flawed, in remarkably the same ways.

This similarity of argumentation around "human nature" is ironic in light of the attitudes expressed towards it. In general, and especially so in 37 – PaO, critics use this collocation disdainfully – partly a byproduct of the latent irony. In this way it is used

ostensively: the critic recognizes the collocation's 'special' status as an *alien* piece of discourse. The collocation does not speak for itself, but for another critic. And it is this feature, this ostension, that implicates it as 'echoic' in Sperber and Wilson's sense of the term. In terms of Relevance Theory, an utterance is echoic when it is made manifest that what has been said is what others have said – real or typified imaginary persons, producing real or imaginary utterances. An utterance is echoic if it is Relevant *because* of this interaction in the context of another's speech or thought.

My proposal is that collocation, as echo, is a method for generating the search for Relevance: the motive for pairing two terms, the act of collocation, is an ostensive act to have the hearer recognize the collocation as the words of others and consequently search for the speaker's attitude as it relates to others' attitudes. This ostensive use of form, and focus on word-pairing rather than word-meaning, seems to resonate with Firth's separation of form from semantics: "Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words" (1957 [1951b] p. 196). Collocations *mean* through their form – a necessary condition for generating this type of meaning, since in these cases of collocation it is the form that is ostensive. This perspective removes from our concern "the conceptual or idea approach to the meaning of words" and replaces it with a primary concern for more accountable, more intersubjective approaches: contemporary pragmatics, or Bakhtinian dialogism where we speak only through others' words.

In fact, in the corpus a site where "human nature" occurs is also a site of ventriloquizing: not only is "human nature" relegated to another's words, but "human

nature" (which is the critic's term, but assumes the consciousness of the author) is what allows Golding to speak those words *through* a character:

> [27 – MFF]: And a few pages later Simon, the convulsion-afflicted mystic, says of the Beast: "What I mean is . . . maybe it's only us." This rather subtle interpretation of **human nature** from a small boy demonstrates further that Golding is so intent on his moral message that he will not hesitate to make the youngsters dance to his tune.

This is quite complex, because here the critic presupposes not only an "interpretation of human nature," but that "human nature" refers to the author as a speaking subject. Ironically, here "human nature" itself dances to the critics' tunes: the implicatures to be immediately derived are that Golding not only holds such a claim but that this claim in fact represents his presence in the text, and that this disembodied collocational-representation is of poor moral character for ventriloquizing a "small boy" (or a "youngster"). But further, it is therefore the critic playing the same tune – but the same tune for a different song: by presupposing an underlying tenet of *LOTF* scholarship the critic has a 'small boy' speak for Golding, but where Golding further speaks for his own representation *by* critics, and therefore the critic arranges a familiar chorus and claims this arrangement—not necessarily the tune—as novel.

This summoning-distancing is common. Even when the collocation does *not* hang on the edge of another's speech, it hangs on the edge of the critic's:

> [16 – UnPop]: […] Golding's symbolism emanates from a desire to support the conclusions rather than from a total commitment to his subject, whether that

> subject be defined as the fate of a handful of boys after a nuclear attack or the defects of society and **human nature**.

This example, like the previous, participates in a common type of criticism in the corpus. These criticisms identify as predetermined *what* Golding claims, and therefore the issue to be explored is not his message but his style: 'it's not what he says, but how he says it.' In this example, "human nature" hangs at the end of the thought, seemingly compulsively appended after another common collocation, "defects of society" (8 instances). As the critic makes clear, the *definition* of Golding's concern (as well as *who* should define it) is unnecessary, as it can be summoned and dismissed in the same thought.

Finally, the critics themselves recognize and write about collocations as intertextual. Like Golding's Wilfrid, this critic tacitly condemns repetition of stock phrases:

> [32 – DooT]: With the resolution of the structural antithesis it is clear that to try to explain the novels with phrases like 'the darkness of man's heart', 'man's fall into guilt', or 'man's loss of innocence', is far too reductive, *Lord of the Flies* and *The Inheritors* are more comprehensive than this. They are also more obscure.

Here, this critic cites three examples of types of utterances, utterances others might say, and distinguishes his work from those criticisms of the corpus in tune with the chorus. The critic casts a wide net: by offering three types the critic not only gains a rhetorical advantage by a pleasing parallelism but makes more manifest that it is not these *exact* phrases that are wrong but these *types* of phrases that embody poor scholarly thought. As Sperber and Wilson write, "An echoic utterance need not interpret a precisely attributable thought: it may echo the thought of a certain kind of person, or of people in general"

(*Relevance*, p. 238). In fact, these particular phrases—"the darkness of man's heart," "man's fall into guilt," or "man's loss of innocence"—never actually occur in the corpus. Interestingly, though, "loss of innocence" does occur (8 instances) as a frequent description of a poignant point of the novel: the end, where Ralph weeps "for **the end of innocence, the darkness of man's heart**, and the fall through the air of the true, wise friend called Piggy" (Golding, 1954, p. 237, emphasis mine). So if this passage from the novel might be deemed a textual source for the images conjured by the typified expressions represented above, the precise attribution is elided by concatenating "man" with the phrases. Once again, the character Ralph is not speaking for himself but for the critical imagining of Golding.

4.3 'Not'

The above selection from the corpus (32 – DooT) represents a common method of argumentation in literary studies: the 'it's not this, it's that' proclamation, where the target is made ironic and the critic's exclusive knowledge—undistorted by whatever flaw afflicted the previous analysis—about the topic can be revealed.[40] Closely related to this is the charge—again, enacted above—of reduction: '*x* states *y*, but it's not **just** *y*…'. These types of arguments, here within proximity of "human nature," bring me back to the collocation I noted from the chart at the beginning of the chapter: "is not," ranked the 28th most frequent bigram with 205 occurrences in the corpus. Although it is difficult to

[40] Golding's *The Paper Men* dramatizes this. In the following passage Barclay ironically describes his experience at an academic conference, listening to a presenter critique Barclay's fictional work and deride the academic work of a peer: "Prof. Tucker, still toneless, was now pointing out the significant difference between his graph and the one constructed by a Japanese Professor Hiroshige (that was what he sounded like), for Professor Hiroshige, it appeared, had not done his homework, to our surprise, and had also been guilty of the gross error of confusing my compound sentences with my complex ones. In fact Professor Hiroshige should get lost and leave the field to the acknowledged expert, who had heard from the author's own lips that he did not tolerate so overly broad an interpretation in his iconography of the absolute, or words to that effect" (p. 23).

claim definitively, of course, this phrase seems to hold certain weight in *LOTF* criticisms just based on its relative frequency of use. According to the *Corpus of Contemporary American English* (*COCA*) the bigram "is not" occurs with the greatest frequency in the Academic writing category, at 540.75 occurrences per million words, and highest in the Philosophy / Religion sub-category (857.4 per million), second highest in Law and Political Science (732.84 per million), and third highest in Humanities writing (652.85 per million). In the *LOTF* corpus, "is not" occurs at an extrapolated frequency of 1044.3 per million words.

"is not," therefore, stands out as a collocation possibly particular to the *LOTF* corpus. And this stands to reason: given the propensity for the corpus to harbour repetitive, mutually recognizable collocations, it should also employ a substantial amount of negation in the form described above. My reasoning for this is that, as I have been arguing, some sets of collocations are Relevance generating devices, where the motive for the utterance is to have the hearer recognize them as the words of others and therefore help guide the hearer to look for Relevance in the speaker's attitude to the other speakers' position. Negation seems to be a close-cousin of this, since negation relies almost entirely on the success of the proposition being 'made mutually manifest' to the hearer. And indeed, we do see 1R collocates of "is not" that provide evidence for this function in the corpus: 5 occurrences each of "is not just," "is not only," and "is not that":

> [32 – DooT]: We have shared Ralph's perspective for most of the novel and sympathize with him. The change in point of view to the officer **is not just** to reveal the officer's lack of insight; it is designed to shock us out of our

rationalistic complacency by revealing that *Lord of the Flies* is condemning the point of view it has made us read from.

So, while it is interesting in itself that "is not" is itself a commonly-occurring collocation in the corpus, I sought to uncover its Relevant collocates: what is being ostensified by this pairing? The following figure is my Relevance scatterplot with "is not" as the seed term:

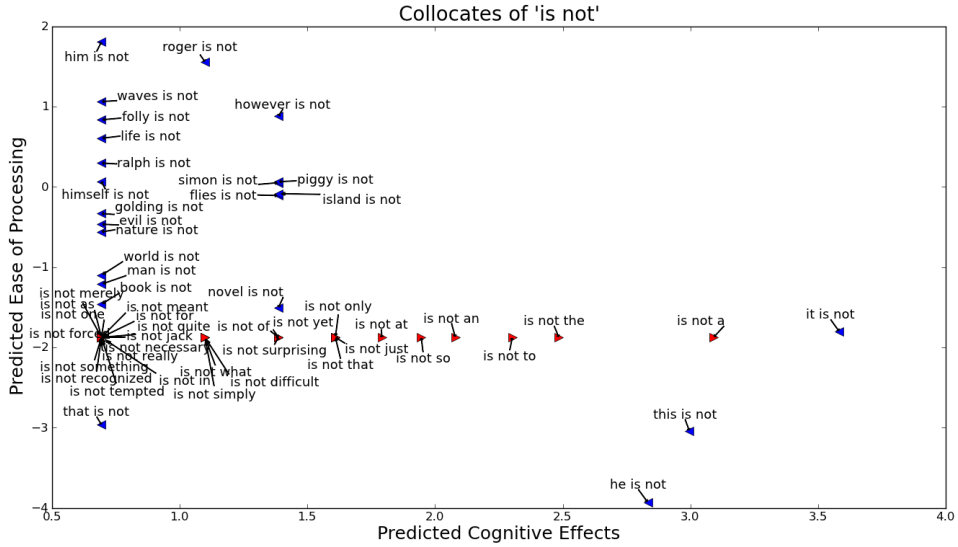

Figure 4.3: Relevance plot of "is not"

The first thing striking about this chart is that, proportionately, many more items are higher on the y-axis on this chart than on others – the 'human' chart, for example. This suggests that the collocates of "is not" are more 'specific' to it: in terms of Relevance Theory, they are therefore also easier to process as relating to this collocation than to others. (Admittedly, though, this may just be an artifact of using a bigram as the seed term as opposed to a single word.) So, we have few function word collocates to contend with (which would appear at the bottom of the graph, recalling the triptych schema from the last chapter) and see, in the central area, the profusion of genre-specific terms

common to literary criticism: “is not simply,” “is not surprising,” “is not just,” “is not so,” etc. This supports my claim, from the last chapter, that this central area is where we will find terms common to broad categories of writing (e.g. literary criticism), as opposed to language in general—the bottom half of the chart—and specific topics, like *LOTF* criticisms, at the top.

The upper half of the chart is interesting because it contains every major character (Ralph, Piggy, Roger, Simon) in the same phrase (‘*x* is not’), except—oddly—Jack. But perhaps this is not so odd: a common point of discussion in these criticisms revolves around ‘who symbolizes what’ (e.g. which Biblical figure is Simon?), where this symbolism is taken further to represent the theme / character / moral of the text. As we observed with “human nature,” these collocations elicit a mutual understanding of other scholars’ work—often with tacit disdain—and critics’ positions are also established with this phrase (‘character names + *not*’). Jack, of course, is *Lord of the Flies*’s antagonist and the most obvious, near-literal embodiment of evil. The lack of contest over his character suggests either that there really is nothing to dispute, or that critics disagree with one another only speciously: they might critique one another’s treatments of the text through the characters, but they really do agree with the ‘evil humanity’ thesis – Jack is simply uninteresting or indefensible.

Admittedly, the left side of these charts is the area in which only relatively few realizations of the collocations occur (there are two instances for Ralph, three for Roger, and four each for Simon and Piggy) – though, for a small-sized corpus, and taken in aggregate, these occurrences do seem meaningful. Furthermore, Jack does appear as a collocate of “is not” – just as a 1R collocate as opposed to those of the same form for the

remaining trio of characters. The corresponding excerpt from this realization of "Jack is not" both dampens and bolsters my above hypotheses:

> [22 – LOTF2]: What is wrong on the island **is not** Jack, as Ralph and Piggy think, or a Beast or Devil to be propitiated as Jack thinks. What is wrong is that man is inherently evil, as Simon has already maintained; the "ancient, inescapable recognition" is of something in Ralph and Piggy, and Simon himself, as well as Jack and Roger.

In this selection, the critic writes the opposite of what I proposed: Jack's character is in dispute, which at least presupposes the obvious conclusion about him but then denies it. However, within the context of this presupposition-negation we do find our poignant collocations: "inherently evil," as well as a quotation from the novel—the "ancient, inescapable recognition" (with which the Beast looks upon Simon)—and a profusion of character names and their perceived thoughts. The phenomenon of negation is complex, for in going against the grain this critic re-affirms the critical body's status quo: an acceptance of the terms of the discourse, resulting in vociferous opposition but consequently presupposing near total agreement. Negation, in this context, animates the speaker (real or hypothetical) who has made (or will make) the claim being denied, even if such a speaker exists only in the imagination of the reader as a possible response to the fiction. This speaker, made manifest within the history of an institution (literary criticism) and a piece of art (*LOTF*), expediently establishes the possibilities of what might be said and creates the context in which the critic can work. In the following (4 – ALOTF), in denying Golding the author-type of 'social novelist' the critic reframes both Golding's *and the critic's own* 'type': the critic is not one who is interested in pursuing

superficial investigations, and consequently this work is not superficial but substantial – a substantial intervention in metaphysics and literature. This negation therefore helps to establish the genre of scholarship in which the criticism is situated.

> [4 – ALOTF]: He is **not** simply a social novelist attempting to see man's response to a given society, but a metaphysical writer interested in states of being and aspects of survival. . . . Golding is interested **not** in the superficial capabilities of man but in those long-buried responses the latter can suddenly evoke in order to satisfy or preserve himself.

4.4 The Character of a Name

From "is not," which we find collocating with character names, we now turn exclusively to character name collocates themselves. This will also serve as an example of comparative collocation analysis, as we examine the main characters' 'name collocates' in the context of each other. To begin, Figure 4.4: a plot of Jack's collocates.

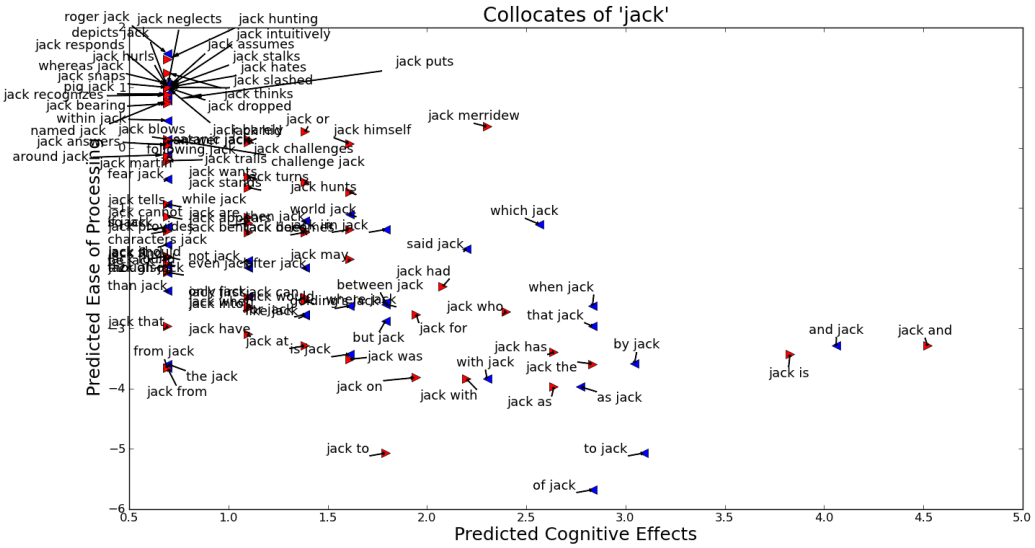

Figure 4.4: Relevance plot of "jack"-1

Two features of this plot are apparent, and are found in all of the proper noun collocation plots: the first is the cluster of terms in the top left of the chart. These will be collocations that have few occurrences (a minimum count of 2), but are terms that occur almost exclusively with that name (character). For example, for "Jack" such collocates are "hates," "stalks," and "snaps." This 'cluster of personality' is unique because these terms tend to be extremely candid descriptors of the critic's attitudes to the characters. Further, since this candid corner tends to aggregate multiple, low frequency terms (some of which are nearly synonymous with one another) we are able to effectively evaluate the critics' attitudes by way of 'triangulation' – a dialogic mode of analysis, since this phenomenon of the plot is dependent on a sampling of multiple voices. Technically speaking, this is a two dimensional representation since two measures, independent of one another, locate the orientation of terms graphically. But a third dimensions is implicit: the multiple voices that speak with or against one another. In the criticisms, this

conversation is only implicit – this graphical representation makes the conversation explicit.

Unlike the collocations occurring elsewhere (the function words at the bottom of the plot, for example) the terms that frame this personality portrait can apply only to these literary figures. So "Jack stalks," "Jack hates," and "Jack slashed" certainly do not occur in all criticisms, but criticisms do cluster these *types* of proper-noun terms in like manner. This corner offers a quick reference guide to the critical community's representation of characters.

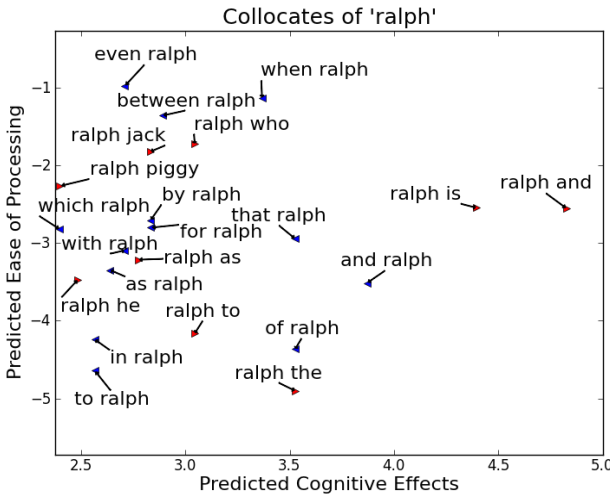

Figure 4.5: Relevance plot of "ralph"

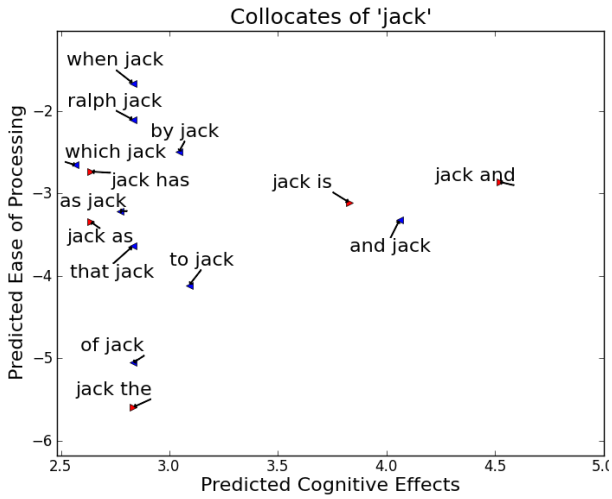

Figure 4.6: Relevance plot of "jack"-2

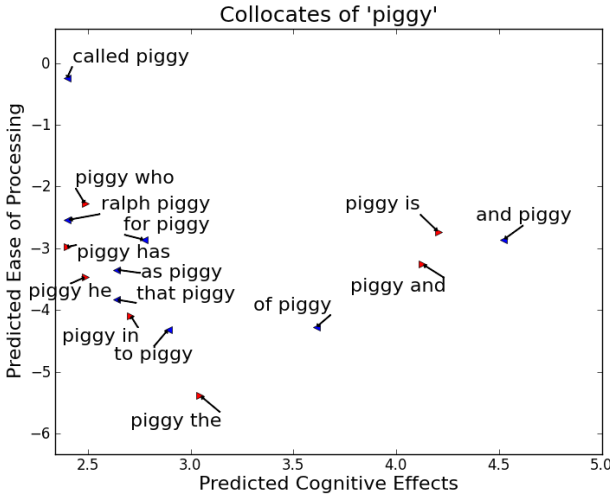

Figure 4.7: Relevance plot of "piggy"

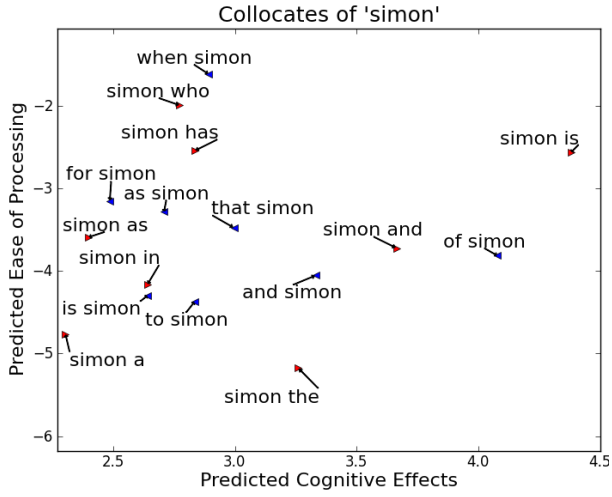

Figure 4.8: Relevance plot of "simon"

Of primary importance to this section, however, is the point of the pennant: the area of the chart with the most Relevant collocations. Above (Figures 4.5, 4.6, 4.7, and 4.8) are collocation plots for the four main characters: Ralph, Piggy, Jack, and Simon. These plots are zoomed in to depict just the tip of the chart. Interestingly, the collocate at the very tip of the pennant is not the verb *be*, as might be expected since a primary focus of criticism is to explain characters' *symbolism*, but the conjunction "and." This finding is relevant on its own, but I believe acquires new meaning in generic terms. Texts

typically within the British public school story tend to have 'trios' as their main characters. An abbreviated list: Tom, East, Arthur in *Tom Brown's Schooldays*; Stalky, Beetle, M'Turk in *Stalky & Co.*; Ender, Bean, Petra in *Ender's Game*; Harry, Ron, Hermione in *Harry Potter*; Ralph, Jack, Piggy in *Lord of the Flies*. As I consider *Lord of the Flies* a public school story,[41] I found these graphs striking and hypothesized that the *and*-ed terms would be the other character names. This hypothesis proved correct:

"Ralph and"…		"Piggy and"…		"Jack and"…		"Simon and"…	
43	Piggy	**17**	Ralph	**26**	his	**11**	Piggy
34	Jack	**9**	Simon	**10**	the	**9**	the
8	the	**5**	the	**9**	Roger	**3**	Tuami
2	Peterkin	**2**	his	**8**	Ralph	**1**	to
2	his	**1**	Jack	**5**	Peterkin	**1**	thinks

Table 4.2: Collocates of *LOTF* character names

From Table 4.2, depicting collocates of each character name and "and," we see an interesting network appear, reminiscent of character-networks in other public school stories. Ralph and Piggy are paired together, and Ralph's ties are stronger with Jack than are Piggy's. Jack is paired weakly with Roger, and does not reciprocate Ralph or Piggy's relationship. Simon is nearly cut out of the fraternity: his only connection is with Piggy (and Tuami is a character from another of Golding's novels, *The Inheritors* [1955]). Moreover, predictably, Simon is *equated* more than he is *associated* with anyone else: rather than "and," his most common collocate is "is" (e.g. [46 – OnSym]: "**Simon is** the incarnation of goodness and saintliness"). A common charge of the critics is that Golding makes his characters 'dance to his tune,' but in this analysis we find Simon as the veritable whipping-boy of critical commentary.

[41] A few others have made this claim, as well: Kirstin Olsen (2000) calls *Lord of the Flies*, "an extreme version of the school-story in which the school has been removed" (p. 56).

But here, after rehearsing the standard generic theme of 'the trio' in the public school story and equivocating collocation of terms with a broad concept of relationships characters have to one another in a text, we should pause. Just because these collocations sound natural does not mean they are, and I should emphasize that these *and*-pairs are not found to nearly the same degree in the novel. In fact, the collocation distributions from the novel are radically different. Ralph's is below (Figure 4.9):

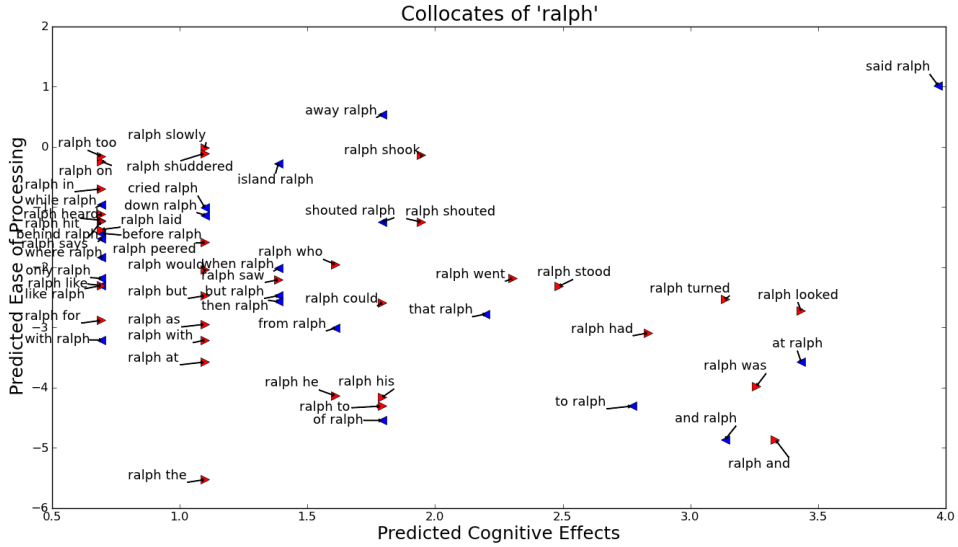

Figure 4.9: Relevance plot of "ralph" in *Lord of the Flies*

In the collocations in the novel, the narrated action words are, not surprisingly, represented as substantially more Relevant to "Ralph": "was," "looked," "turned," "said." Indeed, this figure gives a sense of how fiction really does not take the shape of 'natural language' – for the first time we see the collocate and collocation with absolute optimal Relevance, in the top right of the plot: "said Ralph." This collocation is optimally Relevant because, statistically, it is the most frequent and "said" occurs in few other collocations than with "Ralph." However, this outlier defies patterns found in non-fiction text: in non-fiction, the pennant shape we observe suggests that there is a constant

tradeoff at work, where collocations tend to be either commonly occurring or more or less 'focused' in their co-occurrence. The different distribution in fiction, and particularly fiction, suggests something really is socially / cognitively different—perhaps the system is more designed—and this difference is manifest as a hyper attentiveness to a fictional character's actions, in keeping with the modern dictum of fiction: 'to show, and not tell.'

My observation is that the collocates of character names in the critics' writings are highly reminiscent of my experience with a literary genre – a genre in which I claim *Lord of the Flies* participates. However, the features eliciting this connection are present in the literature's *criticism*, and not the literature itself. A similar 'triangulation,' or trio-forming does occur in the novel (there are six instances of "Ralph and Piggy" – two of "Ralph and Simon" and one of "Ralph and Jack") so there is likely both some translation and substantial innovation. But this is curious: is it not odd that a feature possibly associated with a genre show up in a story's criticisms? Are these collocations simply present in most criticism – are these markers of discourse, or register? On the other hand, perhaps critics of *LOTF* are motivated to write about the text in like manner, and through this repetition form the collocations assumed to be markers of genre? I suspect the answer to each is a qualified 'yes.' This form is overdetermined institutionally and textually, and consequently it would be difficult to call this form of collocation—meaningful as it might be—'communicative' in the same way as the collocations presented above. Critics from similar places, engaged in similar practices, respond to fiction in similar ways. None of this should be surprising, but it does, of course, trouble the question of how we know genre: is the fact that we attribute trios to the public school

story an experience with genre, or an epiphenomenon mediated by critics? (If so, what we experience as genre would actually be what Janet Giltrow [2002] calls meta-genre, mentioned in the introduction, and why I can only say the feature I have identified is 'reminiscent' of genre.) And even if such critical activity does instigate some sort of experience, we ought to consider that a *constellation* of such critical activity is necessary: if critics do indeed respond in similar ways they certainly *do not* write in the same way, as the following three plots, generated using "Ralph" as a seed term and taken from three of the longest essays in the corpus, show. It is only when the criticism is aggregated as one body, taken *en masse*, that we observe the phenomena depicted above.

[28 – OnSym]:

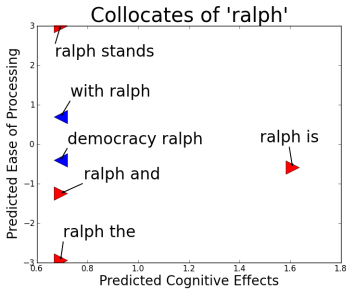

Figure 4.10: Relevance plot of "ralph" in a single essay-1

[40 – NatB]:

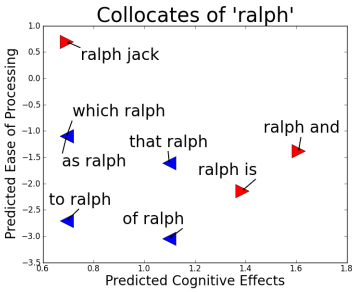

Figure 4.11: Relevance plot of "ralph" in a single essay-2

[25 – Ant]:

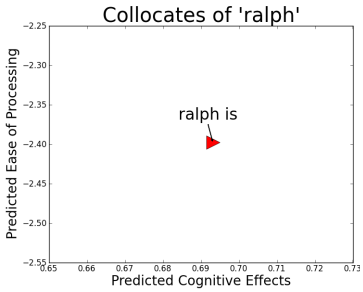

Figure 4.12: Relevance plot of "ralph" in a single essay-3

4.5 The Linearity of Language: Or, The Devil's Advocate

At this point, a caveat is warranted, as well as a reflection on the linearity of the English language and its impact on methodology in studies of collocation. In the previous section I presented what I believe to be strong evidence against the conventionalization of collocations' meanings: since "human nature" appears as an echoic utterance, a second order interpretation, its meaning cannot be conventionalized as regular and stable – it is inherently volatile as an utterance that primarily depends on

intertext, presupposition *and* the speaker's orientation around these. However, to make this argument I claimed that "human nature" appears in the vicinity of a certain feature. In the present section I question this mode of argument, turning from collocation as I have been using it—immediate co-occurrence of terms—to co-occurrence of terms across spans of text.

Here, I am reminded of Alan Partington's (1998) comment on textual collocation, collocation as "a consequence of the linearity of language, or, conversely, if we view text as a process rather than product, it is the principal method, together with syntax, with which this linearity is constructed" (p. 15). The implication is that in English, any *n-gram* has a certain spatial relationship to any other *n-gram* through *x* number of terms: in other words, *all* terms have a spatial relationship with one another and therefore what *is* the category of collocation—as a 'special', or 'interesting' category—of co-occurrence? Closeness? Meaning? Frequency? We have seen some answers to these questions and different perspectives in chapter two, but here I will present data from an experiment in term co-occurrence. The gist of my argument is that although this work is based on the 'interesting,' unique properties of collocation—this is to say, meaningful properties—what we find is that these relationships are structured, orderly, and predictable. This brief comment will note quantitative observations and potential drawbacks of certain methodologies in studies of term co-occurrence.

In the preceding section I presented what I believe to be the strongest piece of evidence supporting an argument against the semantic standardization of collocations' meanings: a large number of occurrences of 'human nature' were attributed as another's speech / thought. In order to bolster my claims of proximity, I set out to graphically

represent one term's extended textual proximity to another term. Rather than contend with the complexity of reporting verbs, I began with a search using the negating particle "not." To see how many times the term "not" occurs around "human," I wrote a Python script[42] to count occurrences within a range of 'word windows.' That is, it begins by counting occurrences of "human" within a very large span of text (500 words, for example), and then for each smaller window until it gets to a window of one (where the occurrence of "human" and "not" within a one word window, or one word either side of the seed term, "not", would indicate co-occurrence or collocation). Frequency of occurrence is represented as a percentage of the total occurrences of the term.

My hypothesis was that I would find a clustering tendency: a graph that 'peaks' or otherwise has distinctive trends where terms cluster around one another – but not the following (Figure 4.13):

[42] The concept for this script is rather simple, but implementing an optimized search algorithm was not: my thanks to my colleague Nigel Meyers for developing an efficient list comparison algorithm.

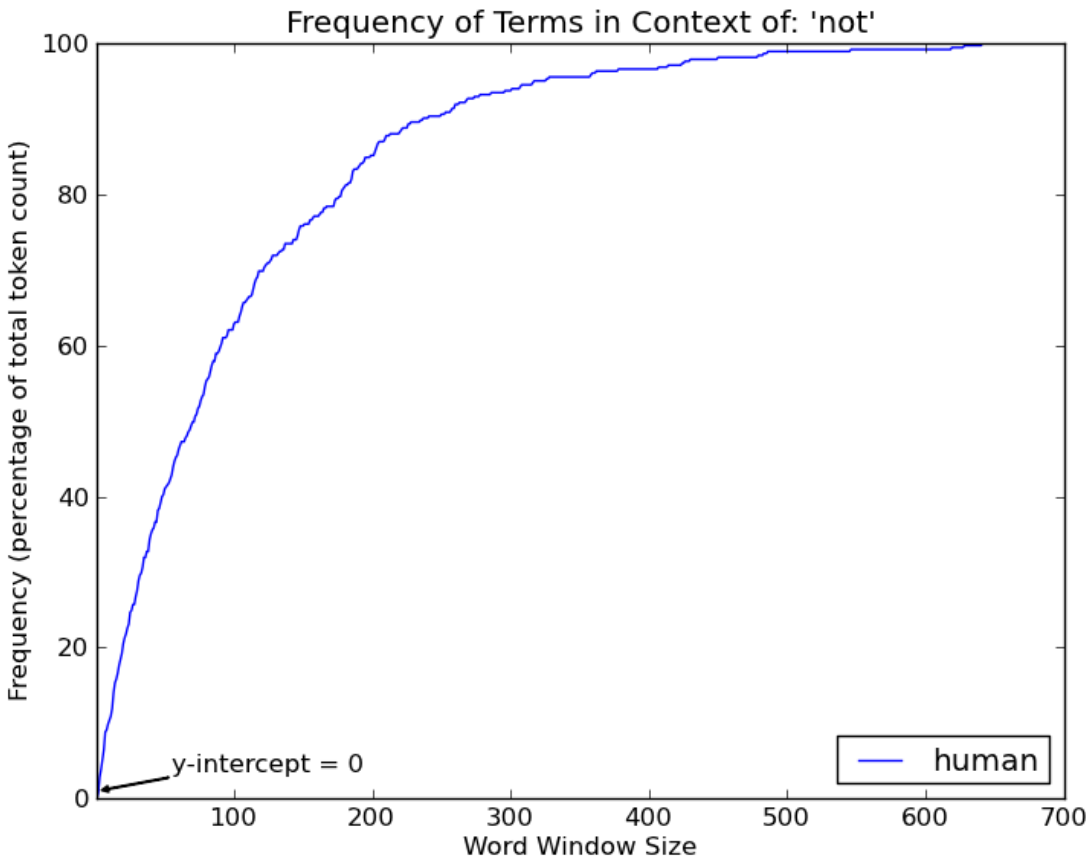

Figure 4.13: Frequency of "human" in context of "not"

Figure 4.13 chart plots the frequency of "human" in the context of "not" within a word window range of 1 to 700. All, or 100%, of the occurrences of "human" can be found at the top end of this range (the precise upper limit, where all occurrences of "human" are found in the context of "not," is a window of 640 words) and no occurrences of "human" are immediately adjacent to (within a 1 word window of) "not."

This was somewhat surprising to me, especially since a collocation we know to be quite strong ("human" and "nature") produced a similar distribution (Figure 4.14):

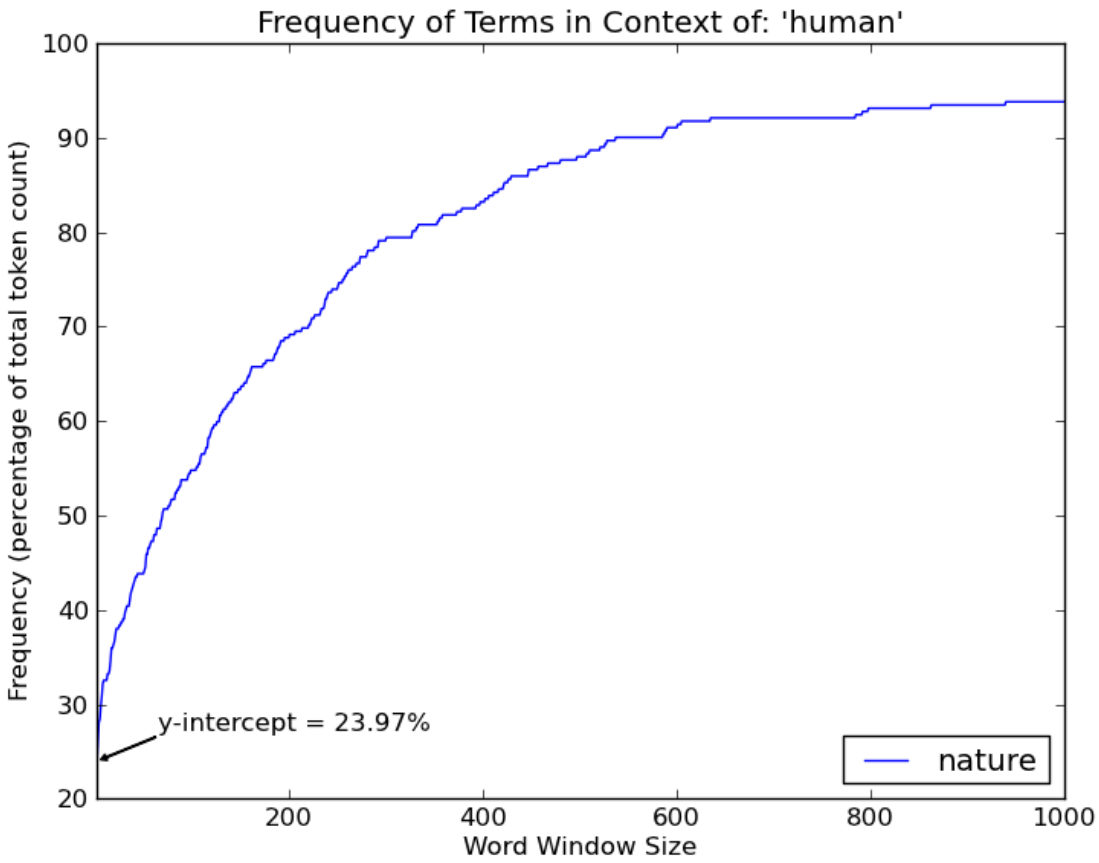

Figure 4.14: Frequency of "nature" in context of "human"

In Figure 4.14, the top-end of the chart does not include all occurrences of "nature" because there are criticisms in the corpus where "nature" appears and "human" does not, and vice versa. The two charts' overall shapes are obviously very similar, and the greater slope (the 'steepness') of the first chart might be simply attributed to a larger number of occurrences of "not" than of "human" (992 versus 385). Given the same seed term (i.e. 'not' or 'human'), the shapes of the charts are very similar.

In fact, even taking five random terms from the corpus[43] and comparing these terms with a frequently occurring term—"not", again—produces very similar distributions.

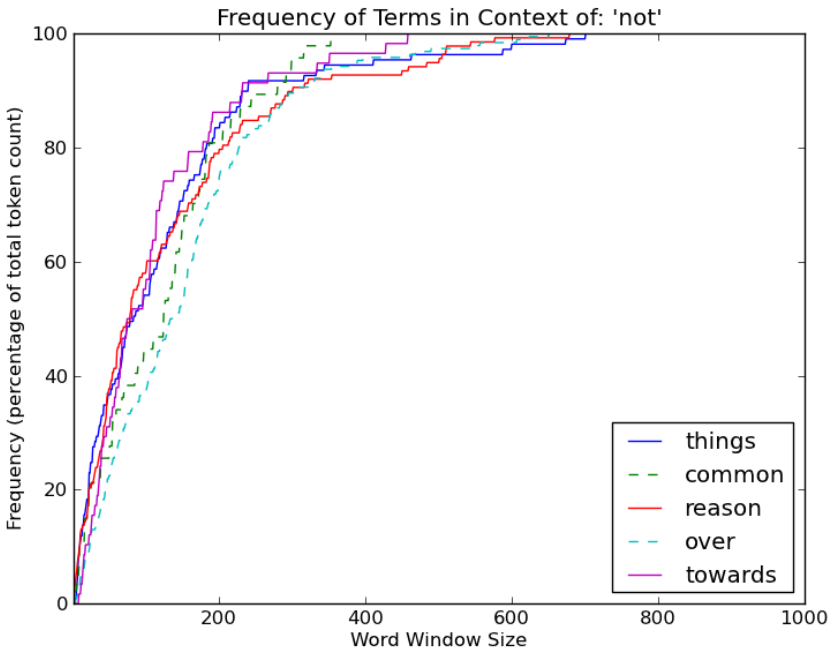

Figure 4.15: Frequency of randomly selected terms in context of "not"-1

[43] Randomly selected terms qualifying for evaluation had to also appear in every document in the corpus.

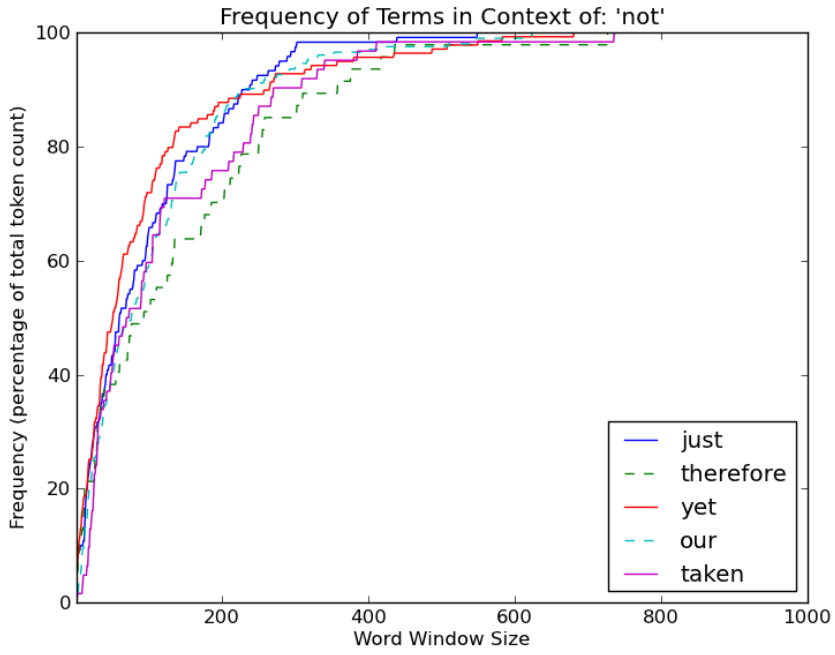

Figure 4.16: Frequency of randomly selected terms in context of "not"-2

These distributions suggest not design but natural, distributed formations: language composed of universal laws. Like Zipf's law (that states a list of ordered term frequencies rendered from a sample of natural language will decrease on a logarithmic scale) this distribution shows the natural distribution of terms around other terms (but distribution in space, as opposed to distribution of magnitudes). These plots occur because the terms searched are distributed throughout the corpus: "not" occurs 992 times, and is present in every file in the corpus. As we can see from Figure 4.16, approximately 80% of randomly selected terms occur within ~180 words of "not," and the long 'tail' of the graph implies that fewer terms occur within this window as the window increases. More could be said using better statistics, probability testing, representations with scatterplots and regression lines, etc., but I will spare the reader: for now, I will make a guess concerning the implications of these data and collocation in general.

The curved distributions for each term approximate one another closely, following a natural distribution that appears patterned. What is not patterned, however, is the y-intercept, the 'originating point' of the graph where the line intersects with the ordinate. As I mentioned previously, this area is interesting because this is where collocation—in the sense of strict co-occurrence—is indicated: the 'word window' here is only 1, and therefore any term appearing in the context of another term, here, is its immediate collocate. The y-intercept of the "human" vs. "nature" plot is ~24% – and 24% of 300 is 71, our count for "human nature" in the corpus. But interestingly, the y-intercept appears to have little effect on the rest of the graph: just because a certain term's frequency 'starts high' does not mean its shape will be substantively different from a term that collocates less frequently with the seed term (the y-intercept does not appear to affect the slope). From all of this, I reach a tentative conclusion: the collocability of one term with another has little effect on its collocability in an extended sense (i.e. across a span of text). Thus, in an extended sense of collocation (one which we have avoided), purely statistical measures using word count frequencies seem to be uninformative since the 'trajectory' of such terms is so patterned. All of this is based on a simple fact of English and its linearity: one word comes after or before another.

On the one hand this makes intuitive sense: 'human nature' is just very different from 'nature […] human.' "Human" does not 'prime' the text for the co-occurrence of "nature" – it will appear around that term, based on frequency. But on the other it suggests—as just one piece of evidence, a small but pointed part of the picture—that term co-occurrence as one type and extended collocation as another type are irreconcilable as categories based on quantified frequencies. At the very least this might

serve to caution claims based on term co-occurrence across spans of text. In the last section I claimed that "human nature" tends to occur around another feature in text. In that case it was a pragmatic feature—representations of speech and thought—and not lexical items as in this present discussion, but in general we might be wary of claims based on term co-occurrence as a special category when it appears to be, as Partington mentions, a consequence of the linearity of language. Of course, this warning against statistical sampling is undermined by a like flaw – a lack of rigorous statistical sampling in this comment. But I hope here to illustrate the limits and liabilities of such claims, and even call into question what we can really say about one n-gram appearing within *x* words of another. This feature *is* likely meaningful, but the ways and means of establishing and communicating this meaning ought to be carefully considered.

5 Conclusion

'Collocation' in this work has been used loosely, purposefully avoiding a pursuit of this term's definition and instead focusing on the several implications (implicatures) of term co-occurrence. But in this light, it has also been used restrictively: with the exception of the latter parts of the final chapter, only terms that occurred immediately adjacent to one other have been investigated – these are similar to the lexical bundles of Biber, rather than collocation in an extended sense as used by Halliday. But 'collocation' has still held somewhat of a special meaning. Overall, the lexicon used to denote term co-occurrence has been small, using 'n-gram' or 'chunk' when little more has been meant than one word appearing before or after another, and reserving 'collocation' for one or more words where their very occurrence together is exceptional.

So what is this special meaning? In terms of Relevance Theory, it would seem that collocation achieves Relevance by drawing attention to (ostensifying) a phrase that *has* been said, *could have* been said, or *will* be said, by someone else. In fact, more often than not it is a hypothetical utterance, and therefore its meaning is derived not from its content but its attribution. In this sense it is an echo, an echoic utterance that apprehends the situation of its elocution and frames (makes manifest) the speaker's voice within this context.

This has consequences for the interface between collocation, invention, and banality, first, because a collocation cannot *be* a collocation without repetition and reformulation. Lack of reformulation would render it susceptible to banality as a very shallow, specious repetition. As an ostensive utterance, collocation constructs context and therefore sets the stage for innovation: a collocation is first recognized as a *formal*

repetition, but its singularity lies in recognizing but breaking with its history. This is not a general poststructural comment on the instability of referential meaning, but a very specific note on the *type* of attitude made manifest by a collocation: in short, the attitude is essentially ironic. A collocation that too heavily subscribes to its past (or potential) context, or might be deemed to naively recreate (repeat) such contexts, might be deemed 'banal' owing not to a specific recreation but the sheer scope of potential recreations. In the imagery of echoes, this is what I mean by echoes echoing diffusely – "human nature," for example, might be said to resonate too weakly, too widely.

But this type of collocation—the ostensive type—does not mean that all pairings of words are simply intentional, and by consequence of this intent meaningful. The 'character name + *and*', for example, cannot be said to convey meaning by the intentional pairing of a proper noun and a conjunction for the simple reason that not every pairing of these terms *could be* intentional. This particular collocation is at once banal and striking: banal, for obvious reasons of the necessities of syntax, but striking because the pairing of these character names resonates well with the genre of which the criticized text—*Lord of the Flies*—is a part. *LOTF* criticisms invoke trios of characters as do other texts of the public school story. I have supposed that the 'character name + *and*' collocations *are* traces of a genre, and that their pairings are not intentional but the result of multiple voices, taken together, speaking from similar backgrounds and contexts.

However, ultimately, perhaps these two types of collocation are not so different. First, 'ostensive' does not precisely mean 'intentional,' but that there is a consciousness manifest behind the utterance and directing its interpretation. In the case of the non-

intentional collocations of 'character name + *and*' there is very clearly a motive for the critic concatenating characters with other characters in their arguments. In the case of *Lord of the Flies*, if I might take part in some literary criticism myself, this motive would seem to be the tacit recognition of the rhetorical pressures exerted in, specifically, trios of children: such groups are inherently unstable since—aside from pure consensus—any dispute will result in two people ganging up on one, and it is this rhetorical 'leveraging' type action that is too easily exploited, too heavily favoured by rampant liberalism and its privilege of the 'majority,' that is responsible for the conflict described by the text.[44] To be clear, though critics do not say these things explicitly—perhaps because they are too obvious, or are simply tangential to the critics' argumentative purpose—they are present, observable, and therefore purposeful. The implicatures to be derived from such a collocation are therefore weak and many, and certainly have different properties from more heavily ostensified collocations (hence Sperber and Wilson's category of "poetic effects"), but are not of a completely different type than the collocations described above.

Chapter two presented a short history of 'collocation' and suggested that the two dominant views—the statistical view, and semantic view—are irreconcilable perspectives and, further, defy what Firth intended as a method for analysis based on form. My work on discovering the language formulas in the corpus, building on O'Donell's (2011) adjusted frequency list, led to what I have been calling a 'saliency metric.' This metric is based on dividing the logged frequency of a term by the number of top-level chunks in which the term appears: *log(f)* / #chunks. The result is intriguing, since, when terms are rank-ordered by this metric, function words fall to the bottom and the corpus' 'salient'

[44] The critics of the corpus would likely not describe this behaviour in rhetorical terms, and would further explain this as fundamental to, of course, human nature. I would never make these claims.

terms rise to the top – the output is so clear-cut it is almost as if a regular ordered frequency list is sorted according to grammatical function. But only almost: mixed into the function words at the bottom of the list, we do find our typically 'banal' terms, like 'human,' since it does occur frequently but also collocates widely, in many chunks. In terms of raw computation, this might be the closest measurement found in this investigation that approximates banality.

Chapter three built upon this metric—the conceptual basis of a term's scope of association—and mated this with Relevance Theory and Howard White's (2007a, 2007b) use of the theory in concert with bibliometrics. I emphasized RT's lack of use, glossed its flaws, and suggested that its biggest flaw is precisely its lack of use. My use of the theory took the form of a computer program that inputs a seed term, reads through a corpus, and represents the seed term's collocates based on my further adaptation of White's work: the collocates' position on the x-axis is determined by the frequency of occurrence in the corpus, and position on the y-axis by the number of other top-level chunks in which the term appears (based on the saliency metric from chapter two).

Chapter four presented an analysis of the corpus using Relevance Theory and these tools, and resulted in the conclusions regarding collocations presented above. These conclusions, of course, are dependent on a range of suppositions, such as the saliency of terms, and the two axes along which Relevance is predicted for corpus collocations: ease of processing and cognitive effects. Though I endorse these as legitimate uses, I also propose further research, some of which should be experimental, to test the psychological plausibility of these assumptions.

All chapters run a gamut of analyses, sliding from literary to computational methodologies. Through this formal arrangement I have hoped to convey not only the benefit of employing both qualitative and quantitative methodologies, but also that sometimes—when investigating literature, for example—different methodologies might be *counterproductive*, working at cross purposes. Sometimes having 'more perspectives' is *not* helpful, since not every perspective is appropriate. The literary works briefly described certainly imply so: each presents the problem of criticism, that being the mastery of the work over the critic, but none present the solution. My solution is a reading practice that does not consume the text it critiques. To enable this solution, a significant portion of this work has been devoted to the development and deployment of text tools to analyze a collection of criticisms. This work is, of course, historically and disciplinarily situated, and as such is also subjected to trends and trajectories of research. I would like to conclude by making an observation on this trend.

Current trends in computer-aided language investigations favour analyses of larger sets of data – with the potential product of 'larger' conclusions. 'Culturomics,' for example, is a neologism coined to describe the historical investigation of culture through massive collections of language: in 2010 a multi-disciplinary, multi-institutional group published research using "a corpus of digitized texts containing about 4% of all books ever printed" ("Quantitative Analysis of Culture Using Millions of Digitized Books", *Science*, 2010). The lab responsible for this work is dubbed the "Cultural Observatory" – as an observatory, their work uses heavy-duty equipment to study large, complex, and ever-expanding sets of data. Though to a lesser extent, corpora in linguistics are also favoured when they are large: *COCA* advertises itself as "the largest freely-available

corpus of English, and the only large and balanced corpus of American English" (Davies, 2011). These methodological concerns contrast with mine, of course, in many respects, and most obviously in scope. Not only is my corpus comparatively small (200,000 words) but so is my object of analysis: it is almost atomic, in that I am interested in the micro-forces attracting one word to another. Further, this work is concerned with the differences between the single author (critic) and the dialogic ramifications of a corpus – univocal production is by necessity smaller than polyvocal.

However, I suggest that an analysis of smaller data does not necessitate 'smaller' tools: my work, too, uses heavy-duty equipment. Not to hyperbolize, but instead of a telescope we have built here an atom smasher. My claim is this: the complexity of the tools lies in exploring the intersubjectivity that drives the very production of an utterance – trickier to do in a small corpus, but worthy. Larger corpora exhibit this dialogism, but as only a circumstantial associate of their size. In fact, I would suggest that the biggest reason—and perhaps the only reason—for studying large quantities of data is to capture dialogic interaction. In *LOTF* criticisms, multiple authors speak with and against one another, using phrases mutually recognizable as the words of others. The single author, too, imports a chorus of voices by repeating formulas. As a dialogic phenomenon, the tacit apprehension of another's language, collocation is a molecular-scale offering to this large problematic.

References

Anthony, L. (2011). *AntConc* (Version 3.2.4m) [Software]. Available from http://www.antlab.sci.waseda.ac.jp/antconc_index.htmltp://athel.com/product_info.php?products_id=80

Bach, K. (2006). Impliciture vs. explicature: What's the difference? Online Publication. Retrieved from: http://userwww.sfsu.edu/~kbach/Bach.ImplExpl.pdf

Bakhtin, M. M. (1981). *The dialogic imagination*. (C. Emerson & M. Holquist, Trans.). M. Holquist (Ed.). Austin, TX: U of Texas P.

Bartsch, S. (2004). *Structural and functional properties of collocations in English*. Tübingen, Germany: Narr.

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

--- and Conrad, S. (1999.) Lexical bundles in conversation and academic prose. In H. Hasselgård and S. Oksefjell (Ed.) *Out of corpora: Studies in honour of Stig Johansson* (pp. 181–189). Amsterdam: Rodopi.

Blakemore, D., and Carston, R. (2005). The pragmatics of sentential coordination. *Lingua 115*(4): 569-589.

Burke, K. (1923). Engineering with words. *The Dial 74*(4): 408-412.

Card, O. S. (1984). *Ender's Game*. New York, NY: Starscape.

Carston, R. (2002). *Thoughts and utterances: The pragmatics of explicit communication*. Oxford: Blackwell.

Carter, R. (1987). *Vocabulary: Applied linguistic perspectives*. New York, NY: Routledge.

Cohen, R. (1986). History and genre. *Neohelicon 13*(2): 87-105.

Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3: 43-57. doi: 10.3366/E1749503208000063

---. (2004). Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes* 23: 397–423.

Danielewski, M. (2000). *House of leaves*. New York, NY: Pantheon Books.

Derrida, J. (2007). Psyche: Invention of the other. In P. Kamuf and E. Rottenberg (Ed.), *Psyche: Inventions of the Other* (Vol. 1) (pp. 1-47). Stanford, CA: Stanford UP.

---. (1988). *Limited inc*. (S. Weber & J. Mehlman, Trans.). Evanston, IL: Northwestern UP.

Davies, M. (Online Resource). *Corpus of Contemporary American English: 425 million words, 1990-2011*. Retrieved from: http://corpus.byu.edu/coca/

Davies, M. (Online Resource). *Corpus of Historical American English: 400 million words, 1810-2009*. Retrieved from: http://corpus.byu.edu/coha/

Dooley, R. A. (2008). Relevance theory and discourse analysis: Complementary approaches for translator training. *GIALens 2*(3): 1-11.

Fernando, C. (1996). *Idioms and idiomaticity*. Oxford, UK: Oxford University Press.

Firth, J. R. (1951 [1957a]). The Technique of Semantics. *Transactions of the Philological Society*. Reprinted in J. Firth. *Papers in linguistics 1934-1951* (pp. 7-33). London: Oxford University Press.

---. (1951 [1957b]). Modes of Meaning. *Essays and Studies*. Reprinted in J. Firth. *Papers in linguistics 1934-1951* (pp. 190-215). London: Oxford UP.

---. (1957 [1968]) A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis*

(Special Vol.), Philological Society. Reprinted in F. Palmer (Ed.). *Selected papers of J. R. Firth 1952-59* (pp. 168-205). Bloomington & London: Indiana UP.

George, W. (1926). *Children of the morning*. London: Chapman and Hall.

Giltrow, J., and Stein, D. (2009). Preface. In J. Giltrow & D. Stein (Ed.), *Genres in the internet* (pp. 1-26). Philadelphia, PA: John Benjamins.

---. (2002). Meta-genre. In R. Coe, L. Lingard, & T. Teslenko (Eds.), *The rhetoric and ideology of genre: Strategies for stability and change* (pp. 187-205) Cresskill, NJ: Hampton Press.

Goatly, A. (1994). Register and the redemption of Relevance Theory: The case of metaphor. *Pragmatics 4*(2): 139-181.

Golding, W. (1984). *The paper men*. London: Faber and Faber.

---. (1961). *The inheritors*. London: Faber and Faber.

---. (1954). *Lord of the flies*. New York: Faber and Faber.

Götz-Votteler, K., & Herbst, T. (2007). Introduction: The mystery of collocation. *Zeitschrift für Anglistik und Amerikanistik 55*(3): 211-215.

Gutt, E. A. (1990). A theoretical account of translation – without a translation theory. *Target* 2: 135-164.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Essex: Pearson Education.

Herbst, T. (1996). What are collocations: Sandy beaches or false teeth? *English Studies 77*(4): 379-393.

Hickey, T. (1993). Identifying formulas in first language acquisition. *Journal of Child Language* 20: 27-41.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 47-69). London: Language Teaching Publications.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford UP.

Howarth, P. (1998.) Phraseology and second language proficiency. *Applied Linguistics 19*(1): 24-44.

Hughes, T. (1857). *Tom Brown's Schooldays*. London: Macmillan and Co.

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

---. (2000). 'It might be suggested that . . .': academic hedging and student writing. *Australian Review of Applied Linguistics* 16: 83–97.

Kipling, R. (1919). *Stalky & Co.* London: Macmillan and Co.

Kjellmer, G. (1991). A mint of phrases. In K. Aijmer and B. Altenberg (Eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik* (pp. 111–27). London: Longman.

Leech, G. (1974). *Semantics*. Hammondsworth, UK: Penguin.

Martínez, A. S. (2010). Collocation analysis of a sample corpus using some statistical measures: An empirical approach [Supplemental material]. *Escuela Oficial de Idiomas, Murcia*. Retrieved from http://www.um.es/lacell/aesla/contenido/pdf/6/sanchez3.pdf

Mateo, J. (2009). Contrasting relevance in poetry translation. *Perspectives: Studies in Translatology 17*(1): 1-14.

Mey, J. L., & Talbot, M. (1988) Computation and the soul. In A. Kasher (Ed.), *Cognitive Aspects of Language Use* (pp. 239-285). Amsterdam: Elsevier.

Meyer, C. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge UP.

Miller, C. (1984). Genre as social action. *Quarterly Journal of Speech, 70*(2): 151-167.

Moon, R. (1998.) *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Clarendon Press.

Nabokov, V. (1962). *Pale fire*. New York: NY: G.P. Putnam's Sons.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

O'Donnell, M. (2011, June). *The adjusted frequency list*. Paper presented at the International Society for the Linguistics of English (ISLE2) Conference, Boston, MA.

O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge UP.

Olsen, K. (2000). *Understanding* Lord of the Flies: *A Student Casebook to Issues, Sources, and Historical Documents.* Westport, CT: Greenwood Press.

Oxford English Dictionary [webpage]. (2010). Retrieved from http://www.oed.com

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins.

Pilkington, A. (2000). *Poetic effects. A Relevance Theory perspective*. Amsterdam: John Benjamins.

Routledge Dictionary of Language and Linguistics. (1998). Gregory Trauth (Trans. and Ed.). London: Routledge.

Rowling, J. K. (1997). *Harry Potter and the Philosopher's Stone*. London: Bloomsbury.

Schmid, H. J. (2007). Non-compositionality and emergent meaning of lexico-grammatical chunks: A corpus study of noun phrases with sentential complements as constructions. *Zeitschrift für Anglistik und Amerikanistik 55*(3): 313-340.

Schneider, E. W. (2003). The dynamics of New Englishes: From identity construction to dialect birth. *Language 79*: 233-281.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

---. (1991). *Corpus, concordance, collocation*. Oxford: Oxford UP.

Siskin, C. (1998). *The Work of Writing: Literature and Social Change in Britain, 1700-1830*. Baltimore: Johns Hopkins UP.

Sperber, D., and Wilson, D. (2005a). Relevance theory. In L. Horn and G. Ward (Eds.), *The handbook of pragmatics* (pp. 607-632). Maiden, MA: Blackwell.

--- and ---. (2005b). Pragmatics. *UCL Working Papers in Linguistics* 17: 353-388.

--- and ---. (2002). Pragmatics, modularity and mind-reading. *Mind and Language 17*(1-2): 3-23.

--- and ---. (1986/1995). *Relevance: Communication & Cognition*. Cambridge, MA: Blackwell.

Swinden, P. (1987). [Review of the book William Golding: Novels, 1954-67: Lord of the flies; The inheritors; Pincher Martin; Free fall; The spire; The pyramid Ed. N. Page]. *Notes and Queries*, 4: 570-572.

Thiher, A. (1997). *The power of tautology*. Madison, NJ: Fairleigh Dickinson UP.

Townsend, R. C. (1964). *Lord of the Flies*: Fool's gold? *The Journal of General Education 16*(2): 153-160.

Uchida, S. (1998). Text and relevance. In R. Carston & S. Uchida (Ed.). *Relevance Theory: Applications and Implications* (pp. 161-178). Amsterdam: John Benjamins.

Waugh, A. (1983, December 10). Tale of two authors. *The Spectator*, p. 6.

White, H. D. (2007a). Combining Bibliometrics, Information Retrieval, and Relevance Theory, Part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology, 58*(4): 536-559.

---. (2007b). Combining Bibliometrics, Information Retrieval, and Relevance Theory, Part 2: Some implications for information science. *Journal of the American Society for Information Science and Technology, 58*(4): 583-605.

Wikipedia [webpage]. (2010). Collocation. Retrieved from http://en.wikipedia.org/wiki/Collocation

Wittgenstein, L. (2009 [1953]). *Philosophical investigations*. (G.E.M. Anscombe, P.M.S. Hacker, and J. Schulte Trans.). West Sussex, UK: Wiley-Blackwell.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, MA: Cambridge UP.

---, and Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication* 20: 1-28.

Xiao-chun, G. (2007). Lord of the flies: A survey of evil humanity. *Sino-US English Teaching, 4*(12): 61-65.

Yus, F. (September 15, 2011). Relevance Theory online bibliographic service [Webpage]. Retrieved from http://www.ua.es/personal/francisco.yus/rt.html

---. (1998). A decade of relevance theory. *Journal of Pragmatics 30*: 305-345.

Zipf, G. K. (1965). *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.

Appendices

Appendix A: Corpus Files

Number	Name	Year	Word Count
1	Disquieting Story	1954	1806
2	Significant Motifs	1954	1526
3	Fiction or Fable	1957	2627
4	Assessing LOTF	1960	2231
5	Lord of the Flies--Cox	1960	2859
6	Modern Allegory	1960	3008
7	Coral Island Revisited	1961	2779
8	Smaller Growth	1961	5052
9	Bacchae	1964	915
10	As Fable	1965	2837
11	Beelzebub Revisited	1965	5542
12	Beelzebub	1965	6996
13	Impossible Categorize	1965	1794
14	Meaning of Beast	1965	2934
15	Obscure Setting	1965	1483
16	Unwarranted Popularity	1965	2076
17	Classical Themes	1966	1469
18	Lord of the Flies	1967	19628
19	Irony in LOTF	1968	3829
20	Several Interpretations	1968	3371
21	Metaphor of Darkness	1969	6264
22	Lord of the Flies2	1970	7748
23	Ranking LOTF	1970	1965
24	Mythical Elements	1971	4094
25	Questioning the Merit	1972	2671
26	Rhythm and Expansion	1978	4822
27	Myth Fable Fiction	1980	4908
28	Explicator2	1983	937
29	Law and Order	1983	2824
30	Resolution of Antithesis	1984	4043
31	Boys Accurate	1986	1812
32	Doorways Through Walls	1986	8006
33	Beelzebub's Boys	1988	10685
34	Fictional Explosion	1988	7003
35	Nature of the Beast	1988	8608
36	Lord of the Flies3	1990	6040
37	Pride as Original Sin	1992	5280
38	A Suggested Reference	1993	421
39	Christian Interp	1993	3590
40	Grief Grief Grief	1993	4827

Number	Name	Year	Word Count
41	Government of Boys	1997	3272
42	Mature World	1997	6282
43	Explicator	1999	791
44	Golding and Huxley	2000	6778
45	Desert Island Reading	2005	4656
46	On Symbolic Significance	2009	3213