PHRASEOLOGY AND EPISTEMOLOGY IN SCIENTIFIC WRITING: A CORPUS-DRIVEN APPROACH

BY

Garry Lee Plappert

A thesis submitted to The University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

Department of English

School of English, Drama, Canadian and American Studies

July 2012

Abstract

This thesis uses the tools and methods of corpus linguistics to study the process of knowledge encoding in a corpus of texts from the scientific discipline of genetics. It is argued here that the approach taken fits into the tradition of corpus-driven approaches to linguistic questions in that no assumption is made about the linguistic form that this knowledge encoding will take. Instead the study proceeds by identifying a set of keywords using the concept of lexical chains to identify items of terminology. The investigation of these uses the cluster function of *WordSmith Tools* (Scott 2004) and is qualitative, following Sinclair (1991; 2004) in attempting to develop a picture of the typical linguistic nature of the patterns surrounding these clusters inductively through a process of studying collocation and colligation patterns and identifying phraseology. It is argued here that such an approach is required to discover linguistic aspects of epistemic encoding that have as yet not been identified by those working in the related fields of discourse analysis or corpus linguistics.

Table of Contents

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Introduction

This thesis considers the social construction of knowledge in scientific texts. In this introductory chapter an outline of the intellectual context for this research will be provided. Initially the principal research problem(s) will be outlined (1.2), setting out the initial purpose of the study by stating the key questions that this study will seek to answer. Secondly, a justification for the research will be provided (1.3), with a brief adumbration of recent related literature on scientific texts allowing and an outline of gaps within that literature that this thesis will attempt to fill.

Thirdly, an outline of the following chapters will be provided (1.4), giving an overview of the connections between various related disciplines (the study of epistemology in philosophy, social epistemology, previous linguistic study of scientific writing, corpus linguistics), and the process of obtaining data and the methodology used to investigate that data.

Fourthly, (1.5) some of the key concepts that this research relies upon will be set out, providing a brief theoretical context for this study that will be discussed in more detail in chapters 2 and 3.

Finally, (1.6) the key assumptions and limitations of this study are anticipated, since an awareness of these is crucial in providing the correct intellectual context for the study, thereby allowing for the correct interpretation of the scholarly contribution made by this thesis.

1.2 Stating the research problem(s)

This research explores the social construction of scientific knowledge. However, the social construction of knowledge is a multi-faceted theory, involving a range of theoretical and empirical claims relating to different aspects of that theory. The starting point of this thesis is that the theory of the social construction of knowledge, if it is a valid one, might feasibly be expected to entail certain textual features, and furthermore that this texturing of the purported social aspect of knowledge might be amenable to linguistic analysis. Indeed, a considerable body of research that makes just this assumption already exists (discussed in sections 3.2 and 3.5 below), and a number of linguists have proceeded to investigate scientific texts in order to shed light upon the means by which the social construction of knowledge is realised textually. In response to this body of work, an urgent initial research question can be phrased as follows: to what extent can the further linguistic study of the social construction of scientific knowledge contribute anything in addition to what has already been discovered?

It will be suggested in what follows that there are answers to this first challenge to this study related to each of the empirical, theoretical and methodological aspects of the social construction of knowledge: there are specific areas of knowledge construction that are worthy of study in their own right, there are aspects of the theory of the social construction of knowledge that are underrepresented in the linguistic literature thus far, and there are methodologies in the study of language that have as yet not been applied systematically to this particular area.

Each of these partial and thematic answers to the initial question of how this study can contribute to the linguistic study of the social construction of scientific knowledge raises a number of related research questions that this study could address. For instance: which of the various aspects of the social construction of knowledge can most plausibly be expected to be revealed textually? What is the relationship between the encoding of knowledge across different genres of 'scientific' writing? To what extent does the further study of these features support or even falsify what has been claimed before in previous studies of this type?

From these possibilities the key question that this research seeks to answer is methodological in nature and can most simply expressed by asking the following related questions: to what extent can a corpus perspective improve upon previous linguistic analyses? What method can be proposed to investigate the linguistic nature of epistemic signalling in genetics? Will corpus methodology alone suffice to make progress in the rigorous analysis of scientific texts or ought one to take a broader approach, combining complementary methods and taking a wider theoretical perspective? Each of these related issues will contribute to solving the key research question of how the linguistic study of the social construction of knowledge can benefit from the use of corpora.

What must be admitted from the outset is that these questions are embedded within a certain theoretical perspective towards the study of language: one which proceeds upon the belief that language study of this kind must be primarily and rigorously empirical; that the analyst though competent in interpreting the data is in no privileged position in respect of that data, and must work towards as full a description of the data as is possible; and that the understanding of the analyst must develop in

accordance with what is discovered, rather than imposing *a priori* assumptions on the data and analysing language merely in an attempt to corroborate those assumptions. In order to proceed within this paradigm two corpora were created that it was hoped would constitute a firm empirical basis for the study, and these were intended to be as representative as possible of the phenomena that are to be studied. A pilot corpus, *genepilot*, comprising a number of texts listed on the Human Genome Project website as key texts in the field was used to explore potential methods for the study, and this process is discussed in detail below (section 4.2.1). The final corpus for the study, *genecorp*, was then created using 2,979 texts from the journal *Nature Genetics* (this process is discussed in detail in section 4.3 below) for the purpose of providing sufficient data for the identification and study of the linguistic nature of knowledge encoding in genetics.

The creation of these corpora unsurprisingly led to further interrelated research questions: Would a corpus investigation of the linguistic features of the social construction of knowledge corroborate the more local claims made by Hunston (1989;1993;1994), Myers (1989;1990;1991;1992;1994), Hyland (1998) and others? Which of the many current approaches to corpus investigation would prove most fruitful in approaching the data for this study? Would a corpus perspective alone be sufficient to develop a rigorous description of the data, or would a more mixed methodology be more productive, for instance a combined text and corpus perspective, acknowledging the importance of the behaviour of language within certain text types in addition to considering linguistic forms across a whole corpus. This thesis aims to explore knowledge encoding in genetics through an exploration of *genecorp* in order to attempt to assess whether those linguistic items typically studied in order to investigate

epistemology (discussed in detail in section 3.5 below) can be added to or revised on the basis of a detailed qualitative corpus investigation. Rather than providing a corpus-based study of features that are already known to be epistemically relevant, such as grammatical modality or modal adjectives, I will explore *genecorp* with a view to identifying new linguistic objects of study; in order to make a global assessment of such devices this study will use the term *epistemic signalling* to cover all of the types of linguistic variation that are epistemically significant, and this notion will be discussed in more detail in section 3.6.2 below.

These theoretical and methodological questions will be explored in greater detail in the main body of this study. At this point this number of smaller and contributory research questions can be organised into two overarching question for the thesis to answer. Firstly, **what method can be proposed to achieve findings about the linguistic nature of epistemic signalling in genetics**? Crucially this question if it is to build upon what is already known from studies of scientific writing by extending our understanding of the linguistic means of epistemic signalling must also answer a second major question: **Can the methodology employed produce findings about the linguistic nature of epistemic marking in genetics that are not wholly predictable**? This thesis must achieve both of these things if it is to be a worthwhile contribution to the field.

1.3 Justifying the research

It is relatively uncontroversial to say that the traditional analysis of the concept of knowledge is almost universally held to be flawed by philosophers, and the reasons for this will be discussed below (2.1, 2.2). Finding the correct replacement for the traditional

model, however, has proved to be one of the most vexing problems in the history of philosophy, and remains a live issue in contemporary study. A range of promising answers to this ancient question that are currently being suggested in disciplines as diverse as sociology, philosophy and psychology takes a descriptive approach to the concept of knowledge. The problem of knowledge is said to arise because certain prescriptive positions are argued to be inherent to knowledge which upon examination are an inadequate description of the actual process of knowledge building in society; the resolution to this problem, it is argued, is to foreground the actual practice of knowledge building and to tailor any theory of knowledge to fit with that description. Moreover, a descriptive approach leads to a more specific and localised perspective on knowledge creation, stressing that what is judged to constitute knowledge will vary across disciplines and contexts and over time, with a more nuanced and less generalised view of how 'knowledge' as an over-arching concept functions, or even whether such an overarching concept can really be said to exist. It is argued here that such speculations provide a timely justification for a localised, descriptive study of the knowledge building practices evident in scientific texts.

A further justification for this study derives from a specific aspect of the theory of the social construction of knowledge: the study of the dissemination of scientific ideas and 'facts' into wider society. Whilst claims concerning what actually is involved in the popularisation of scientific texts have been common in language studies few studies have proceeded by utilising what is perhaps the most convincing methodology for providing a thorough empirical study of a given linguistic feature: corpus linguistics. Whilst previous linguistic analyses have produced plausible claims about the use and function of

particular linguistic features in a few or even just one text, what cannot be so plausible is the move often made to then generalise this claim to broader contexts or even to scientific language in the widest sense, without first establishing whether such a discourse can genuinely be said to exist, and whether, if it does, it can be said to have a nature consistent with the small scale studies carried out. That is not to say that such a claim is false per se, but rather to acknowledge that it is an empirical claim, and a rather significant one. Moreover, as will be illustrated below, one of the most salient criticisms of the social constructionist position is that the theoretical conclusions that are drawn within that theory are based upon scant and unrepresentative evidence. Thus, a rigorous and sustained empirical study of scientific texts provides a context within which the theoretical claims of this position can be more thoroughly considered in terms of the actual empirical entailments generated by that theory.

In terms of the popular dissemination of scientific 'facts', therefore, a clear justification of the research is that it will enable the examination of commonly held empirical claims about the nature of the popularisation of scientific texts. The view that popular texts are mere simplifications of original research ought in principle to be open to empirical verification, and claims that the role of scientific texts is much more complex ought to be verifiable or falsifiable through a process of careful study. Whilst it is not the purpose of this study to contrast popular texts with those found in *Nature Genetics* it is worth noting at the outset that a further potential justification for the study of this process of popularisation is provided by a more nuanced perspective of the likely audience for popular scientific texts, since the audience for popular science is also disputed. Indeed, what might be called the 'naïve' view that popular science is for the 'non-scientific'

general public fails to acknowledge that a significant part of the audience for popular science is actually likely to be scientists themselves, since scientists interest in 'the scientific' in general is unlikely to be entirely sated by the highly specialised and restricted texts they experience within their own particular research field. It might be speculated that texts aimed at scientists, even scientists who are not specialists in the field of any given article, might look quite different from a text aimed at the 'non-scientific' community, and one justification for this study is that in taking a highly specific and localised view of scientific discourse practices it may provide for the further study of the differences between texts aimed at a somewhat scientific readership, and those aimed at the non-scientific general public, through a comparison of texts from research articles, articles found within specialist science magazines and those found in newspapers or magazines with a more general readership.

1.4 Outline of the study

This work is organised into five parts. Part I comprises this introductory chapter, outlining the initial motivation for the thesis and briefly discussing the research questions, justifications and conceptual starting points for the research.

Part II (chapters 2 and 3 comprises a literature review discussing the most relevant aspects of the three fields of study that I attempt to combine in this investigation: the social construction of knowledge, previous linguistic studies of scientific language and corpus linguistics.

In Chapter 2 I discuss the traditional view of the analysis of knowledge in western philosophy, arguing that the Gettier type cases discussed there have wrought a profound

crisis in the traditional normative view of knowledge and one that requires a radical response. Chapter 3 presents such a response, outlining the development of the theory of social epistemology, considering precisely why a 'social constructionist' approach to knowledge, whilst undoubtedly being controversial, is nonetheless an attractive and useful alternative to the traditional prescriptive model. The potentially fruitful relationship between a descriptive view of the study of knowledge and a descriptive view of study of language is then discussed.

Chapter 3 provides an outline of previous linguistic studies that have already contributed to our understanding of the linguistic nature of knowledge signalling in scientific texts, and this is broadly divided into two sections, with approaches that do not employ the methods of Corpus Linguistics discussed in section 3.2 and those that do being discussed in section 3.3. A very brief adumbration of the impetus behind corpus linguistics is then provided, focussing particularly on the theoretical commitments that motivated the development of this approach to language and are inherent in certain of the methodologies that are currently popular within this research paradigm. The 'corpus-based/corpus-driven' distinction is then discussed in some detail, and an explanation is provided as to why in this study it is the 'corpus-driven' approach that is preferred.

In chapter 4 I set out the methodology subsequently employed in this study, both in terms of corpus construction and corpus investigation. The process of corpus construction is presented in detail, and the criteria for choosing texts for the corpus are made clear. The difficulties (both practical and technical) in producing a corpus are acknowledged, and any responses that were deemed necessary in order to ameliorate these difficulties are justified at this point. In the section devoted to corpus investigation

the methodology of corpus interrogation that was chosen is discussed and justified in full, with particular reference to the overall theoretical framework for this study, with any difficulties or limitations inherent in such an approach also being acknowledged at this point.

In chapters 5, 6 and 7 the results of this study are discussed in detail. Chapter 5 presents the data from the cluster analysis of a number of keywords, as described in Chapter 4, allowing for the corpus-driven identification of the linguistic nature of epistemic signalling around these node phrases. In chapters 6 and 7 an attempt is made to build upon the cluster analyses in order to identify two types of epistemic processes found in *Nature Genetics*; the discovery of causative relationships (chapter 6) and the process of ontological categorisation (chapter 7).

Finally in chapter 8 a discussion of the findings of this thesis is provided. The overall conclusions of this study are drawn, and the success of this study in answering the initial research questions outlined in above is assessed. In addition to this, the implications of this study are discussed, this time from the post-research perspective, and some suggestions for potential further exploration of this topic are made.

1.5 Some assumptions and limitations of this research

There are a number of assumptions and limitations related to this study that can and should be acknowledged from the outset in order to place this study accurately and honestly within the correct intellectual framework within which to assess its level of success as an academic enterprise.

Firstly, it is acknowledged from the outset that this study proceeds on the assumption that previous studies of the social construction of knowledge have plausibly indicated that there is a significant linguistic aspect of the social construction of knowledge. Though it might be argued that there would be some scholarly merit in a study that systematically falsified the view that the social construction of knowledge can be revealed textually, such an objective would rule out all of the remaining research questions that this study sets out to answer, and would therefore constitute a very limited achievement when one considers the scope of what was originally intended.

A second, closely related assumption is that the linguistic aspect of the social construction of knowledge is a significant enough element of the overall phenomenon to merit study in its own right. This is stated as being a separate and significant assumption since it might be quite possible, in principle, for the 'social construction' of knowledge to appear in texts to some degree whilst the crucial elements of the process of 'construction' remain as 'extra-discursive' elements. This would, for instance, be the case when texts make crucial use of referencing extra-discursive elements that are not open to inspection in the discourse, or at least not to this researcher. This would in turn appear to be the most threatening limitation of the present study: that (arguably) crucial elements may not form part of the (available) discourse.

Thirdly, this thesis assumes that a worthwhile investigation can proceed on the basis of a researcher lacking in any technical knowledge of the scientific texts in question, and that since what is being studied is the linguistic encoding of the scientific data, rather than that data itself, a specialist scientific analyst is not required. In defending this assumption it is worth noting that a non-scientific analyst may provide certain advantages

over an insider perspective, since certain researcher biases may be avoided, allowing the linguistic features of the texts to be considered in their own right.

Fourthly, this study proceeds on the assumption that a purely descriptive study into the linguistic nature of the data collected will prove worthwhile. Whilst a more evaluative perspective, investigating the relationship between purportedly 'good' and 'bad' examples either of scientific practice, or of the reporting of that practice, might prove fascinating, it is simply outside of the scope of this study to attempt such an analysis.

In terms of the limitations of the present study it must be admitted that a great deal of further investigation of the final corpus *genecorp* would be possible, and that this study merely highlights and discusses a number of the most salient features. Indeed, a key limitation of a corpus linguistics approach (as discussed in more detailed below) is that corpus studies, whilst being very good at picking out the most salient aspects of a body of texts in *statistical* terms may not be similarly good at picking out features of a text which are interesting or important but comparatively rare. This limitation is mitigated somewhat by the initial research outlook and justification for the study having focused upon surveying what typically or commonly occurs in the texts examined, but it must be acknowledged that it is entirely possible that the social construction of science, assuming it exists, could proceed on the basis of occasional but highly significant socially constructed aspects which are then followed by a multitude of procedures that are in some sense more 'objective', and yet rely on and follow on directly from the socially constructed aspect; in such a case the socially constructed element would be rare, but

highly significant, and it might plausibly be suggested that corpus methodology would be unlikely to uncover such phenomena.

In addition to the assumptions outlined above a significant limitation of the present study may be the inability to identify accurately those elements of the construction of scientific knowledge that are omitted textually simply because they are taken as given. Indeed, it has been suggested by some researchers studying the creation of 'facts' in scientific discourse that the final stage in the process of the social construction of a 'fact' occurs when it is no longer deemed necessary to state that fact. This might prove problematic. Given the non-specialist nature of the present researcher it will by no means be obvious when certain 'facts' are being omitted simply because it is assumed that everyone agrees upon them. Although it might be possible to infer this if the development of a fact is traced over time and can be seen no longer to appear in the discourse at some advanced stage any such process of inference will have to proceed very carefully, particularly again bearing in mind the non-specialist status of the present researcher.

Finally it must be noted that the relatively limited size of thesis and constricted amount of time that is allowed to be taken on it constitute clear practical limitations to the present study. Whilst there are undoubtedly further aspects related to this study that merit additional investigation it is simply not possible within the limited scope of a PhD thesis either to carry out further investigation of the data or indeed to report that investigation within the word-limit constraining the length of the written work. However, an outline of some potential further work is provided here, in the section of the concluding chapter entitled 'Recommendations for further study'.

Chapter 2: Epistemology, social epistemology and the philosophy of science

Whilst it is difficult to provide a linear overview of the three interrelated topics of this chapter, they combine to provide a thorough intellectual context for the present study. In this chapter I attempt to set out this context by placing the development of social epistemology firmly within the philosophical tradition of normative epistemology, but as a discipline that has arisen in order to respond to a profound intellectual crisis within that tradition. In section 2.1.1 I provide a brief description of the history of epistemology, before going on to describe the seminal work in analytic philosophy of Gettier (1963) (2.1.2) which has been the catalyst for a range of attempts to 'save' traditional normative epistemology from its critics. In section 2.2 I review a radical response to this crisis in the traditional analysis of knowledge; the academic sub-discipline of social epistemology. I will also discuss some of the ways in which this discipline is a more suited companion for empirical linguistics than traditional normative epistemology, in particular by attempting to draw out some of the empirical consequences of such as view. Finally in section 2.3 some of the key works in the philosophy of science are discussed, and in particular those that have lead to the study of science as a social practice rather than an abstract set of intellectual behaviours.

2.1 Traditional Epistemology

2.1.1 Epistemology in analytic philosophy

The normative analysis of knowledge that has become ubiquitously known as the traditional or tri-partite theory of knowledge can be traced at least as far back as Plato; in his work *Theaetetus* the eponymous figure suggest a third and final solution to Socrates' questioning about the nature of knowledge; that it is true belief accompanied by an account or logos (Plato and Waterfield 1987). This 'account' or 'logos' is understood to be necessary (as a result of Socrates' questioning) because mere true belief is itself shown not to be sufficient for a proposition to count as knowledge

2.1.2 A crisis in the traditional view of knowledge: the Gettier case

The philosophical counter-example provided by Gettier (1963) produced what can be seen as a crisis in the traditional view of knowledge, with two counter-examples to the notion that a true justified belief is sufficient for knowledge. To take just one; Gettier asks us to imagine that a man (Smith) is at a job interview with another man (Jones). Smith can see that Jones is the better candidate, and develops the justified belief that Jones will get the job. He combines this belief with another justified belief (Jones has ten coins in his pocket) that leads him to a final, crucial justified belief- that a man with ten

coins in his pocket will get the job. In fact Smith gets the job- and it transpires that, unbeknownst to him, he in fact also had ten coins in his pocket. As a result his final justified belief- that a man with ten coins in his pocket will get the job- transpires to be a justified true belief. Yet we would clearly not want to equate Smith's justified true belief that the man with ten coins in his pocket would get the job with knowledge.

This original counter-example by Gettier has led to a great deal of discussion of this original example (cf. Shope 1983; Pollock 1986; Moser 1986 and Dancy 1987), with suggestions that there are flaws in Gettier type examples (eg. Kirkham 1984) and reformulations of even more convoluted versions of the Gettier type case that avoids its supposed faults (Feldman 1974, for instance). The volume of such literature expresses how serious a problem this is felt to be and the attempt to rescue the traditional view of knowledge from Gettier type examples by trying to reject them is one type of philosophical response. What the Gettier examples show is that the support supplied by 'justification' for true belief is not straightforward. Sturgeon (1998) provides a clear schema of the necessary and sufficient conditions now required after the acceptance of Gettier type examples:

S knows P iff (if and only if)	(a) S believes P,
	(b) S's belief in P is fallibly justified
	(c) P is true,
	(d) (b) ensures that (a) and (c) are not jointly

an accident

(Sturgeon, 1998: 17)

It is in condition (d) that the greatest problem lies for the traditional view of knowledge since, as Gettier has identified, this connection between (a) and (c) that can ensure that the two are not jointly accidental is problematic indeed. In our day to day existence there is a great deal that we want to retain as being 'knowledge', have a set of justifications for believing and yet cannot be certain that our justification has no accidental role. Yet the 'ensures' of condition (d) would appear to be a subtle reintroduction of certainty into the conditions of knowledge. The next section will outline attempts to move away from the traditional model of epistemology by rejecting the normative tradition and suggesting instead a positive alternative: social epistemology.

2.2 The move towards Social epistemology

The role of others in the development of our knowledge has long been an ignored area within the western philosophical tradition, though it has been revived somewhat under the heading of 'testimony'; the name given within that tradition for the role that others play in our acquiring knowledge. In work which in the context of a two-thousand year scholarly conversation is relatively recent Coady (1973; 1975; 1981; 1992) has drawn attention to this neglected area of study, identifying elements of the social transmission of knowledge throughout the history of western philosophy. The work of Kuhn (cf.1962) which will be discussed in more detail below (2.3) also contains elements of the social in its explanations of the development of knowledge and although Kuhn distanced himself

from some of the more radical conclusions drawn from his work (cf. Kuhn 1977; 1983; 1996) others have developed radical theories of epistemology based on an understanding gained from Kuhn (1962) of the social nature of the scientific process (eg: Longino 1990a).

Other writers from outside of the western tradition of analytic philosophy have also contributed to this burgeoning field and the work of Foucault (1980; 1989) is perhaps best known in this sense. What connects both traditions is an understanding of the process of knowledge development that (in often explicit contradistinction to the work of Descartes (cf. Descartes and Cress 1993; Descartes and Clarke 1999). From these disparate intellectual traditions the discipline of social epistemology has emerged in its own right, viewing the social aspect of knowledge encoding as an indispensible and central aspect rather than as a 'problem' created through counter-examples and thought-experiments such as the Gettier case to be 'resolved' in order to rescue either the traditional analysis of knowledge or something very similar. Central to this new discipline has been the work of Goldman (1986; 1987; 1999; 2001; 2004) who has attempted to generate new theories of knowledge based on this social aspect, drawing distinctions between group knowledge and group rationality and tackling the pressing question of the role of the expert in the knowledge community, one that is directly relevant to issues such as the popularisation of scientific texts. The wide range of approaches to knowledge that is developing in this field ranges from analysis of our belief forming practices and social interaction (Alston 1994) to work discussing the role of the ego in epistemology (Foley 1994). Most relevant to the present thesis however is

the seminal study of Latour and Woolgar (1979) that predates much of this increase in interest and plethora of publication in social epistemology but remains an influential text in the field and is the most relevant to the present study in applying the theoretical background of social epistemology to the sociological study of a scientific laboratory. It also provides many possibilities for complementary empirical linguistic research in the claims made about the social processes that impinge upon researchers and in the more achievable sense of the variation in epistemic signalling that can be found throughout what can be seen as the 'life cycle' of a scientific fact. The work is based at least partly upon the first hand observations of Bruno Latour and in fact Latour and Woolgar describe this process under the heading of 'an anthropologist visits the laboratory' (1979: 43-90). Whilst the almost comical anthropological approach is of course labour intensive in the extreme, the observations collected by Latour form a convincing picture of the social practices that influence the formation of texts. Whilst I shall return to the academic sub-discipline of discourse studies later (3.1) it is worth a brief look at a basic schema of such contexts to be clear on the supposed interrelationship of text and social practice- the following example from Fairclough (2003) is sufficient:

Social structure: languages

Social practice: orders of discourse

Social events: texts

(Fairclough 2003:24)

This simple structure provides a straightforward schema into which Latour and Woolgar's study can be placed; the actor in social events takes part in texts (both written and spoken), but is constrained by the social practices (what Fairclough here calls the

'orders of discourse' that he/she is submersed in. Whilst it is difficult indeed to replicate the study of Latour and Woolgar in gaining access to the day-to-day conversations of scientists, to be accepted by such an institution, be able to successfully negotiate various ethical difficulties and be able to cope with the extremely time consuming nature of such an approach (it is no coincidence that Latour and Woolgar produced a collaboration), the written texts that are according to them the product of scientific work can be examined in a much more straightforward way. In their own seminal work Latour and Woolgar provide a case-study of Thyrotropin Releasing Factor (hormone) (herein TRF) and attempt to trace the development of claims around this purported entity over time. In what is virtually a diachronic linguistic study, they are able to show that prior to 1962 there is doubt about the existence of TRF; that this is agreed upon within the scientific community in 1962, after which the question becomes (in their words) 'there is a TRF-what is it?' (147) That it is later agreed upon that it is a peptide (but not until 1966) that there is then doubt raised within the community with the possibility that it might not be a peptide being suggested; finally in January 1969 it is 'established' once again that it is a peptide and that it contains various agreed elements, with the final constituent parts being agreed upon in late 1969. This type of diachronic process is typical of that in social epistemology and the philosophy of science in focussing on a case-study of one given entity (in this case TRF) and tracing the changing epistemic value given to this entity over time (eg: it possibly exists, it does exist, it might not exist etc.). What is fascinating for the present study is the possibility of harnessing the tools of corpus linguistics to explore the linguistic nature of such changes in epistemic status over time, particularly as such a study would allow for a corpus-driven approach (about which more in section 3.6

below) to discovering the actual linguistic structures that encode this changing epistemic status.

Latour and Woolgar attempt to transform the linguistic structures they encounter into a classification of 'statement types' (1979:75-80) which are broadly as follows:

Type 1: statements comprise conjectures or speculations

Type 2: statements contain modalities which draw attention to the generality of available evidence (or the lack of it) sometimes taking the form of tentative suggestions

Type 3: a statement where the modality is constituted by the included reference

Type 4: deletion of modalities leaves a type four statement of fact

Type 5: statements corresponding to a taken-for-granted fact (1979:76-81)

Whilst Latour and Woolgar found in practice that there 'seems to be no simple relationship between the form of a statement and the level of facticity it expresses' this 5 point schema provides a useful abstract structure indicating how scientific facts are understood to become established. As has already been noted in relation to TRF, the epistemic status of facts can however go down as well as up and presumably diachronic study of a large number of entities in just the manner that TRF is analysed by Latour and Woolgar would also reveal some 'facts' surrounding entities that move both up and down this abstract scale; such as when a fact becomes accepted, then doubted again, then accepted again. The attempt to harness the tools of corpus linguistics to establish genuine insights into this process is perhaps the key task of this thesis.

2.3 The analytic tradition of the philosophy of science

The philosophy of science has a long history as a sub-discipline of the western 'analytic' tradition, and the focus upon, for example, metaphysics, the existence and nature of substance and the study of causation can be identified in the work of Aristotle (Aristotle and Acrill 1974; Aristotle and McMahon 2008) Hume (1975) and Hume and Buckle (2007) and Locke (Locke and Pringle-Sattinson 1978 and cf. Tipton 1977) amongst many others. The development of the philosophy of science as a particular area of special engagement over such issues of epistemology and metaphysics is most associated with the work of Kuhn (1959; 1962; 1970; 1977), Lakatos (1968; 1970; 1971; 1976) and Popper (1969; 1972; 1974; 1979; 1983), and these works particularly focused on the supposed 'special' nature of scientific knowledge; Popper in particular (cf. 1972; 1979) seeks to identify falsification as the key property of scientific discovery that separates a scientific discipline from one that is non-scientific and seeks to differentiate between statements that are falsifiable and those which are not. Lakatos (1977) also seeks to identify the difference between what he calls 'science' and 'pseudoscience' but differs from Popper in this, crucially, in arguing that the process of falsification is not in fact a key part of the scientific process, in the following way:

> **Scientists have thick skins. They do not abandon a theory merely because facts contradict it. They normally either invent some rescue hypothesis to explain what they call a mere anomaly or, if they cannot explain the anomaly, they ignore it, and direct their attention to other problems (1977: 4)**

What is particular striking about this argument from Lakatos is the appeal to actual scientific practice in structuring the theoretical distinction between science and pseudoscience. Lakatos rejects the notion that unlike, for instance, the astrologer, the scientist 'proper' is always keenly alive to the possibility of falsifying his ideas and assumptions. Indeed, he suggests quite the opposite: that scientists will endure considerable convolutions and labelling of 'anomalies', finessing hypotheses so that they fit the 'facts' as they have established them. However, Lakatos identifies the activity of prediction as the defining feature of science proper: for him it is the ability to consistently predict what will be discovered and for those predictions to be correct that makes a theory or set of practices scientific; and conversely, it is the failure to make accurate and verifiable predictions that marks other disciplines as pseudoscience.

This appeal to the actual social practice of scientists made by Lakatos is mirrored in, and indeed inspired by, the work of Kuhn and in particular his Structure of Scientific Revolutions (1962). In this seminal work Kuhn focuses not on abstract reasoning for an understanding of the nature of scientific knowledge but rather on the psychological and social processes that underpin and constitute scientific practice. Kuhn's influential theory marshalled the concepts of normal science, revolutionary science and paradigm shifts to explain scientific process , accurately describes actual scientific practice. According to Kuhn normal science occurs within a particular paradigm, practising methodology that is agreed on and established within the scientific community and works as an agglomeration of knowledge within that particular paradigm until such time as that process is deemed unsatisfactory, either in and of itself or more likely because a paradigm shift has

occurred; a piece of scientific work that cannot be assimilated into current practices but instead demands a new and different way of theorising and of practising. Kuhn (1962) gives the example of the discovery of the X-ray as such a paradigm shift; within the field of radiation theory 'the emergence of X-rays necessarily violated one paradigm as it created another' (1962: 93).

For the applied linguist, the theory has profound implications in terms of what might typically be found in any study of scientific writing. Unless a piece of 'revolutionary science' is deliberately chosen for empirical study it would seem unlikely that empirical study of scientific texts will discover examples of such paradigm shifts, particularly if such work takes a corpus approach and works with typical or common patterns; rather, what is likely to be unearthed is the process of normal science; the process of accumulating knowledge within a particular theoretical paradigm. The value of such study is in illuminating the epistemic process through which scientists typically establish and develop new knowledge; the processes that will presumably form the training of scientists and enable them to publish their first papers and establish themselves within a pre-existing field of study. However it must be noted that the seemingly straightforward process of knowledge accumulation in normal science has been argued by many to mask actual scientific practices that are not accurately described in the publications that announce new findings: theorists such as Feyerabend (1975; 1976;1978;1981a;1981b) have argued that the actual scientific process is often based on chance findings rather than the consistent application of a method and indeed Kuhn (1977) describes how the paradigm shift constituted by the discovery of X-rays was based on an accidental

laboratory observation. The work of Kuhn has led to a plethora of such radical theories about the nature of objectivity and methodology of science such as that of Longino (1983;1987;1990b;1994) and Okruhlik (1994) both of whom seek to challenge the established view of objectivity at least partly from a feminist perspective. The work of Latour and Woolgar discussed above (2.2) is also explicitly inspired by that of Kuhn and the attempt to represent faithfully the social nature of scientific discovery that these researchers make is a key inspiration to this study. However it must be clear at this point that there is no ethnographical or sociological aspect to the present work. Rather, what is attempted here is to contribute to the linguistic investigation of scientific writing, and this tradition will be discussed in the next chapter.

2.4 Conclusion

This chapter has provided the theoretical background for the present study, discussing the relevant literature from the related disciplines of epistemology (2.1) and social epistemology (2.2). It has sought to place social epistemology firmly within the discipline of traditional epistemology, rather than being viewed as a slightly alien or outlandish response to a philosophical problem. I have argued that the Gettier type examples cannot be responded to adequately by normative epistemology, with those who seek to defend the traditional view being caught in an unfortunate dilemma between setting an impossibly high standard of what constitutes knowledge that borders on the practically impossible requirement of certainty, or accepting that we cannot in fact establish necessary and sufficient conditions of what ought to constitute knowledge. The answer to this dilemma in my view is to examine the actual practices of knowledge making

positively by asking what it is exactly that (in this case) geneticists actually do in order to signal new knowledge to each other unencumbered by impossibly high standards of what it is for something to be a fact; put simply, to do social epistemology. Finally I have reviewed some of the insights from the philosophy of science (2.3) that can provide further theoretical context for the present study, and outlined the way in which social epistemology can also be seen as a part of the tradition of the philosophy of science when it aims to describe the knowledge building processes of scientific disciplines. Having done this I must now turn to the more empirical part of this literature review and assess the ways in which the discipline of linguistics can contribute to social epistemology.

Chapter 3: The contribution of linguistics

In this chapter I will attempt to place the current study within a range of related works all of which have influenced the methodological or theoretical approach that was finally taken. Whilst there have been few studies indeed that have attempted to take a radically corpus-driven approach to the study of knowledge construction in scientific texts, comparisons can be made to elements of all of the studies I discuss here, and the present work can be seen as forming a contribution to the same academic conversation that is spoken to by these studies. In order to impose some sort of order to what follows these are divided along what are broadly speaking methodological lines. Studies of scientific texts that do not proceed by using the methods of corpus linguistics will be discussed first (3.1) before the overall paradigm shift offered by corpus methods is introduced (3.2). This will also allow for a brief adumbration of the concepts and methods of corpus linguistic (3.3), and particular emphasis will of course be placed on those that will be employed in this thesis. I will then provide a discussion of a number of studies that employ corpus methods in ways that are relevant to the present work, particularly previous corpus approaches to scientific language and academic writing more generally. I will then discuss the potential of corpus-based and corpus-driven approaches (3.5) explaining what is meant by the corpus-based/corpus-driven distinction before discussing the ways in which this is relevant in methodological terms for the present study, arguing that what is required is a corpus-driven approach. I will then provide a summary of the main points raised in this chapter (3.6) before moving on to the methodology section itself.

3.1 The study of discourse

A considerable body of work within applied linguistics attempts to build upon a framework provided by theorists such as Foucault (1972; 1984) and Habermas (1972:1984;1987a;1987b) by identifying the discursive practices which function in particular 'discourse societies' or 'discourse communities'. Such work places language study firmly in the social context and there are multitudinous works in this field (e.g. Widdowson (1979:2004); Sinclair and Coulthard (1975); Hoey (1983;1991;2001); Coulthard (1985;1992;1994a); Cook (1989;1992;1994); Fairclough (1993;1995;2000;2003;2006) McCarthy and Carter (1994; Carter 1997;2004).

Various scholars have pointed out the potential confusions that surround the various uses of the term *discourse* in this field and Groom (2007:23) follows a number of recent works (he cites Pennycook (1994), Cameron (2001) and Baker (2006)) in identifying the distinction between *discourse* as a countable noun and *discourse* as an uncountable noun as the key move needed to bring clarity to the use of this term: as Groom (2007) points out:

> **discourse as an uncountable noun tends to be used to refer to "any naturally occurring stretch of language, spoken or written" (Carter 1995:39), thereby effectively incorporating all and any linguistics phenomena not covered by mainstream Chomskyan linguistics (2007: 24).**

Needless to say it is not this very broad definition of discourse that is most relevant to the present study but the far narrower one represented by the countable discourses, which identifies discourse as:

> **Particular set[s] of beliefs, values and attitudes, which are embedded in social and cultural practices and which shape the identity of those associated with [them]. Discourse in this meaning is manifested in language, *most saliently in the way conversations, arguments, written reports, narratives, etc. are conducted* (Carter 1995:42, cited in Groom 2007: 24, my italics).**

It is in this sense that a considerable body of work in applied linguistics has set out to identify linguistic features that are relevant to the present study. The texts types found in the two corpora used here (*genepilot* and *genecorp*) are inevitably shaped by and therefore reveal the discourse practices of genetics as represented by the journal *Nature Genetics*. Whilst not all of the vast literature of discourse studies is closely relevant to the present work the study of scientific texts has formed a sizable subsection of this field and some of the key works in the study of scientific discourse are discussed below in section 3.3. It is also worth noting that one result of the development of discourse studies into a thriving area of research has been that some researchers have inevitably sought to create explicit sub-disciplines within the field, most notably in the Critical Discourse Analysis of Fairclough (1989;1992;1995;2004;2006), van Dijk (1991;1992;1993a;1993b;2009), van Leeuwen (2004;2005), Wodak (1989;1997;2001) and others. These researchers have argued for the need to extend the focus of discourse analysis in such a way as to create an

approach that actively seeks to engage with and disrupt such discourse where it is in some sense harmful, particularly to minority interests or groups. Bloor and Bloor have provided a helpfully succinct definition of the key aspects of discourse studies and Critical Discourse Analysis that are helpful in elucidating the key differences between the two. They identify they key aims of discourse studies as follows:

> **It has three main purposes: (1) to identify and describe how people use language to communicate; (2) to develop methods of analysis that help to reveal the categories (or varieties) of discourse and the essential features of each; and (3) to build theories about how communication takes place (2007: 12)**

Whereas in contrast to these predominantly descriptive goals the aims of Critical Discourse Analysis are identified as follows:

- To analyse discourse practices that reflect or construct social problems
- To investigate how ideologies can become frozen in language and find ways to break the ice
- To increase awareness of how to apply these objectives to specific cases of injustice, prejudice, and misuse of power (2007:12)

Whilst the present study does not have the goal of 'breaking the ice' of ideological language or raising awareness of the misuse of power it should be noted that such work has been undertaken with explicit attention being paid to scientific texts. Greenhalgh (1998), for example, examines the 'political, ideological and cultural context' which

compromises the ability of researchers to write in an impartial way; whilst the present study will focus only on the textual output of geneticists rather than the political or ideological notions that might underpin their work the complementary value of such studies must be acknowledged.

3.2 Non-corpus study of scientific texts

A wide range of linguistic studies of scientific texts pre-dates the use of Corpus Linguistics methods for applied linguistics and have already contributed a great deal to the empirical understanding of knowledge signalling in scientific texts. Perhaps most obvious amongst these is the impressive body of work of Myers (1989;1990;1991;1992;1994) which in particular stands out as a contribution to the understanding of scientific writing. Myers speaks directly to the social construction of knowledge in several of these studies, though his focus is less on the expression of knowledge between scientists working in a given field and more on the popularisation and wider dissemination of scientific findings. Through his work both in assessing the language of textbooks in the process of socialising scientific knowledge (1992) and the comparison of texts from specialised scientific writing and popular versions of these texts (1991;1994) he is able to illuminate the linguistic processes involved in popularisation, arguing persuasively that the different sets of texts (scientific versus popular) encode different narratives of the scientific process, instantiating thoroughly different views of the nature of scientific practice; with scientists viewing their work 'as much more tentative and mediated than does the public' (1994:189). Though popularisation is not directly relevant to the present study this view of scientists' own perception of their work

accords with the placement of hedging at the very centre of the linguistic study of scientific writing argued by Hyland (1998).

Myers' study of the writing process of a scientific research article (1990) also provides a compelling if very localised example of the threats to the integrity of the writing process that can be wrought by anticipation of the peer-review process. His description of the scientists materially altering their article and in particularly reformulating claims so that they are more cautious, until they are less representative of what they actually want to say but more reflective of what they believe will be accepted in the wider discourse community has profound implications for any approach which seeks to treat the final published article as the object of study. The drawback of a such a study of course is that the necessarily small scale and labour intensive process needed to monitor the drafting process makes such work very difficult to achieve in practice; and that is without even considering the difficulty of needing to get a scientist, or, in practice, a collaborating team of scientists, to agree to such work taking place. However there are concerns with such an ethnographic practice that can be addressed by an approach that uses a corpus of final texts, and there may even be advantages to such a method. Whilst Myers undoubtedly gains great insights into the scientists' thought processes during writing and to their perceptions of the reception by the wider discourse community, what he is partially studying in such work is just that: perceptions. Scientists' intuitions about what will be accepted by the discourse community may not be infallible and there may be great gains to be made from studying what actually and typically takes place in texts that are published in genetics rather than the drafting process. The gains in scope and breadth of

study are more obvious, and I note that both the arguments and conclusions of Myers and Hyland are, however, based on examples chosen from a relatively small number of texts; this acts as both an inspiration and challenge to attempt to find a way to put such findings to the test in a more empirically rigorous way.

A potential complexity that any attempt to develop a corpus-driven approach to scientific writing must face is that introduced by the considerable literature on the differing rhetorical roles of the separate sections of scientific research articles. Hyland (1998) neatly summarises this research as having established, amongst other things, that what he calls greater 'writer intrusion' (25) occurs in Introduction and Discussion sections, and he asserts that the previous research in this area has established some precise and substantive linguistic claims about the grammatical and lexical forms found in these parts of research articles:

> **'pronominals, verbs of reasoning, "that- nominals", and adverbs, adjectives and modals qualifying assertions tend to cluster here (West 1980; Bernhardt 1985; Butler 1985)' (1998:25)**

Whilst it is not the intention of this work to divide texts into their constituent parts and compare the frequency of lexical or grammatical items therein, this is clearly an empirical claim that can be tested and one that has consequences for a corpus-driven approach; if items are to be identified in the corpus as a whole and then investigated it will be important to be mindful of the (arguably) differing linguistic nature of the different

sections of the text in drawing any conclusions from these findings. The work of Swales (1981;1990) has also been highly influential in this area and has led to a number of studies by others in other disciplines (e.g. Brett 1994) which follows Swales by reporting on the communicative 'moves' in results sections of sociology articles). If it is accepted as a result of such studies that each section within a research article has its own goals, this then will presumably also be reflected in the epistemic processes that are found in each section. As an exploratory study the present work will proceed by making no assumptions about where precisely within the texts in the corpus the most valid objects of study are to be found, or about whether dividing a corpus of articles into sub-corpora based on the different sections (ie. Introduction, Methods, Results and Discussion) will reveal these differences, but will instead proceed from a corpus-driven methodology that takes high frequency lexical words as a starting point; the reasons for this will be discussed below (4.4.4) but at this stage it is worth noting that the differing goals in the different sub-sections of scientific texts may strongly influence the linguistic features present.

The work of Hunston (1989;1993;1994) on evaluation in scientific writing has also contributed to our current linguistic understanding of scientific writing practices within a framework of social epistemology. Drawing upon, for example, the insights of Latour and Woolgar (1979) that the aim of a piece of scientific research is to persuade the academic community to accept the new claims (1994: 192) Hunston identifies three kinds of evaluation, performing three distinct functions in the text; status, value and relevance. Whilst the approach of Hunston is again carried out on a relatively small group of texts she is able to forge a clear link between the evaluative forms found, their organisational

patterns in the text and the role that such forms and patterns play in making the text 'acceptable' to the academic community at large.

Finally it is worth mentioning a number of pieces of work on the popularisation of genetics (cf. Condit 2001) that have a marked tendency to decry the linguistic form of reporting of genetics with no actual reference to or comparison with original texts in the way that Myers (1991;1994) approaches this topic. Typical of such work is the *Frame That Gene* project (Carver et al. 2008) which takes 300 texts from the popular reporting of genetics and attempts to identify the semantic frames present in media reporting. Whilst the project identifies and decries popular forms such as 'classical determinism' in genetics reporting which is identified with 'key words such as 'cause', 'control', 'blame' and 'disease'' (Carver et al. 2008: 946) there is never any comparison with the original scientific texts or attempt to explain why such terms as 'cause' and 'disease' have come to form what they claim is a prominent popular misconception of the nature of genetics. Part of the motivation of this thesis is to improve upon this position by arriving at a more detailed picture of the way in which geneticists typically encode knowledge in their discipline, and the attempt to do this will employ the methods of Corpus Linguistics.

3.3 Corpus Linguistics: concordances, collocation and keywords

It is clear that the study of the social construction of knowledge is by its very nature empirical. Whilst the normative approach to epistemology discussed above (2.3) sought to find out what the necessary and sufficient conditions are for a proposition to constitute knowledge or be given the status of a fact, the positive approach seeks instead to view what is actually done: for the purpose of this thesis this means how scientists working in genetics actually signal the epistemic status of a claim to each other. This process requires an empirical method for investigating the writing of geneticists, and the methodological framework that will be employed in this thesis is broadly that of Corpus Linguistics.

Whilst the approaches outlined in section 3.2 are clearly all empirical in nature and can all be seen to have contributed to the understanding of scientific language, they all suffer from the limitations inherent in being focussed on one or just a few texts. The process of selecting items for study falls entirely to the intuitions of the analyst and this is a particularly worrying aspect when one considers that those working in this field are generally linguists and not scientists and therefore have a very limited knowledge of the discipline that they are studying. There is little way of knowing, for instance, how typical the wordforms they discuss are, how often the phenomena they identify actually attach to those particular wordforms or what other significant linguistic forms might have been missed in their analysis of just a handful of texts. It is to address just these types of

weaknesses that the discipline of Corpus Linguistics is often applied to an area of linguistic study.

Perhaps most often associated with the work of Sinclair (1987a;1987b;1991;1992), Sinclair and Carter (2004) and Biber and his co-workers (1988; 1992; 1993;1995;1996a; 1996b;Biber et al. 2007;Biber 2009), Corpus Linguistics has come to encompass a range of approaches so diverse in their main point of focus and methodological practices that researchers working in the field may have almost nothing in common with each other apart from the fact that they call the collection of texts that they use a *corpus*: an electronic collection of texts or parts of text that are authentic examples of language use. Corpora rather than analyst intuition are regarded as being the font of linguistic knowledge. Whilst there is not space to provide a history of Corpus Linguistics in the present thesis (which can in any case be found elsewhere eg: Leech (1991)) it is perhaps worth dwelling momentarily on the following exchange involving Noam Chomsky that is cited in McEnery and Wilson (2001):

Chomsky: The verb perform cannot be used with mass word objects: one can perform a task but one cannot perform labour

Hatcher: How do you know if you don't use a corpus and haven't studied the verb to perform?

Chomsky: How do I know? Because I am a native speaker of the English Language.

(Hill (1962), cited in McEnery and Wilson (2001))

As McEnery and Wilson point out, Chomsky is simply wrong: 'One can perform magic, for example, as a check of a corpus such as the BNC reveals' (2001: 11). It would be hard to imagine a more devastating blow to a linguistic practice than its most famous and lauded proponent blithely insisting on the primacy of his 'native speaker intuition' over actual empirical evidence whilst simultaneously making such a demonstrably false claim about the English language; as McEnery and Wilson observe, 'Native speaker intuition merely allowed Chomsky to be wrong with an air of absolute certainty' (2001:11).

In place of native speaker intuition, the empirical approach employed by Corpus Linguistics allows the computer to select objects for study, reducing researcher bias. Whilst in practice the ways in which these possibilities are harnessed vary greatly, the potential for a corpus to show us new information about language use is impressive. In the most radical form this can lead to new claims about the nature of the language itself, such as that of Hunston (2002) that meaning ultimately belongs 'to the whole phrase rather than to individual words in it'.

In this thesis the Corpus Linguistics concepts of concordance, span and collocation will be used to explore the linguistic nature of knowledge encoding in scientific writing. These concepts are interrelated and can be most easily defined together: the *span* (Sinclair 1991:175) is a window of a number of stipulated places either side of a search

word or phrase. Words found to be 'occurring within five words either way of the headword with a greater frequency than the law of averages would lead you to suspect' (Krishnamurthy 1987) are labelled *collocates*. Whilst the identification of collocates within the span can be carried out automatically by corpus investigation software the concordancer is a tool that enables more manual analysis by organising examples of a keyword in context (also commonly known as KWIC). This alignment of many examples vertically allows the analyst to identify patterns of both form and meaning surrounding a node word (cf. Sinclair 1991). A further term can also be introduced at this point: that of colligation. This according to Sinclair and Carter (2004) is the extension of the phenomenon of collocation to look also at the grammatical patterning around a node phrase. Sinclair and Carter (2004) show how the systematic investigation of a node phrase can proceed on the basic of a process of identifying the patterns around such a node phrase, and this approach will be taken in this study in order to identify the linguistic nature of epistemic encoding around terms in scientific texts.

3.4 Applications of Corpus Linguistics

The methodological approaches that constitute Corpus Linguistics have already been applied to an impressively wide range of sub-disciplines of applied linguistics including lexicography (Sinclair 1987a;1987b;1990;1991; Biber 1993;1996a; Moon 1998) , the study of grammar (Halliday 1993; Francis and Hunston 1996; Francis, Hunston and Manning 1998; Biber et al. 1999; Hunston and Francis 2000; Mindt 2000; Francis 2003; Mahlberg 2003; Hoey 2005), translation studies (Baker 1993; Teubert 1996; Baker 1999),

language pedagogy (Johns 1991; Bernadini 2004; Sinclair 2004; Lew 2009), literary studies (Louw 1993; O'Halloran 2007), forensic linguistics (Coulthard 1993; 1994b), sociolinguistics (Lee and Ziegeler 2006; Beeching 2006; Mautner 2007; Culpeper 2009; Millar 2009;) and CDA (Teubert 2000; Mauntner 2005; 2009; Orpin 2005; Baker 2006; Baker et al. 2008). It is difficult to provide an overarching generalisation of what it is exactly that identifies this diverse body of work as constituting Corpus Linguistics but Hunston (2002) provides an illuminating summary of what she calls the 'emphases' of approaches that seek to apply Corpus Linguistics thus:

- **An emphasis on frequency**
- **An emphasis on collocation and phraseology**
- **An emphasis on variation**
- **An emphasis on lexis and grammar**
- **An emphasis on authenticity (2002:96)**

This list applies to most if not all of the studies named above in that they approach linguistic investigation with one or more of these emphases strongly influencing methodological considerations. The relevance of these to the present study will be discussed throughout chapter 4, but it is worth providing a brief adumbration at this point of the ways in which the present study is to be understood as an application of corpus linguistics with reference to these emphases. Thus the corpora generated (*genepilot* and *genecorp*) comprise authentic texts reporting findings in genetics in the journals *Nature* and *Nature Genetics* for the pilot corpus and *Nature Genetics* for the final corpus. In

order to represent these publications as authentically as possible every text type found therein has been retained for further study, meaning that the final corpus has the same diverse range of text types found in the actual journal. The keywords that are studied are also the ten most frequent lexical words in *genecorp* and once the study had focused upon clusters with at least three lexical elements the most frequent of these were then used for further investigation. The approach to investigating these then uses collocation and phraseology as the principal objects of study, identifying common collocates and phraseological patterns using *WordSmith Tools* (Scott, (2004)). In doing so I hope to add contribute to those studies that have already taken a Corpus Linguistics approach to scientific texts, and some of the key publications in this area will now be discussed in some detail.

3.5 Corpus linguistics, scientific texts and academic writing

Perhaps most relevant to the present study are other works that have employed the methods of corpus linguistics in studying scientific language. A key text amongst these is Hyland (1998) which reports on an investigation of a small corpus of 26 research articles 'in the field of cell and molecular biology' (p.96). Whilst corpus size and scope has increased exponentially since this work was carried out it remains an influential text in describing in detail what he calls 'hedging' in scientific research articles. Following Hyland, hedging is to be understood in this thesis as 'one part of epistemic modality; it indicates an unwillingness to make an explicit and complete commitment to the truth of propositions' (1998:3) and Hyland names 'compromisers […] downtoners […] softeners […] back-grounding terms […] and pragmatic devices' as various terms that have

appeared as labels for roughly this phenomenon. Hyland's approach is of course only focussed on a relatively small number of texts but results in an impressive if somewhat predictable list of devices that can be used to hedge a scientific claim, including modal auxiliaries, what he calls 'epistemic lexical verbs' and 'epistemic adjectives, adverbs and nouns' with wordforms such as *essentially*, *relatively*, *generally*, *most*, *slightly* and *presumably* and their various frequencies in his research articles being compared to their relative frequency in the JDEST corpus (a corpus of 2,000 texts of approximately 500 words each totalling around 1 million words and comprising English texts from ten scientific disciplines) and the Brown/LOB corpus once the frequencies had been adjusted to 75,000 words. Hyland also provides a discussion of the hedging of numerical data and what he terms 'non-lexical hedges'. This latter category is of particular interest in that whilst Hyland presents these as fairly abstract 'strategies' (the frequency of which he also attempts to judge), the actual linguistic details of these are far from obvious or predictable and include phrases such as 'one cannot exclude a possibility that', 'cannot presently be ruled out' and the perhaps more predictable 'it is not known whether'. Hyland sub-categorises these strategies as 'reference to limiting experimental conditions', 'reference to a model, theory or methodology' and 'admission to a lack of knowledge' and provides plentiful examples of these from a corpus of just 26 research articles. Whilst Hyland attempts to judge the frequency of these by extrapolating a figure from his own small corpus it might be expected that an approach based on a larger corpus could reveal similar epistemic devices in patterns identifiable through corpus methods rather than Hyland's more manual technique which was of course entirely necessary at that stage of corpus methodology. In his more recent work Hyland has used corpus methods to

contribute to work in disciplinary discourse (eg: Hyland 2004; 2008) but it is perhaps this earlier work on hedging in scientific articles that is most relevant to the present study.

An approach that applied a slightly more corpus-oriented methodology to a corpus of scientific articles of a similar size is the work on collocations in scientific writing done by Gledhill (2000). Taking a very different methodological approach to that of Hyland, Gledhill (2000a) investigates the collocates around grammatical words in a corpus of approximately 500,000 words made up of cancer research articles. This innovative method is undertaken in an attempt to uncover what Gledhill calls the 'most typical expressions' in the corpus" (Gledhill 2000b: 117), and this unusual technique takes the most frequent of grammatical words (such as *of*, *the* etc) and attempts to draw conclusions about the nature of scientific writing on the basis of these, arguing for example that what he calls 'extraposed *that*- clauses' have an important role in epistemic signalling in scientific writing. This approach has since been refined by Groom (2007) in order to attempt to construct a more far-reaching and systematic way of comparing disciplinary epistemology, and Groom's corpus-driven application of this to a set of research articles and review articles in history and literary studies makes a persuasive case that a quantitative comparison of the use of such grammatical items across academic disciplines can uncover systematic differences in disciplinary epistemology.

A range of other studies have employed corpus methods to identify very specific pre-selected features in scientific writing, constituting what will be described below as 'corpus-based' studies. Thus Ferguson's (2001) study of *if*-conditionals compared their use across three genres (research articles, journal editorials and doctor-patient

consultations), with the attempt to compare their use across speech and writing apparently being a key aim of the study. Whilst this type of study can prove illuminating in the fine details of if-conditional use (and, of course, gains considerable ease in methodology by pre-selecting an easily identifiable object of study) its conclusions are severely limited by the scale of the corpus used, which totals merely 177 examples of *if*-conditionals from an approximately 100,000 word corpus.

Away from the focus on specifically scientific language there has been plentiful study of academic language using corpora and in ways intended to reveal important linguistic details that are not immediately available to an analyst through introspection or intuition. Indeed work in this area has become so popular that Groom (2007) was able to identify (somewhat despairingly) the 'usual suspects' of corpus study for this purpose. As he states:

> **'A glance at the recent literature identifies report clauses and other attributive forms […] modal verbs and other hedging devices […] and extraposed complement clauses and other kinds of that- clause […] as being amongst the usual suspects' (2007: 40)**

To this list we could probably add the study of various kinds of semi-fixed phrases known variously as lexical bundles (eg. Biber (2009), Cortes (2004)), fixed collocation patterns (Oakey 2008) and also, named differently according to the particular piece of software used, such as clusters in *WordSmith Tools* (Scott 2004) and c-grams in *W—*

Matrix (Rayson 2009). What these works have in common, and indeed what they have in common with Groom's 'usual suspects' of report clauses, modal verbs and extraposed *that*- clauses is that these linguistic items are again pre-selected as the linguistic device for study. Whilst there is most certainly a place for such studies they have the drawback of limiting the feature to be studied at the outset, limiting in turn what can be discovered about academic or scientific writing. Cortes' (2004) pedagogically motivated study into student writing is a typical example of such approaches. The overall aim is again disciplinary comparison, with the pedagogic motivation coming from an EFL perspective where students writing in a language other than their first language are being taught academic writing. By pre-selecting lexical bundles for study Cortes assumes that the construction of 'target bundles' (which are derived from professional writing in the fields as represented by published research articles) is what is needed for improved student writing and the study proceeds from the identification of these bundles using automatic corpus methods. A recent corpus-based study of the pronoun 'we' in scientific texts (Noguchi et al 2006) provides a further example of this type of study; in this case it is one particular pronoun form that is chosen for detailed study; whilst there can be little doubt that such a study represents a successful use of a corpus to identify the actual, authentic use of a linguistic phenomenon in scientific texts it could be argued that such a study again fails to add to the list of 'usual suspects'.

Corpus-based approaches such as those outlined above have the profound benefit of being able to identify the objects of study quickly and indeed automatically, and it must be said that the disciplinary comparison work that is often carried out on the basis of this

work can produce impressive findings regarding the relative use of such bundles in different disciplines. However, it is hard not to view such work as contributing little that is new or unexpected in its findings in terms of the actual linguistic items identified. Thus whilst the recent study by Biber (2009) of multi-word formulaic sequences is arguably exemplary in terms of the clarity and replicability of the methods used it is hard to see how the actual findings (that the strings used in conversation tend to be fixed sequences whilst the strings in academic writing tend to contain what Biber calls an 'intervening variable slot' which can contain differing content words) genuinely constitutes more than a very slender advance in our understanding of academic writing.

If we accept that the most important gain from a corpus approach is the enhanced ability to discover facts about language that are not immediately obvious or even available to or achievable through intuition it seems a shame to focus instead upon the capacity of corpus methods to measure language, using frequency and by comparing items across different (and large) bodies of data. Whilst these latter abilities are undoubtedly strengths of approaches using corpora it is argued here that corpus methods can at this point in time best be exploited to search for epistemically significant relationships in scientific language that do not at present occur within the list of 'usual suspects' identified by Groom; the type of corpus approach that might allow this to be done- the corpus-driven approach- will now be considered.

3.6 The potential of corpus-based and corpus driven approaches

3.6.1 The corpus-based/corpus-driven distinction

Whilst the studies discussed in the previous section all made use of corpora, they vary in both the theoretical and methodological role that they accord the use of the corpus within their study. The much discussed corpus-based/corpus-driven distinction (Tognini-Bonelli 2001) remains a useful way to categorise such studies, particularly with a view to understanding how the ultimate research goals of each study are to be met by their use of corpora. In the original formulation of this distinction Tognini-Bonelli defines a corpus-based approach as one which uses a corpus 'as a repository of examples to expound, test or exemplify given theoretical statements' (2001:10). In this sense each of the studies discussed above in section 3.4 takes a corpus-based approach (and sometimes explicitly so) in the sense that the actual linguistic item that is supposed to be of interest in the study of scientific language (such as personal pronouns in the case of Harwood (2005) or grammatical words in the case of Gledhill (2000b)) is pre-defined before the study begins. The corpus then functions as a database to be searched for examples of such items. This differs from the corpus-driven approach, where the linguistic items to be studied, and indeed ultimately any theoretical statements based on observation of these items, are determined as a result of corpus investigation. Thus this is a fully a posteriori approach, where 'a theoretical statement can only be formulated in the presence of corpus evidence and is fully accountable to it' (Tognini-Bonelli, 2001: 11). Classic examples of the corpus-driven approach would include Sinclair's (1991) oft quoted discussion of *eye* versus *eyes* where, contrary to what might have been assumed *a priori*, very significant differences in pattern and meaning surrounding the two wordforms emerge after

observing corpus data. Far from 'simply' constituting the plural of *eye* and therefore behaving in only a superficially different way in semantic terms, Sinclair famously demonstrated that *eyes* typically had a different set of collocates from *eye*. Other commonly cited examples of corpus-driven research include the pattern grammar work of Francis et al. (1996; 1998) which attempts to create a grammar that accounts for semantic constraints inherent in grammatical usage; the work of Mason and Hunston (2004) that attempts to show how the understanding facilitated by the work of Francis et al. could be operationalised into an automatic analysis of grammar patterns; the work of Groom 2007 (discussed in more detail in section 3.5.2 below) which attempts to find a corpus-driven methodology for the study of disciplinary epistemology looking in particular at research articles and review articles in history and literary studies; the work of Oakey (2008) in identifying the form and functional of what he calls 'fixed collocational patterns' in research articles in a range of different academic disciplines and perhaps most famously of all the work of Biber in register analysis and the study of formulaic phrases in spoken language and in academic English (1999; 2009). What these approaches all have in common according to Biber is what he calls 'the nature of their central research goal: to uncover new linguistic constructs through inductive analysis of corpora'. (2009:278). What precisely this type of approach means and how it might be useful in the present thesis will now be explored.

3.6.2 Epistemic signalling and the corpus-driven approach

In a thorough recent discussion of this issue Biber (2009) defines his own approach to the study of formulaic language as corpus-driven in careful contradistinction to what he formulates as being the corpus-based approach:

> **'Corpus-based' research assumes the validity of linguistic structures derived from linguistic theory; the primary research goal is to analyze the systematic patterns of use for those pre-defined linguistic features. Thus, in corpus-based studies of formulaic language, the researcher pre-selects formulaic expressions, and then analyzes the corpus to discover how those expressions are used' (Biber, 2009, p. 276)**

This formulation by Biber neatly expresses how this distinction is generally understood; a corpus-based approach is one that pre-selects the linguistic item of study, whilst a corpus-driven approach will allow the corpus to 'show' the analyst what to investigate further through some form of frequency measure. Biber goes on to list three desiderata for what he calls a 'radical corpus-driven approach to formulaic language' which he says is based on a synthesis of previous theoretical discussions, and his criteria (are as follows:

> **1. it would be based on analysis of the actual word forms that occur in the corpus (not lemmas)**

> **2. it would be based on analysis of sequences of word forms, with no consideration given to the grammatical/syntactic status of those words**

3. it would focus on frequent, recurrent combinations of word form (Biber, 2009: 281)

Thus through this approach Biber arrives at a set of recurrent combinations of words (such as *it is clear that*, *is likely to be* and *it is possible that*) which he has in no sense pre-selected himself. However, as the keen reader has no doubt already spotted, in studying such items we have returned once again to Groom's 'likely suspects' discussed in section 3.4 above. Whilst Biber's approach may well be a radically corpus-driven approach to formulaic language, it is not a radically corpus-driven approach to academic language, since it again pre-selects the linguistic item of study (in this case 'frequent, recurrent combinations of words' (Biber, 2009: 281)). For the present study this is a crucial distinction because a study such as that of Biber (2009) has already identified the crucial linguistic form: formulaic phrases. In the present study the argument is that study of scientific language is currently being limited by the pre-selection of such linguistic items and that if there is to be any possibility of moving beyond what Groom (2007) calls the 'usual suspects' a more radical approach than even that of Biber is required. Biber himself argues that in practice the corpus-based/corpus-driven distinction is not a simple one, and that the term 'corpus-driven' has encompassed 'a fairly wide range of methodologies' (2009:278). Indeed, he even argues that the pattern grammar studies of Hunston and Francis (eg. Hunston and Francis (2000)) constitute what he calls a hybrid approach to the methodology, combining both corpus-based and corpus-driven techniques:

> **The pattern grammar studies are instructive here because they are often cited as the best developed example of corpus-driven research, but in practice they employ both corpus-driven and corpus-based methodologies. The studies are corpus-driven because the lexical associations of each pattern are discovered through corpus analysis. However, the studies are corpus-based because the analyses are in part determined by pre-defined linguistic categories (including basic grammatical categories like 'noun' and 'verb', phrase types, and even syntactic structures). (2009:287)**

For Biber the pure corpus-driven approach differs from such a hybrid approach in that it assumes no categories at all prior to analysis, with the process of selecting linguistic objects for study having no criteria other than frequency. For the present study this is a profound issue because a corpus and indeed a computer is not capable of automatically identifying objects that extend our understanding of epistemological signalling in scientific texts. Part of the reason why the 'usual suspects' are commonly studied is that it is a fairly straightforward task to identify them in a corpus. One can identify all of the instances of closed class items such as those that make up grammatical modality with great ease. However, when charged with the task of identifying epistemically significant items whose linguistic nature is not yet known the task is much more difficult. Biber can proceed with a frequency based approach to selecting multi-word patterns because he has already decided that it is such patterns that he wishes to explore. However, whilst the study of such linguistic items undoubtedly improves our understanding of some of the formulaic aspects of scientific writing, and indeed constitutes well-defined items of study for the type of quantitative cross-disciplinary comparison that Biber is interested in, they are arguably not very useful in extending our understanding social epistemology or of the

linguistic forms that signal epistemic meaning in genetics. I would argue that whether corpus-based or corpus-driven study is appropriate should not be a lifelong commitment by an analyst but rather should be decided by what is required within a given area of study at a given time; where a corpus-driven approach should be used in my view is when linguistic findings in a given area have become somewhat stagnant or banal. In order to take a broader view that seeks to discover additional linguistic forms of knowledge construction this thesis will use the term *epistemic signalling* as a generic term for any variation that carries epistemic meaning. No assumption will be made from the outset as to what forms this epistemic signalling might take and therefore it is hoped that the list of 'usual suspects' will be added to and improved upon here. Whilst there is no doubt that structures such as *possible*, *probably*, *it is likely that* and *it is possible that* play a role in epistemic signalling in scientific writing what is needed is an approach which can produce findings that are more likely to be of interest to those working in social epistemology whilst simultaneously extending our understanding of the linguistic nature of epistemic signalling. It would be remiss at this point not to discuss one such method that has taken a radically different approach to epistemology and phraseology: that of Groom (2007). In a corpus-driven study of a set of grammatical words across two corpora Groom studies the semantic sequences constructed in order to identify disciplinary differences. Semantic sequences are understood here as coselections of elements of meaning that occur in a regular order that can be revealed through corpus analysis (cf. Hunston (2008) for a detailed discussion of the study of this phenomenon). Groom begins by identifying differences in grammatical words across corpora from different disciplines and then identifying the semantic sequences surrounding them. This ingenious technique

allows him to identify differences between disciplines in a quantitative way before studying the nature of these differences qualitatively, revealing an extraordinary amount of semantic detail showing the epistemic preoccupations of each discipline. Whilst there is a great deal to be admired in this approach it undoubtedly has a number of drawbacks in terms of implementing it for the purpose of the present thesis.

Most crucially the methods developed by Groom support cross-disciplinary comparison in disciplinary epistemology; whilst these findings reveal a wide-range of semantic sequences that are of epistemic import in the disciplines studied, these structures are divorced from or abstracted from any particular terms or constructs in the corpus. Whilst this is perhaps a strength within Groom's work this approach does not sit comfortably with the history of the study of epistemology within the philosophy of science, where the epistemic status of a given fact or purported entity is studied over time. In the present thesis what is sought is an approach that can combine the insights of the philosophy of science and social epistemology with those that can be wrought from a corpus of texts from one specific scientific discipline. The attempt to find such an approach will be presented in Chapter 4.

3.7 Conclusion

In this chapter I have reviewed and discussed the works that I believe are most relevant to the linguistic aspects of this thesis. This thesis has been located intellectually as an attempt to build upon previous linguistic studies of scientific language such as those of Halliday (1988), Myers (1989;1990;1991;1992;1994), Hyland (1998) and Hunston (1994) and to do so in such a way as to make linguistic findings that can contribute towards the wider understanding of social epistemology. In order to do so I will employ the concepts and methods of corpus linguistics and in particular the notions of concordance, collocation and colligation were discussed (3.3); I will return to these in the following chapter where the methodology of the study will be set out. In order to connect the study fully to previous similar approaches a number of studies applying the techniques of corpus linguistics to scientific writing and academic writing more generally were then discussed. The corpus-based and corpus-driven distinction was then introduced and it was argued that a corpus-driven approach is preferable if genuinely new findings about the linguistic nature of knowledge signalling in genetics are to be achieved in the present work. It has been established that an attempt will be made to connect the scholarship in the fields of philosophy of science and in particular social epistemology to a study of texts from genetics that utilises the concepts and methods of corpus linguistics; the attempt to find such an approach will form the subject of chapter 4.

Chapter 4: Methodology

4.1 Introduction

Perhaps the key challenge for this thesis was to create a plausible methodology linking the theoretical aims of social epistemology with the tools of corpus linguistics. In order to trial such a methodology a pilot study explored connections between pattern and epistemology and provided some extremely promising example of the kinds of finding that the final thesis might hope to deliver. A discussion of this is therefore provided as an illustration of the type of analysis that is being sought in this thesis and forms in effect the rationale for what takes place thereafter: the production and investigation of a corpus that is highly specific to a narrowly defined discourse community that it is hoped can provide sufficient data, spread across a sufficient period in time, to draw plausible and empirically supported conclusions about the knowledge signalling practices of that community.

4.2 Pilot study

4.2.1 The pilot corpus

The corpus used for the pilot study, hereafter called *genepilot*, was arrived at through the official Human Genome Project website, which contains a section entitled 'research archive' with links to 'landmark papers'. Wherever possible, these links were followed and the papers they lead to were copied and pasted into raw text files in the unicode format, and then saved into a file named *genepilot*. By following this process it was

hoped that corpus could be arrived at which could be said to be genuinely representative of the discourse of the Human Genome Project, whilst at the same time being of sufficient size to enable the investigation of lexical items in a way that utilises the understanding of pattern and meaning developed in corpus linguistics.

As a result of this process *genepilot* contained 107 texts from the journals *Nature* and *Nature Genetics* from 1999 to 2006 that specifically related to the Human Genome Project. The texts were apparently of many different types, falling into as many as nine different categories according to the headings given to them in *Nature*; *article*, *concepts*, *news feature*, *analysis*, *letter*, *review*, *brief communication*, *news and views*, and *short report*. Despite this high number of categories, most of the texts fell into just three of these *letters* (51), *articles* (30), and *brief communications* (15), with their being just one example of each of the remaining categories, with the exception of *news and views* (4 texts) and *review* (3 texts). I decided to use all of these texts, despite the apparent variety in genre, but resolved to be aware of the potential issues surrounding using such a wide variety of text types, with the most obvious being that any variation in language use discovered might simply be explicable by this presence of different genres, rather than by factors surrounding the construction of knowledge. In order to allow for this each text file was labelled not only by the name of the author (or, where there were multiple authors, which is typical in scientific writing, by the surname of the first named author), but also by the category into which the editors of *Nature* had placed the piece. In addition to this each text was also labelled by the year it was published, so that the filename has in effect three parts, as shown by these examples: 'LAN01_A', 'RAG04_L', and 'NUS06_L'.

This format allows the analyst to see both the genre of the text and the year the text was published when looking at concordance lines. In addition to this, the corpus was also organised into nine separate folders, corresponding to the nine different text types, giving the analyst the opportunity to quickly isolate all of the texts of a particular type. The files within each folder were then also subdivided into folders based on the calendar year that they were published, again allowing the analyst to isolate a particular set of texts quickly, but this time on the basis of publication date rather than genre.

Once *genepilot* had been compiled and organised in this way the next step before an investigation could begin was to ask whether any 'clean up' of the data was required. Since the texts had been cut and pasted in exactly the form that they were found in the journal *Nature* they contained a number of features that might be thought irrelevant to the current study, including certain headers and footers, labelling of diagrams and bibliographical data, and it is a fairly common practice in corpus linguistics to remove certain of these elements before serious study of the data begins. Given that the position of this thesis, generally, is that any element of a text that contributes to the reading practices within which that text is understood ought to remain present in a corpus if the corpus is to be truly representative of the text it purports to contain, it is clear that very little 'clean up' would be expected to take place. Indeed, since the texts were obtained manually by cutting and pasting the texts in as close a form as possible to that which they appear in the electronic form of *Nature* it is clear that, in order to be consistent with this principal, nothing should be removed from the texts at all. However, in practice, once corpus investigation began it quickly became clear that the inclusion of the

bibliographical section of the texts created a number of problems. Principal amongst these was that reference to certain texts within the bibliography section was so common amongst the 107 texts present that certain modes of investigation, such as cluster lists and collocation profiles, would be dominated not by features from the main body of information in each text, but by data from the bibliography. This problem is compounded by the formulaic nature of the bibliography entry: given that every reference to a particular text will (or certainly should) appear in exactly the same form, it is clear that corpus analysis tools, which were essentially created in order to recognise consistent formal patterns within large numbers of texts, will generate data that recognises these structures as being amongst the most salient features of the corpus overall. Whilst the study of the citation and referencing of other scientific texts may be highly informative, and is no doubt a valuable field of enquiry in its own right, the intention of this thesis is to focus upon the construction of discourse objects within the main body of the texts themselves. As such, it was decided that the main version of the corpus that would be used for this study would have the list of references removed from it, though the original corpus, including this data, would also be retained, allowing for future reference to or study of this data. Unfortunately this left only a small corpus comprising merely 560, 972 tokens. Whilst this is probably not a large enough set of data upon which to draw conclusions about the behaviour of lexis it was sufficient to enable a heuristic process such as the one intended.

4.2.2 Extracting an item for further investigation

In order to make a start in investigating the data Keywords were used, with the *British National Corpus* being used as a reference corpus to generate keyness statistics using the wordlist for *genepilot*. The following is a list of the top 18 items generated by the Keywords tool from the WordSmith Tools 4 (Scott 2004) suite of corpus investigation software:

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness
1	#	46204	8.2144	2E+06	1.613	77843
2	CHROMOSOME	2498	0.4441	445		23399
3	SEQUENCE	3168	0.5632	4211		22809
4	GENOME	2224	0.3954	201		21669
5	GENES	2249	0.3998	2073		17351
6	GENE	2267	0.403	2231		17289
7	HUMAN	3191	0.5673	19275	0.0194	14937
8	MOUSE	1239	0.2203	1849		8702.1
9	MB	815	0.1449	143		7640.1
10	CHROMOSOMES	862	0.1533	391		7382
11	SEQUENCES	1033	0.1837	1423		7379
12	FIG	1473	0.2619	7762		7248.5
13	KB	774	0.1376	335		6666
14	REGIONS	1170	0.208	4189		6548.2
15	SUPPLEMENTARY	859	0.1527	981		6370.4
16	GENOMIC	603	0.1072	151		5495.3
17	MUTATIONS	648	0.1152	458		5219.8
18	DNA	932	0.1657	3369		5200.4

Figure 4:1: *genepilot* Keywords

In order therefore to proceed with some (limited) empirical investigation the remainder of this project focused exclusively upon the Keyword *genome*, chosen partly due to its position close to the top of the 'keywords' list, partly due to its obvious centrality to the subject matter of the Human Genome Project. As figure 4:1 above illustrates, there are 2224 tokens of the keyword *genome* in *genepilot*. Whilst this is by no means an unmanageable amount of data in the sense that the Keywords list above might be, it is hardly a trivial exercise to analyse 2224 examples of a lexical item. Moreover it would seem likely that genome will participate in a number of lexical items within the 2224 examples, rather than each of these constituting the same complete lexical item. In order to search for such lexical items the clusters feature of *WordSmith Tools* was used to investigate the existence of formulaic phraseology within which *genome* might be discovered to function. The following (figure 4:2) is a list of 3-gram clusters within 5:5 range of the node within the concordance of *genome* in the *genepilot* corpus:

N	Cluster	Freq.	Length
1	THE HUMAN GENOME	458	3
2	DRAFT GENOME SEQUENCE	192	3
3	OF THE HUMAN	183	3
4	THE DRAFT GENOME	165	3
5	OF THE GENOME	152	3
6	IN THE HUMAN	138	3
7	THE MOUSE GENOME	116	3

8	IN THE GENOME	68	3
9	THE WHOLE GENOME	52	3
10	HUMAN GENOME SEQUENCE	49	3
11	IN THE DRAFT	49	3
12	OF THE MOUSE	49	3
13	SEQUENCE OF THE	46	3
14	OF THE DRAFT	43	3
15	THE GENOME WIDE	43	3
16	ACROSS THE GENOME	41	3
17	HUMAN GENOME THE	40	3
18	THE GENOME SEQUENCE	40	3
19	HUMAN GENOME PROJECT	36	3
20	REGIONS OF THE	33	3
21	THE DOG GENOME	30	3
22	THE CHIMPANZEE GENOME	29	3
23	HUMAN GENOME AND	29	3
24	GENOME SEQUENCE AND	29	3
25	A GENOME WIDE	28	3
26	HUMAN GENOME SEQUENCING	27	3
27	WHOLE GENOME BAC	26	3
28	WHOLE GENOME SHOTGUN	26	3

Figure 4:2: 3-part clusters from *genepilot*

The first thing to be noted about such a list is that many of these three word clusters or tri-grams are very likely to form part of larger clusters, and might therefore also be seen as sections of four or five word clusters. For example, the cluster *in the human*, which is present 138 times in the corpus, has *genome* as an R1 collocate in 133 of the 138 instances found, and can thus be seen rather as forming part of the four-part cluster *in the human genome*. Of the remaining clusters the frequency of some would appear to be explicable due to their frequent use in titles or names of either the overall project or certain sections of the project: examples of these would be clusters such as *human genome project*, and the examples that refer to the creature that is being focussed on within a particular study such as *the dog genome*, *the chimpanzee genome*, *the mouse genome* (and also *of the mouse*). Even *in the human* would appear to fall into this apparently very common semantic sequence. Although clusters such as these might well prove interesting upon closer examination, it was decided that the most interesting cluster for further investigation in a small study such as this was *whole genome shotgun*: principally because 'genome shotgun' would appear prima facie to be a piece of scientific terminology, the study of which has traditionally been regarded as crucial in epistemology and in the philosophy of science. However, it is again acknowledged that the basis of this choice is somewhat arbitrary, with the need to choose an object that is both of theoretical significance and of a manageable size for small-scale analysis being central at this stage.

4.2.3 Exploring an item using the techniques of corpus linguistics

In order to find a methodology for the main thesis two main elements were of interest, each due to their highly significant explanatory power. Firstly, synchronic analysis of a corpus can bring very frequent elements to our attention, with the assumption being that such elements are likely to be highly significant. By employing a text perspective on the data found, the analyst can remain sensitive to variation in data that exists as a result of the textual position of the individual datum. Whilst some findings in both the study of scientific writing and the social construction of knowledge relate to features of texts that do not necessarily vary in their discourse function over time (strings such as 'it is argued that', 'it was found that', 'it is unlikely that'), much of the study of scientific knowledge in both sociological and philosophical contexts has centred on the development of scientific concepts and purported entities over time. The claim made by Latour and Woolgar (1979) that 'facts' develop and become entrenched in language over time by a linguistic process of demodalisation can provide us with a basis to proceed empirically, albeit that Latour and Woolgar acknowledge that there is no straightforward relationship between linguistic form and the level of facticity that they assign with their statement types (1979: 76-81). This claim by Latour and Woolgar can in principle be investigated by looking at the discourse, and the purpose of the pilot study was to combine these three perspectives in order to generate robust claims about the construction of knowledge in the discourse. Indeed, given that these two different elements in a sense constitute separate tests as to the significance and verity of any claim about the construction of knowledge it is further hoped that the triangulation provided by these approaches will form a very

rigorous basis within which claims about the data can be made, since where a linguistic feature is supposed to constitute a significant example of the social construction of knowledge the analyst can variously ask the following questions: is this object of study common or typical in the data? Can a number of significantly frequent examples of alternate construction of this object be found in the data? (synchronic perspective); is the variation found genuinely a case of the epistemological status of the object changing over time or (another possibility) being differently constructed by different groups, or is it merely a textual feature dependent upon the various positions in the text in which it is situated? (text perspective). Claims about the linguistic construction of knowledge that could meet all of these requirements would be subject to what could be seen as a rigorous process of potential falsification, and would constitute an empirically robust basis on which to make claims about the development of knowledge in a discipline that could in principle be applied to any discipline. It is to the search for such techniques that I now turn.

4.2.4 Investigating an item: The synchronic perspective

The next step taken was to examine the salient collocates of *shotgun*. On the right side of the node word *shotgun* by far the most common collocate was *sequencing* (47 instances as a right side collocate), with *sequence* (16 instances as a right side collocate) also figuring prominently. In these instances *shotgun* appears to be acting as a classifier of *sequencing* or *sequence*, denoting the type or form of sequencing that is carried out. Other R1 collocates such as *strategy* (9 instances as a right side collocate), *approach* (9

instances as a right side collocate) and *coverage* (9 instances as a right side collocate) would appear to form a unit with *shotgun* that is more or less synonymous with *shotgun sequencing*, and *data* (7 instances as a right side collocate) would appear to refer to the findings generated by that procedure. This gives a total of 102 instances of this group of collocates within the 5 span to the right of the 109 instances of *shotgun*. Though in some cases more than one of these may be present in a single concordance line, since some of these collocates may be mutually inclusive, as in the case of *shotgun sequencing data*, it is certain that at least 78 of the 109 instances of *shotgun* collocate with one of these, since this is the total number of instances in which they appear as an L1 collocate, where they must of course be mutually exclusive, since only one word can appear in the L1 position in any given concordance line. Given that these collocates appear to be more or less synonymous, and that they occur in at least 78 and possibly as many as 102 of the right side contexts of *shotgun*, it seems reasonable to assume that the right side collocates of *shotgun* function to form a unit of meaning that is more or less monosemous.

However, the left side collocates of *shotgun* appear to demonstrate a greater variety of meaning. Given that *shotgun* was chosen from the concordance data of *genome* it is of course not a surprise that *genome* features prominently. It is also not surprising that the frequency of *whole* in the L2 position matches that of *genome* in the L1 position, due to the high frequency of the cluster *whole genome shotgun*. However, what can be observed in the collocation data that cannot be gleaned from a cluster list is that *whole genome shotgun* accounts for 26 of the 28 occurrences of *whole* within a 5:5 span of the node word *shotgun*, apparently indicating a highly formulaic relationship between *whole* and

shotgun. Moreover, inspection of the two remaining occurrences of *whole* reveals that all instances of *whole* occurring within a 5:5 span of shotgun concern *whole genome shotgun*. Firstly, the L4 occurrence of *whole* occurs in the context 'whole-genome or hierarchical shotgun sequencing', where *whole-genome shotgun* can be seen as being 'split' by an alternative modification (indicated by *or*) of *hierarchical*. Secondly, the R4 occurrence of *whole* occurs in the context 'clones are then selected for shotgun sequencing and the *whole* genome sequence is reconstructed'. In this context the occurrence of *whole* is again in the L1 position of *genome* (as indeed all 28 instances of *whole* in the concordance of *shotgun* are), and can be seen as clearly and explicitly related to the meaning present in all other instances of *whole* since it is functioning here as explaining the meaning of the cluster *whole-genome shotgun*.

In addition to *whole* it must be noted that the *WGS* that also appears in the collocate list is an abbreviation of *whole genome shotgun*. Thus we find that *whole genome shotgun*, in the form *WGS*, is itself a collocate of *shotgun*. Moreover, since *WGS* straightforwardly refers to *whole genome shotgun* it can be seen that there may be additional instances of *shotgun* in *genepilot* in the form *WGS*. Concordance data for *WGS* was subsequently generated and it was discovered that there are 55 instances of *WGS* in *genepilot*; considerably more than the 7 instances that occur as collocates of *shotgun*. This discovery demonstrates the reiterative nature of corpus linguistics: as more discoveries about the lexis are made, and more connections are formed, new avenues of potential enquiry appear. I decided that due to the scope of the pilot study *WGS* would not be analysed in the same way as *shotgun*, particularly since the 55 instances of *WGS* are accounted for by

just 4 texts (she04, tar05, lin05 and chi02). Rather, *whole genome shotgun* will be taken as being representative of both *whole genome shotgun* and *WGS*, though it is clear that certain of the conclusions that may be drawn with regard to *shotgun* will require careful formulation in respect of the existence of *WGS*. Indeed, cross-referencing results with the *WGS* concordance lines may be essential if claims about the occurrences of *whole shotgun sequence* throughout the corpus are to be made. There also remains the question as to whether *WGS* is used in the same way as *whole shotgun sequence*; a plausible hypothesis is that *WGS* occurs in texts where *whole genome shotgun* has already occurred, effectively referring back to the full-form.

A second prominent left side collocate of *shotgun* is *hierarchical*, appearing 16 times in that context. All 16 of these occurrences of *hierarchical* occur in the L1 position, marking them as mutually exclusive (at least in form, if not in meaning) from the 26 instances of *whole* that occur in the L1 position. Similarly to *whole*, *hierarchical* occurs almost exclusively as a left side collocate. Indeed, the one instance of *hierarchical* being on the right side of the node word, is in the context 'to perform shotgun sequencing on these intermediates (*hierarchical shotgun*)', where *hierarchical* is actually in the position of an L1 collocate of *shotgun*, but within a 5:5 span of the previous token of occurrence. Given that this is the case, *hierarchical*, like *whole* and *WGS*, would appear to occur only in a highly formulaic way in the context of the node word *shotgun*.

Finally a third prominent left side collocate of *shotgun* is *clone*, occurring 19 times as a left side collocate. Since *clone-by-clone* was identified as one of the most frequent 3 word clusters in the 5:5 span of *shotgun* it is no surprise that inspection of the concordance

lines reveals that 14 of these 19 occurrences are accounted for by the sequence *clone-by-clone shotgun*, and this is again reflected in the collocation data since *clone* appears 8 times as an L3 collocate and 7 times as an L1 collocate, with *by* occurring 8 times as an L2 collocate. It should again be noted that the collocation list proves a more useful tool for analysis than the cluster list in that it demonstrates the proportion of the overall occurrences of *clone* that occurs in the cluster *clone-by-clone*: which is 14 out of 19 instances. From this it can be inferred that *clone-by-clone* is indeed clearly the most salient context within which *clone* occurs, and that it can again be said that like *whole*/*WGS* and *hierarchical* the word *clone* appears in a very formulaic relationship with the node word *shotgun*.

From the above it would appear at this point that there are three principal textualisations of shotgun: *whole/WGS*, *hierarchical* and *clone-by-clone*. Moreover since these occur in fixed patterns to the left side of *shotgun*, apparently modifying or classifying the type of *shotgun sequencing* that takes place, these would appear to be strong candidates for alternative construals of *shotgun*, with the different multi-word items each disambiguating *shotgun*, and thus revealing a monosemous unit: *whole* shotgun, *hierarchical* shotgun and *clone-by-clone* shotgun. At the synchronic level corpus investigation appears to indicate these three principal ways that *shotgun* is likely to be used in *Nature*, with *whole/WGS*, *hierarchical* and *clone* accounting for 64 of the 109 instances of *shotgun*. What has not been indicated by the collocation data is the relationship of meaning between these different construals of *shotgun*: in principle it is possible that these are all synonymous, or that two of the three are synonymous with one

having a different meaning, or again that all three have different meanings and refer to different forms of '*shotgun sequencing*'. Whilst a geneticist with previous knowledge of the discourse surrounding shotgun may already know what the relationship between these discourse objects is, the non-specialist linguist must either look to the corpus or the wider discourse concerning these lexical items (such as that in dictionaries of terminology) in order to further comprehend the relationship between these three items, and it is to the former of these two possibilities that I now turn.

4.2.5 The synchronic perspective continued: qualitative analysis of expanded contexts

In order to attempt to understand the semantic relationship held between these competing forms of *shotgun* I carried out a qualitative investigation of *shotgun*. In order to do this I created a text file containing expanded contexts of each of the 109 instances of *shotgun*, with the hope being that some wider patterning of these discourse objects would distinguish their semantic relations. The expanded context in this sense refers to the entire sentence that the node word occurred in, rather than just that part of it that is visible in concordance lines. The examples below (4:1-4:4) illustrate how this file of expanded contexts appears to the analyst:

> 4:1 In **the whole-genome shotgun method**, sequence would
> necessarily come from two different copies of the human genome
> lan01 15/02/01

4:2 A biotechnology company, Celera Genomics, has chosen to incorporate **the whole-genome shotgun approach** into its own efforts to sequence the human genome lan01 15/02/01

4:3 combining some coverage with **whole-genome shotgun data** generated by the company lan01 15/02/01

4:4 If the raw sequence reads from **the whole-genome shotgun component** are made available lan01 15/02/01

Examples 4:1-4:4: Expanded contexts surrounding the node word *shotgun*

This qualitative analysis revealed the semantic relationship between the main three different forms of *shotgun* in a surprisingly straightforward way: that relationship was explicitly stated in the expanded context. Thus it was subsequently discovered in the cotext of these concordance lines that *whole* and *hierarchical* are mutually exclusive alternate versions of *shotgun sequencing*, whilst *clone-by-clone* is synonymous with *hierarchical*. Examples of expanded context that reveal this relationship include (all italics are mine):

4:5 There was lively scientific debate over whether the human genome sequencing effort should employ *whole-genome or hierarchical shotgun sequencing* lan01 15/02/01

4:6 A principal issue in the sequencing of large, complex genomes has been whether to perform *shotgun sequencing* on the entire genome at once (*whole-genome shotgun*, *WGS*) chi02 05/12/02

4:7 or a combination of WGS and *hierarchical shotgun sequencing* (including those of Drosophila melanogaster50, human2 and rice51) chi02 05/12/02

4:8 This issue is better addressed through *hierarchical shotgun* than *WGS* sequencing chi02 15/02/02

Examples 4:5-4:8- Examples of expanded context that reveal the relationship between *whole, hierarchical and clone-by-clone*

Even in the four examples given above a great deal of information about the relationship between the different construals can be seen. Whilst *whole*/*WGS* and *hierarchical* are principally presented as mutually exclusive, as in the example above from the text lan01 and the text chi02 it also becomes apparent that these approaches can be complementary, as indicated in chi02 'or a combination of *WGS* and *hierarchical shotgun sequencing*'. What is particularly promising about these examples is that in each case where the relationship between the two different construals is being made explicit, the different construals occur within a 5:5 span of each other. This suggests that, once potentially competing discourse objects such as these have been discovered, the semantic relationship between them can be discovered by examining instances where the

purportedly competing discourse objects appear as collocates of each other. This is a very straightforward task when using software such as *WordSmith Tools*, since all that is needed is to generate concordance data for one of the discourse objects, and then search the collocates list or concordance data for the alternative construals. Though this phenomenon may of course not be present in all discourses and for all types of discourse objects, the general point, that competing discourse objects will appear as collocates of each other when the relationship between them is being made explicit, is one that is likely to be extremely useful to discourse and corpus analysts of many different types.

In addition to this it was also discovered that the semantic relationship of *hierarchical* and *clone-by-clone* is one of synonymy. This was again discovered from looking at the expanded context of concordance lines, where the following statement was discovered 'second is the '*hierarchical shotgun sequencing*' approach, also referred to as '*map-based*', '*BAC-based*' *or* '*clone-by-clone*' *'* (lan01). Moreover, as can be seen, two more apparently synonymous terms are discovered, namely *map-based* and *BAC-based*, demonstrating a key difficulty in studying terminology in (arguably) any field: the existence of (supposedly) synonymous words referring to the same (discourse external) object. In this case the discovery that *BAC-based* and *map-based* are additional textualisations of *hierarchical* and *clone-by-clone* has only been made after careful investigation of every (extended) concordance context and even then only because one writer, on one instance, explicitly drew attention to these alternatives. What this demonstrates is that the linguist must be extremely careful about the claims that are made on the basis of investigating words at the formal level only, and without recourse to the context within which those words are placed. Moreover, though this demonstrates the usefulness of such qualitative

analysis, it is clear that the problem of synonymy is a significant one for the corpus linguist, and that the claims of the corpus linguist must be carefully limited to the particular form studied.

4.2.6 Summary

In the pilot study it was possible to arrive at both clearly expressed epistemological findings (such as the nature of the relationship between the set of shotgun sequencing terms). In methodological terms what is crucial here is that the techniques used to investigate any given item must be highly flexible. When one sets out from a 'corpus-driven' approach one must inevitably vary the exact method of exploration used for each item. There must be a symbiotic relationship between data and methodology in corpus-driven corpus linguistics, where methodology is best viewed as a set of techniques which may or may not be employed depending on what is discovered when one turns to the data. Each technique that is used will generate a new subset of data from the overall corpus, and what is discovered at each stage may vary considerably from one study to another, and from one type of linguistic object to another. In the above case what was discovered was a terminological item pertaining to genetic methodology, where the epistemic issue present was a range of competing methodology. As such what was needed was a technique (in this case collocation analysis of the 'root' part of the cluster *whole genome shotgun*) that revealed this underlying relationship. The non-specialist analyst cannot have any inkling of either the linguistic or the epistemic relationship that will be

discovered; instead, a process of exploration must reveal the specific linguistic patterning, with the role of the analyst being to choose the correct tools to reveal this pattern and then suggest a plausible epistemic interpretation of what has been found. By moving through the steps of collocation analysis and qualitative analysis of expanded contexts it is to be hoped that similarly significant findings will be found on examination of the final corpus.

4.3 Final corpus construction

Here I follow Groom (2007) by using a framework proposed by Flowerdew (2004) which poses the follow questions which provide a useful list of considerations when designing a specialised corpus:

1. What is the purpose for building a specialised corpus?
2. What genre is to be investigated
3. How large should the specialised corpus be?
4. Is the specialised corpus representative of the genre?
5. How will data be collected?
6. How will the specialised corpus be tagged/marked up?
7. What kind of reference corpus would be suitable to contrast with the specialized corpus?

In what follows each of these will be discussed.

4.3.1 Purpose

The purpose of constructing the corpus is to provide a source of data that will allow for a detailed investigation of the knowledge signalling practices of genetics. The corpus will be fit for purpose if it provides data of sufficient scope to detect significant patterns in the discourse and in sufficient number to provide the basis for detailed study from a synchronic perspective. In each of the following considerations it is the requirement that the corpus fulfils this aim that determines how exactly what will be needed from the corpus construction stage.

4.3.2 Genre

There is a considerable tradition focussing on the research article in investigating scientific discourse, and indeed the research article is seen as the key genre in which new knowledge is communicated. In this study I take a slightly more nuanced approach to genre, based on an investigation of the text types found in the leading journals in the field, *Nature* and *Nature Genetics*, that took place in the pilot study. The position taken here is that whilst there may be significant generic differences between text types such as *Articles* and *Letters* when compared to others such as *News and Views* or *Analysis*, all of these differing genres are potentially of interest when the purpose of the study is to discover how geneticists signal the epistemic status of their findings. However, the study does not extend to including other generic types such as lecture materials, textbooks or articles in the popular press such as newspapers or magazines. The rationale for this is

straightforward; this thesis seeks to explore not the dissemination of scientific knowledge to the wider community, but the original process of encoding that knowledge within the research community.

4.3.3 Size

The position of this thesis is in a sense straightforward as regards corpus size in that it follows Sinclair (1991; 2004) in assuming that there is no maximally ideal size for a corpus, particularly when lexical features will be the focus of study. However the corpus building process is of course only one small element of the doctoral research project and the amount of time devoted to corpus construction must be commensurate with this. In practice approximately three months of corpus building yielded a corpus of 2,979 texts and approximately 10 million tokens and it was felt that this would be sufficient to provide sufficient data to sample and investigate lexical items in some detail. Moreover this constituted everything that had been published in the journal *Nature Genetics* over the ten year period 1999-2008 and this was in that sense the maximum obtainable whilst focussing narrowly on just one journal. Groom (2007) has already drawn attention to the argument of Sinclair (2005) suggesting that a much smaller corpus is needed for the study of specialized technical vocabulary as opposed to general language studies. This being the case it is merely necessary to ensure that the frequency of the given lexical items under study is sufficient for detailed analysis, and in what follows I attempt to show that this is indeed the case.

4.3.4 Representativeness

In terms of representativeness the needs of the present study are to ensure that the data chosen is a plausible representation of the research community of genetics. This goal is usually achieved in corpus linguistics by means of sampling a number of different journals in order to counterbalance any biasing effect of the house style of any one journal. The approach adopted here, however, has been quite different from this in focussing on just one (albeit very prestigious) journal in the field. Whilst texts for the pilot corpus were taken from two journals (*Nature* and *Nature Genetics*) the final corpus for the thesis will focus on just one, namely *Nature Genetics*. The reason for this is very simple. The aim in a lexically (terminologically) driven approach is to maximise the instances of terminological lexis rather than sample and compare more general linguistic features (such as the use of modal verbs or the passive) across different journals. The question is then not what features are similar to all journals but rather what features typically occur in the vicinity of terminological items. It would be almost miraculous if a sampling of journals achieved the required effect of maximising the number of instances of any given item. In terms of representativeness this potentially poses a problem for this thesis. However, it is surely not implausible to suggest that the practices of the most prestigious journal in the field are likely to be highly representative of the field generally. Meanwhile the issue of whether the corpus accurately represents the discourse of *Nature Genetics* is straightforwardly solved since it contains every single such text over a ten year period. Perhaps the only approach that could be argued to achieve the intended data capture whilst being even more representative of the discourse of geneticists would have

been to collect the maximum amount from a number of or even every journal of genetics; but this is simply outside of the scope of this doctoral study.

4.3.5 Corpus annotation

This thesis follows the raw text 'corpus-driven' approach of linguists such as Sinclair (1991), Tognini-Bonelli (2001), Hunston and Francis (2000) Groom (2007) and Oakey (2008) in assuming that annotation practices may obstruct the process of discovering new aspects or details about a language variety. This position is particularly crucial in the present thesis since it is intended to explore the language of a highly specialised discipline in order to arrive at both linguistic and theoretical advances in how the construction and signalling of scientific facts takes place. It is not at all clear how a corpus-based approach could deliver such a goal, since the items for study are typically chosen in advance.

4.3.6 Reference corpus

Whilst the whole corpus keywords used do not in fact differ from the top ten lexical items by raw frequency it is worth outlining briefly the reference corpus used to generate keywords. This thesis uses BNC World as a reference corpus which is made up of 82.82% written texts of a wide range of genres and disciplines. The following table taken from the BNC website summarises the composition of this corpus:

Text type	Texts	Kbytes	W-units	S-units	percent
Spoken demographic	153	4206058	4.30	610563	10.08
Spoken context-governed	757	6135671	6.28	428558	7.07
All spoken	910	10341729	10.58	1039121	17.78
Written books and periodicals	2688	78580018	80.49	4403803	72.75
Written-to-be-spoken	35	1324480	1.35	120153	1.98
Written miscellaneous	421	7373707	7.55	490016	8.09
All written	3144	87278205	89.39	5013972	82.82

Figure 4:3: Composition of BNC World (taken from BNC website)

4.4 Data collection and the corpus

The corpus, *genecorp*, comprises 2,979 texts from the journal *Nature Genetics*. These constitute all of the texts from this journal over the ten-year period 1999- 2008 inclusive. The procedure for collecting and organising the corpus was the same as for *genepilot* as described above in section 4.2.1. and the description of this process is therefore not repeated here. What does need discussion here is the issue of the removal of references from the corpus, which whilst approximately 90% successful did not remove every single example of references sections. The references were removed using the procedure described in 4.2.1 above whereby all elements following the strings *References*, *REFERENCES* and *references* were automatically deleted using a simple command script. Upon inspection of a number of example texts this appeared to have been entirely successful, however it became apparent much later that a small proportion (roughly 10%) of the texts had apparently not been operated on successfully by the command script, despite these always beginning with the strings *References* or *REFERENCES*. Since the process of manually checking for and removing these by hand from all 2,979 texts in the corpus would have been time consuming in the extreme it was decided that the simple expedient of ignoring concordance lines containing references sections would be used in order to maintain a consistent investigation of the corpus.

4.5 Extracting useful items for study

The final methodological issue to be decided upon was the exact technique to be used in choosing objects for further study. Whilst this might seem a relatively straightforward

task there are in fact a great number of different ways in which this might be done, none of which are necessarily the exclusively correct means of generating the data for investigation. In what follows a range of potential techniques are discussed, with the connection between them being that each offers a plausible connection between the wider discourse study aims of identifying epistemologically significant and interesting objects for further study with the corpus-driven aim of allowing the choices to be data-driven and not reliant on the ability of the analyst to intuit which items are significant and interesting.

4.5.1 Whole corpus keywords

The method of identifying lexis for further investigation through whole corpus keywords was considered. The following list of the top thirty keywords generated using BNC World as a reference corpus shows the type of discourse objects that would form the starting point of corpus investigation when a whole corpus keywords approach is used:

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P
1	#	992073	9.11619	1604421	1.613	2E+06	2E-28
2	FIG	36745	0.33765	7762		130762	4E-

							25
3	CELLS	35961	0.33045	7646		127807	4E-25
4	GENE	29058	0.26702	2231		119074	5E-25
5	GENES	28230	0.25941	2073		116162	5E-25
6	MICE	24532	0.22543	1020		105348	7E-25
7	EXPRESSION	28409	0.26105	7231		97230	9E-25
8	CELL	22381	0.20566	5418		77431	2E-24
9	DNA	19999	0.18377	3369		74109	2E-24
1	FIGURE	27015	0.24824	17214	0.017	69662	2E

0			1901		3		-24
11	PROTEIN	17732	0.16294	2898		66034	3E-24
12	MUTATIONS	14895	0.13687	458		65002	3E-24
13	PUBMED	12531	0.11515	0		58069	4E-24
14	GENOME	12959	0.11908	201		58014	4E-24
15	CHEMPORT	11913	0.10947	0		55204	5E-24
16	C	28327	0.26030	31384	0.0316	55186	5E-24
17	ANALYSIS	21125	0.19412	13130	0.0132	55026	5E-24

18	PCR	12060	0.11082	362		52688	5E-24
19	SUPPLEMENTARY	12303	0.11305	981		50215	6E-24
20	MOUSE	12911	0.11864	1849		49076	7E-24
21	DATA	20855	0.191637427	18084	0.0182	46622	8E-24
22	ALLELE	9954	0.09147	84		45171	9E-24
23	SEQUENCE	13739	0.12625	4211		44984	9E-24
24	USING	22585	0.20753	24434	0.0246	44633	9E-24
25	MUTATION	10383	0.09541	468		44352	9E-

							24
26	MUTANT	10377	0.09535	523		43996	9E-24
27	GENETIC	11436	0.10509	1827		42739	1E-23
28	ONLINE	10008	0.09196391	597		41904	1E-23
29	CHROMOSOME	9658	0.08875	445		41196	1E-23
30	THUMBNAIL	8960	0.08233	70		40713	1E-23

Figure 4:4: Top 30 keywords for *genecorp* using BNC World as a reference corpus

The first thing that must be noted from this keyword list is that there are a number of elements present that are entirely trivial and would not constitute sensible items for further study. The most obvious of these is the # symbol, which appears in the keywords list due to the inability of Wordsmith Tools to deal with certain characters including

numbers. This is of course merely a quirk of the software and is of no interest to the analyst. In addition to this there are certain words amongst the keywords that occur due to the standard format or mark-up of *Nature Genetics* texts but are again of no relevance to study of the discourse; examples of these would include *pubmed* and *chemport* which are merely proprietary names and do not form part of the text as such. Another keyword that is likely to be of no interest is *figure*, which, whilst illustrating the fact that geneticists refer to data in the form of tables and diagrams a great deal, is also of only trivial interest. However, these forms are very easily identified, and once these examples have been ignored the analyst is left with words such as the following: *cells*, *gene*, *genes*, *mice*, *expression*, *cell*, *DNA*, *protein*, *mutations*, *genome*, *analysis*, *PCR*, *mouse*, *data*, *allelle*, *sequence*, *using*, *mutation*, *mutant*, *genetic*, *chromosome*, *genetics*, *genomic* and *SNPS* as candidates for further investigation.

One approach to this keywords list would be to take the top ten keywords for further investigation. This would have the advantage of being a principled choice, based on the assumption that terms that are more key have a greater constitutive role, or put more simply are the most important in the discourse. This also has the advantage of making the choice replicable, as it is not reliant upon the choices or intuitions of the analyst. This also yields words that are of very high frequency, as is demonstrated in the following table:

Word	Frequency
cells	35,961
gene	29,058
genes	28, 230
mice	24, 532
expression	28, 409
cell	22, 381
DNA	19,999
protein	17, 732
mutations	14, 895
genome	12, 959

Figure 4:5: Top ten keywords from *genecorp*

The high frequency of these words suggests that they are likely to be present across the whole corpus, allowing for detailed investigation into the salient patterns in which the words occur. One concern about such words is that they may have very established uses in the discourse, with their introduction likely to precede the start of the corpus. However, whilst words that are as frequent as those in figure 4:5 are likely to be present across the entire corpus, this does not mean that there will be nothing to say about the patterns of collocation and colligation into which these words are embedded. In other words, even if the usage of a word appears to be stable across the corpus at the level of frequency, this does not mean that there is no variation occurring in terms of meaning.

Another alternative to choosing words from a keywords list purely on the basis or keyness is to introduce an element of choice from the analyst. The obvious problem with this is that it is difficult to make a principled choice of items, and this process therefore introduces elements of user intuition into the process at a very early stage. However, if investigation proceeding strictly from keyness rankings had proved to be of little interest even this approach could be taken in order to search for discourse objects that are more discipline-specific or that more clearly constitute new terminology in genetics.

4.5.2 Extracting discourse objects using lexical chains

Whilst whole corpus keywords can provide a number of very high frequency lexical items that are likely to be central to the discipline of genetics the problem of too much data still remains at this stage. Frequency statistics of between twelve and thirty thousand for the top ten keywords identified clearly constitute an excess of what the analyst can realistically deal with and therefore what is needed at this stage is to refine this approach focussing on specific aspects of these keywords. Lexical items are of interest since lexical chains are of particular use in identifying terminological items (Rogers, 2007, p. 17); high frequency lexical items would appear to be a plausible starting point for the discovery of terminology for further investigation and conveniently the WordSmith Tools cluster tool can be used in order to identify lexical chains in the vicinity of keywords with a simple piece of manual selection from the analyst. The following is a list of clusters between three and six tokens in length found within the 5:5 span of the node *gene*:

N	Cluster	Freq.	Length
1	OF THE GENE	996	5
2	OF GENE EXPRESSION	828	5
3	THE GENE ENCODING	644	5
4	GENE EXPRESSION IN	524	5
5	IN THE GENE	495	5
6	IN GENE EXPRESSION	442	5
7	MUTATIONS IN THE	380	5
8	OF A GENE	349	5
9	GENE IN THE	251	5
10	FOR EACH GENE	243	5
11	GENE EXPRESSION DATA	218	5
12	GENE EXPRESSION AND	206	5
13	THE GENE EXPRESSION	198	5
14	EXPRESSION OF THE	193	5
15	GENE EXPRESSION PATTERNS	191	5
16	GENE ENCODING THE	182	5
17	A SINGLE GENE	180	5
18	OF THIS GENE	179	5
19	GENE EXPRESSION PROFILES	158	5
20	AND GENE EXPRESSION	155	5
21	TUMOR SUPPRESSOR GENE	154	5
22	OF EACH GENE	151	5
23	THE SAME GENE	151	5
24	CHANGES IN GENE	151	5
25	A CANDIDATE GENE	141	5

Figure 4:6: List of clusters found in the 5:5 span of the node word gene in *genecorp*

This list illustrates the potential use of lexical chains to identify terminological items. Whilst there are strings of varying composition in this list, those that are most likely to fulfil the criterion of constituting discrete lexical items referring to specific terminological entities are those that contain the greatest number of constituent lexical parts. So whilst *of this gene* is undoubtedly an interesting phraseological entity it is not a discourse object in the sense that *gene expression data* or *tumor suppressor gene* are. Moreover, each of the items in this list that contains at least three lexical parts is of sufficient frequency to carry out a satisfactory collocation and concordance investigation. As a result of these observations a list was generated of the clusters surrounding each of the ten most key whole corpus keywords and the results of this process will be presented in chapter 5.

4.6 Conclusion

The foregoing has been an attempt to connect the theoretical and discourse analytic goals of the sociology of knowledge with the linguistic investigatory methods of corpus linguistics. I have argued that whilst there is no exclusively correct way of proceeding in this task there are a number of techniques for both extracting and investigating items that provide a plausible way of progressing. Two crucial considerations still face the analyst once a corpus has been constructed: what items should be studied, and how can they be studied in a rigorous manner? The first of these has been met by focussing on what has traditionally been regarded as the focus of epistemological study: the fact or ontological

entity. This has led to the choice of extracting lexically dense clusters for further investigation. The second of these has required the use both of collocation and concordance data but also the qualitative analysis of the expanded contexts of extracted items in order to reach a satisfactory analysis of the semantic relations found in the context of each items. The results of this process will be presented in what follows.

Chapter 5: Results

In what follows, the findings from the detailed investigation of strings containing at least three lexical elements and located within a 5:5 span of the ten highest keywords in *genecorp* is presented. The results of the concordancing and expanded contexts investigations described in sections 4.4.4 and 4.4.5 are reported below. For each phrase a table illustrating the most frequent twenty collocates is presented. Any notable aspects of this are discussed. This is followed by a discussion of any noteworthy findings from the investigation of expanded contexts. This will be limited to a brief description in cases where no frequent patterns or epistemic marking of interest were discovered. Where the findings are worthy of a more through discussion this will be provided in detail with a short summary at the end of the section to reiterate the key findings. Finally in chapters 6 and 7 there will be an attempt to build on the findings of this study by providing a detailed discussion of causation and ontological categorisation in genetics; to avoid duplication within the thesis the discussion of the clusters relating to these chapters will be omitted from chapter 5

.

5.1 The clusters

A search using the clusters function of *WordSmith Tools* for each of the ten highest keywords in *genecorp* and limiting the search at approximately the 100 occurrences mark gave the following 77 clusters for potential further investigation:

cells

1.	**wild type cells**	**267**	
2.	**embryonic stem cells**		**228**
3.	**cos 7 cells**		**199**
4.	**bone marrow cells**		**155**
5.	**stem es cells**		**145**
6.	**embryonic stem es cells**		**142**
7.	**cd8 t cells**		**109**
8.	**cos 1 cells**		**100**

gene

9.	**gene expression data**	**218**

10.	**gene expression patterns**	**191**
11.	**mutations in the gene encoding**	**175**
12.	**gene expression profiles**	**158**
13.	**tumor suppressor gene**	**154**
14.	**changes in gene expression**	**122**
15.	**analysis of gene expression**	**97**
16.	**variation in gene expression**	**93**
17.	**gene expression profiling**	**89**

genes

18.	**tumor suppressor genes**	**138**
19.	**X linked genes**	**131**
20.	**protein coding genes**	**111**
21.	**differentially expressed genes**	**109**

expression

22.	**gene expression data**	**216**
23.	**gene expression patterns**	**189**
24.	**gene expression profiles**	**159**
25.	**analysis of gene expression**	**98**

cell

26.	**cancer cell lines**	**207**
27.	**lymphoblastoid cell lines**	**205**
28.	**mol cell biol**	**203**
29.	**es cell lines**	**142**
30.	**cell cycle arrest**	**137**
31.	**es cell clones**	**128**
32.	**cell cycle progression**	**112**
33.	**whole cell extracts**	**105**
34.	**cancer cell line**	**89**
35.	**planar cell polarity**	**82**
36.	**breast cancer cell**	**81**

DNA

37.	**DNA binding domain**	**155**
38.	**DNA copy number**	**150**
39.	**southern blot analysis**	**82**

protein

40.	**green fluorescent protein**	**198**
41.	**protein protein interactions**	**151**
42.	**protein blot analysis**	**146**
43.	**fluorescent protein GFP**	**125**
44.	**green fluorescent protein GFP**	**124**
45.	**protein coding genes**	**108**
46.	**protein protein interaction**	**99**
47.	**wild type protein**	**83**

mutations

48.	**loss of function mutations**	**219**
49.	**loss of function mutations in**	**123**
50.	**disease causing mutations**	**92**

genome

51.	**genome wide association**	**824**
52.	**wide association study**	**243**
53.	**wide association studies**	**237**
54.	**a genome wide association**	**172**

55.	**genome wide linkage**	**153**
56.	**genome wide significance**	**152**
57.	**national human genome**	**139**
58.	**human genome project**	**135**
59.	**human genome research**	**135**
60.	**human genome sequence**	**133**
61.	**genome research institute**	**127**
62.	**national human genome research**	**125**
63.	**human genome research institute**	**122**
64.	**the human genome project**	**100**

analysis

65.	**western blot analysis**	**607**
66.	**northern blot analysis**	**599**
67.	**southern blot analysis**	**558**
68.	**RT PCR analysis**	**260**
69.	**southern blot analysis of**	**213**
70.	**northern blot analysis of**	**208**
71.	**western blot analysis of**	**200**
72.	**RTA PCR analysis**	**185**
73.	**by southern blot analysis**	**154**
74.	**RT PCR analysis of**	**149**

75.	**blot analysis using**	**135**
76.	**RNA blot analysis**	**111**
77.	**by western blot analysis**	**107**

Figure 5:1: First list of clusters generated around each keyword using *WordSmith Tools*

As discussed above in the methodology, the urge to use an arbitrary cut-off point of 100 in each and every case was ignored in order to retain potentially interesting clusters that fell just short of this number. As a result the clusters *analysis of gene expression*, *variation in gene expression*, *gene expression profiling*, *cancer cell line*, *planar cell polarity*, *breast cancer cell*, *protein protein interaction*, *wild type protein*, *southern blot analysis* and *disease causing mutations* were all retained for further investigation despite there being between 82 and 99 occurrences of these. Perusal of this initial list of objects for further study shows what appear to be a range of terminological items including: methodological techniques such as *western blot analysis*, *RT PCR analysis*, and *analysis of gene expression*; a range of objects of study including purported entities such as *wild type cell*, *stem es cells* and *tumor suppressor gene*; processes such as *cell cycle arrest* and *cell cycle progression* and clusters already apparently constituting explicit epistemic marking such as a *genome wide association* and *genome wide significance*. As such, this would appear to be a very promising set of data for further investigation, despite several small problems with the initial cluster list that were as follows.

5.1.1 Problems with the initial cluster list

Whilst there is clearly a need for a cut-off point when thousands of clusters can be generated for each of the ten keywords it is equally clear that this cut-off point is ultimately arbitrary. It was argued in the methodology section that a cut-off point of a minimum of approximately 100 concordance lines would be appropriate, but also that there should be no psuedo-scientific commitment to insisting on a minimum of precisely 100 occurrences. Since what is being sought is examples that will provide surprising and enlightening information about epistmic marking in genetics it is clear that this is the priority: 99 interesting examples of a cluster would clearly be more valuable to this study than 100 examples where there is little or nothing to learn from the surrounding language.

Secondly there are a number of clusters that are repeated throughout the list. This has happened because many of the clusters generated contain more than one of the ten highest keywords, and therefore appear in the cluster list for both words, as in the examples *gene expression data* and *gene expression patterns*. This is a relatively small problem though, and in each of these cases any subsequent occurrence of a cluster in the list is simply removed as it would of course simply be the same object of study. There is also some repetition within keywords, where a cluster occurs with such high frequency that a further example of it appears with the most frequent collocate of that cluster forming a part of a new, longer cluster as in the case of *RT PCR analysis* and *RT PCR analysis of* or *southern blot analysis* and *by southern blot analysis*.

Thirdly the keen-eyed reader will note a slight discrepancy at times between the precise number of clusters revealed and the number that *WordSmith Tools* then returns when this cluster is entered into the concordance tool as a search term. Thus for the most frequent cluster surrounding cells (*wild type cells*) the cluster tool reports 267 instances whilst the concordancer reveals 258. Upon inspection of the relevant concordance lines I could find no obvious reason for such a discrepancy; however this was judged not to be a major problem, since the 'gap' between the figure for clusters and that for concordance lines was usually very small (the difference in the case of wild type cells, for example, is between 3 and 4 percent). As such the approach taken was to reveal explicitly in the description where such discrepancies occur (such as in the case of *wild type cells* in section 5.2.1) in the interest of transparency, and then to analyse the concordance lines given by the concordance, even if the number of these differed slightly from the number given by the clusters function.

Finally a taxing problem is that there remains too much data for detailed close investigation in a PhD thesis. Whilst the narrowing of focus to clusters containing a minimum of three lexical elements has significantly narrowed the focus of the study from the thousands of concordance lines of each of the keywords 77 clusters, each having on average over 100 concordance lines for detailed investigation is still likely to provide far more in the way of results than can be reported in a work of this size. One way to solve this problem would be to continue to raise the arbitrary cut off point for minimum number of occurrences of a cluster until such time as a number of clusters that could be reported in detail in the present work has been reached. Whilst offering a neat solution it

is argued here that this process would constitute the limitations of the present work coming to define the process of analysis to too great an extent. Rather, in the present work all of the clusters in the final list will be investigated, but the process of reporting will be far more limited in the case of those clusters which prove, upon further examination, to be of little interest in terms of the epistemic signalling that typically occurs around them.

5.1.2 Final list of clusters for investigation in expanded contexts.

cells

1.	**wild type cells**	**267**	
2.	**embryonic stem cells**		**228**
3.	**cos 7 cells**		**199**
4.	**bone marrow cells**		**155**
5.	**stem es cells**		**145**
6.	**embryonic stem es cells**		**142**
7.	**cd8 t cells**		**109**
8.	**cos 1 cells**		**100**

gene

9.	**gene expression data**	**218**

10.	**gene expression patterns**	**191**
11.	**mutations in the gene encoding**	**175**
12.	**gene expression profiles**	**158**
13.	**tumor suppressor gene**	**154**
14.	**changes in gene expression**	**122**
15.	**analysis of gene expression**	**97**
16.	**variation in gene expression**	**93**
17.	**gene expression profiling**	**89**

genes

18.	**tumor suppressor genes**	**138**
19.	**X linked genes**	**131**
20.	**protein coding genes**	**111**
21.	**differentially expressed genes**	**109**

cell

22.	**cancer cell lines**	**207**
23.	**lymphoblastoid cell lines**	**205**
24.	**mol cell biol**	**203**
25.	**es cell lines**	**142**

26.	**cell cycle arrest**	**137**
27.	**es cell clones**	**128**
28.	**cell cycle progression**	**112**
29.	**whole cell extracts**	**105**
30.	**cancer cell line**	**89**
31.	**planar cell polarity**	**82**
32.	**breast cancer cell**	**81**

DNA

33.	**DNA binding domain**	**155**
34.	**DNA copy number**	**150**

protein

35.	**green fluorescent protein**	**198**
36.	**protein protein interactions**	**151**
37.	**protein blot analysis**	**146**
38.	**fluorescent protein GFP**	**125**
39.	**green fluorescent protein GFP**	**124**
40.	**protein protein interaction**	**99**
41.	**wild type protein**	**83**

mutations

42.	**loss of function mutations**	**219**
43.	**disease causing mutations**	**92**

genome

44.	**genome wide association**	**824**
45.	**wide association study**	**243**
46.	**wide association studies**	**237**
47.	**genome wide linkage**	**153**
48.	**genome wide significance**	**152**
49.	**national human genome**	**139**
50.	**human genome project**	**135**
51.	**human genome research**	**135**
52.	**human genome sequence**	**133**
53.	**genome research institute**	**127**
54.	**national human genome research**	**125**
55.	**human genome research institute**	**122**
56.	**the human genome project**	**100**

analysis

57.	**western blot analysis**	**607**
58.	**northern blot analysis**	**599**
59.	**southern blot analysis**	**558**
60.	**RT PCR analysis**	**260**
61.	**RTA PCR analysis**	**185**
62.	**blot analysis using**	**135**
63.	**RNA blot analysis**	**111**

Figure 5:2: Final list of clusters containing at least three lexical elements surrounding the ten highest keywords in *genecorp*

As can be seen, the final list numbers 63 discrete clusters. Whilst a few of these may subsequently prove to form part of larger clusters (as in the case of *genome wide association* and *wide association study*) no assumption shall be made *a priori* that this is the case. Rather, each cluster will be considered separately. In addition to this, and as anticipated in the methodology section above, there are also a number of clusters (such as *cancer cell line* and *cancer cell lines* or *tumor suppressor gene* and *tumor suppressor genes*) that are identical apart from an element that would be the same if one element of that cluster was taken as a lemma. However, there may, for example, be significant epistemological differences between the marking surrounding a single *tumor suppressor gene* rather than a number of *tumor suppressor genes* and given that this analyst is not assumed to have any privileged understanding of this data the two will be treated

separately for the purpose of further investigation. Finally, in what follows there will not be a detailed report on each of the 63 clusters, for which there would of course not be room in a PhD thesis, since there would be space for just a few hundred words on each phrase. Where there is little in the way of patterning around the node phrase I will not waste space in the thesis saying this. However, where the investigation of these clusters has revealed semantic sequences or, more importantly, patterns that reveal something about epistemic signalling in the corpus, this will be reported upon. Special attention will of course be given to the latter since the study of this is the very purpose of this thesis, and where the patterns revealed constitute a range of epistemic possibilities these will be reported in great detail.

5.2 Clusters containing cells

Overall these clusters tended to report the application of methodology rather than construct new knowledge claims relating to *cells*. As such they were often not found to be of great interest in epistemic terms since they were often not presenting new claims but describing processes. The clusters containing *cells* will not be discussed in detail though a number of them will be discussed in order to illustrate the typical patterns found.

5.2.1 *wild type cells*

The string *wild type cells* was found to occur in 258 concordance lines according to *WordSmith Tools*. The twenty most frequent collocates of *wild type cells* contained few examples of lexical collocates, as illustrated by the following table:

word	number	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
in	143	6	7	16	8	86	0	7	5	4	2	2
of	72	8	14	6	3	23	0	1	3	2	6	6
and	69	5	6	5	2	14	0	12	1	9	8	7
fig	58	0	0	2	1	0	0	42	5	1	0	7
the	55	10	9	3	2	4	0	2	7	3	11	4
with	45	0	2	2	1	15	0	5	8	6	3	3
to	44	3	5	1	2	14	0	4	5	5	1	4
cells	39	6	9	12	3	0	0	0	2	4	0	3
than	36	1	5	6	13	10	0	1	0	0	0	0
that	28	4	0	7	6	0	0	1	3	1	4	2
were	26	4	5	6	0	0	0	4	2	1	1	3
a	22	5	1	0	3	2	0	1	7	1	0	2
compared	20	0	0	2	16	0	0	1	1	0	0	0
not	20	3	0	4	2	1	0	0	8	1	1	0
from	19	1	2	4	2	9	0	0	1	0	0	0
mutant	16	1	4	1	6	0	0	0	1	0	3	0
but	15	0	0	2	2	0	0	8	0	2	0	1
by	13	1	0	3	1	1	0	2	2	1	1	1
was	13	4	2	1	0	0	0	2	2	0	1	1
0	12	1	0	4	0	0	0	0	5	0	0	2

Figure 5:3: Twenty most frequent collocates of *wild type cells* in *genecorp*

The salient epistemic process identified was a process of comparisons between *wild type cells* and other objects of study. The lexical collocate *compared* reflects this process and inspection of concordance lines including both *compared* and *wild type genes* reveals a process of comparing a phenomenon in specific, named cells with that in *wild type cells*, as in the following examples:

5:1. Trimethylated H3-Lys27 **was not more abundant** at telomeric heterochromatin in SUV39DN cells *compared* to wild-type cells (Fig. 3b) (gar04_l)

5:2. Plk4+/- embryonic fibroblasts had **increased centrosomal amplification, multipolar spindle formation and aneuploidy** *compared* with wild-type cells. (ros05_l)

5:3. Wild-type IFNgammaR1 chains probably account for 10-20% of surface IFNgammaR1 monomers in 818del4/wt cells (given **the fivefold global increase of surface receptors** in 818del4/wt cells *compared* with wild-type cells) (jou99_a)

5:4. Figure 2: Chromosomes that disjoin properly in mad2Delta cells **have more crossovers near the centromere** as *compared* with wild-type cells. (lac07_l)

Examples of comparison as an epistemic process in concordance lines containing *wild type cells* and compared in *genecorp*

The linguistic characterisation of the comparison is of particular interest. Inspection of these examples shows a range of these including *increased*, *more*, *was not more abundant* and *fivefold global increase*. It is interesting to note that these are being expressed in a qualitative way with the difference being described textually rather than being quantified in any precise way. It seems reasonable to suggest that such phraseology may occur where there is a rhetorical need for the writer to persuade the audience that the difference is significant. This epistemic process of comparison is also realised by the left side collocate *than*, as in the following examples:

5:5. In fact, **the Ca2+ response to thrombin was significantly greater** in cells that lacked PC1 *than* in *wild-type cells* (nau03_l)

5:6. Expression of luciferase **was consistently lower** in Hmg1 âˆ’/âˆ’ cells *than* in *wild-type cells* (Fig. 3a). (cal99_l)

5:7. In response to PDGF-BB, Sgpl1-/- cells migrate **less far** *than* the *wild-type cells*, whereas BC055757-/- cell lines move farther. (sch07_a)

Examples of epistemically significant comparisons in concordance lines containing *wild type cells* and the collocate *than* in *genecorp*

In the examples shown above there is a claim being made in each clause. However, in many cases these function as justificatory claims supporting a main finding that appears

in a related clause later in the text; whilst inspection of many examples of *wild type cells* and comparisons of the types of claims that are made around them reveals an epistemic practice, a more text based approach would be needed to fully understand the justificatory function of these claims. Example 5:8 below makes this point clearer: here we can see a number of comparative claims made consecutively which then build to (and presumably support an overall finding in the next sentence; that *our observations show that X*:

5:8. We **verified** four predictions of the hypothesis that the spindle checkpoint ensures the proper segregation of chromosomes whose crossovers are far from the centromere: first, there were more crossovers near the centromere of a short chromosome than a long one; second, in cells that successfully segregated all their chromosomes, there were more crossovers near the centromere of a long chromosome in mad2Delta than in *wild-type cells*; third, an artificial tether near the centromere made a long chromosome nondisjoin less often in mad2Delta cells; and fourth, the presence of a tether improved the segregation of chromosomes that had not recombined. ***Our observations show that*** the spindle checkpoint plays a crucial part in rescuing chromosomes that have initially mono-oriented.

5.2.2 *embryonic stem cells*

Concordancing of the cluster *embryonic stem cells* returned 225 examples, with the most frequent lexical collocates often appearing to involve methodological processes around the node phrase. The twenty most frequent collocates of *embryonic stem cells* were as follows:

word	no.	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
in	119	5	9	14	43	29	0	2	7	6	2	2
of	67	5	8	11	9	14	0	1	4	1	6	8
and	67	4	5	5	4	2	0	27	1	6	4	9
the	48	9	6	4	1	5	0	1	5	7	6	4
mouse	46	1	2	0	1	36	0	0	3	1	1	1
to	36	2	3	3	3	0	0	8	2	4	6	5
by	27	3	8	3	0	0	0	6	2	1	1	3
into	24	0	0	2	5	5	0	4	2	5	0	1
from	21	0	0	4	2	4	0	2	5	1	2	1
a	19	4	0	0	0	0	0	1	5	2	4	3
targeting	18	3	5	3	3	1	0	0	0	2	1	0
human	18	0	0	0	0	17	0	0	1	0	0	0
were	18	4	0	0	0	0	0	9	1	2	1	1
recombination	16	0	2	5	5	0	0	0	0	3	0	1
using	15	2	1	0	2	3	0	2	0	1	2	2
homologous	15	2	3	5	0	0	0	0	3	1	1	0
with	14	2	0	1	1	1	0	4	1	4	0	0
derived	14	0	3	0	0	3	0	2	2	4	0	0
for	13	2	1	1	1	2	0	0	1	1	4	0
we	13	0	4	3	3	0	0	2	0	0	1	0

Figure 5:4: Twenty most frequent collocates of *embryonic stem cells* in *genecorp*

One immediately recognisable and somewhat trivial collocational pattern surrounding *embryonic stem cells* involved the classification of *embryonic stem cells* according to their biological origin such as *mouse embryonic stem cells* (46 occurrences) and *human embryonic stem cells* (18). Of the other lexical collocates *recombination* and *homologous* were found to form a method applied to *embryonic stem cells*, in the following way:

5:9. We generated Cdc25b-deficient mice by **homologous recombination** in embryonic stem cells (Fig. 1a,b)13. (lin02_l)

5:10. was introduced by **homologous recombination** (dotted line) into embryonic stem cells (wan05_l)

A number of other named techniques besides *homologous recombination* were found to be present, including *thymadine assay*, *targeted mutagenesis* and *secretory-trap insertions*. Overall the concordance lines containing *embryonic stem cells* were not found to be of great epistemic interest since they generally contained statements recounting methodology.

5.3 Clusters containing *gene*

As indicated by the clusters search already discussed in 5.1 above, all clusters from *genecorp* containing the keyword *expression* also contain the keyword *gene*. As such these must of course be dealt with together and in what follows all clusters containing these two keywords will be discussed. Overall *gene* + *expression* is found to occur mainly in methodological contexts where there is little or no explicit epistemic signalling of note since what is being done is regarded within the discipline as relatively unproblematic. However this material is still of significance in a study of the discourse surrounding geneticists' publication of their findings since it reveals common objects of study and sites of analysis within the discipline.

5.3.1 *gene expression data*

Gene expression data is the most frequent cluster containing both three lexical elements and the keyword *gene*, occurring 218 times in *genecorp* according to the clusters function of *WordSmith Tools*. It is also the most frequent cluster containing both three lexical elements and the keyword *expression*, though *WordSmith Tools* identifies only 216 instances of *gene expression data* around the keyword *expression*. When entered into the concordance of *WordSmith Tools* 216 occurrences were found. The following are identified as the twenty most frequent collocates of *gene expression data*:

word	no.	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
of	100	11	3	6	18	34	0	3	3	7	6	9
the	93	5	6	16	6	17	0	1	15	11	10	6
and	60	6	5	7	7	12	0	10	4	3	4	2
to	47	6	5	2	6	5	0	7	5	1	6	4
from	45	0	2	2	1	11	0	18	6	1	4	0
in	43	3	4	4	5	6	0	7	7	4	2	1
for	33	4	7	3	1	2	0	10	2	0	4	0
a	24	2	2	3	3	2	0	0	8	1	1	2
analysis	20	0	1	7	7	0	0	4	1	0	0	0
we	20	3	2	4	8	0	0	3	0	0	0	0
with	19	1	1	0	1	8	0	3	3	0	2	0
is	18	2	6	1	0	0	0	5	2	0	0	2
be	17	3	1	2	0	0	0	1	7	2	0	1
sets	17	0	0	1	0	0	0	16	0	0	0	0
on	15	0	2	1	0	3	0	5	1	1	1	1
microarray	14	0	0	0	3	9	0	0	0	0	0	2
are	14	1	1	0	0	0	0	12	0	0	0	0
used	13	0	3	1	2	3	0	0	0	1	1	0
large	12	1	3	1	2	3	0	0	0	1	1	0
using	12	0	0	1	3	2	0	5	0	0	1	0

Figure 5:5: Twenty most frequent collocates of *gene expression data* in *genecorp*

Analysis is perhaps the most salient of the lexical collocates of *gene expression data*, present in 20 of the 216 concordance lines and indicating that *gene expression data* is an item that typically occurs in the description of methodology, as is suggested by the presence of *we*, *used*, and *using*. Investigation of the 216 concordance lines of *gene expression data* supports the idea that *gene expression data* revealed little in terms of explicit epistemic marking. Unlike examples such as *candidate tumor suppressor gene*, no hedging of ontological status was present surrounding *gene expression data*, presumably since this refers to a concept that is regarded as unproblematic in the discipline rather than an epistemically contested object such as a gene that plays an unknown role in a studied disorder. The only modification commonly found around *gene expression data* was the use of classifiers to label a specific type of *gene expression data* (eg. *global*, *microarray*: typically found in the L1 position) or qualifiers labelling the source of the data (*from several Drosophilia species*, *of leukemic bone marrow cells*) as in the following examples:

5:11 Using **microarray gene expression data from several Drosophila species** and strains, we show that duplicated genes, compared with single-copy genes, significantly increase gene expression diversity during development. (rif04_bc)

5:12 To address the potential relevance of our findings for human disease, we analyzed **gene expression data of leukemic bone marrow cells** of 285 individuals with AML. (ste06_a)

5:13 **Global gene expression data of 85 locally advanced breast tumors** and control breast tissues3, aCGH data of 37 of the 85 tumors6 and expression data of the fibroblast serum response5 were downloaded from Stanford Microarray Database. (adl06_a)

Examples of *gene expression data* pre-modified to classify type of *gene expression data* and post-modified to indicate source of *gene expression data* in *genecorp*

5.3.2 *gene expression patterns*

Gene expression patterns was found to be also the second most frequent cluster containing both three lexical elements and the keyword *expression,* though *WordSmith Tools* identifies only 189 instances of *gene expression patterns* around the keyword *expression*. When entered into the concordance of *WordSmith Tools* 191 occurrences were found. The following are identified as the twenty most frequent collocates of *gene expression patterns*:

word	no.	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
of	95	2	3	6	7	38	0	19	1	2	5	12
the	87	9	6	7	5	19	0	1	21	11	3	5
in	81	2	5	2	5	17	0	42	4	1	2	1
and	56	0	8	5	2	11	0	12	1	2	13	2
to	43	2	4	2	11	1	0	3	11	2	3	4
that	35	8	1	3	7	3	0	6	4	1	1	1
from	20	2	1	1	3	0	0	4	1	5	3	0
a	17	5	1	0	0	0	0	1	4	3	1	2
by	17	1	3	1	3	1	0	2	2	1	1	2
cell	15	4	0	0	0	0	0	0	1	1	5	4
with	13	1	2	0	0	1	0	2	1	4	1	1
between	11	0	0	1	2	2	0	3	0	1	1	1
are	11	0	1	0	0	0	0	5	0	1	2	2
for	11	0	2	2	2	0	0	2	0	2	0	1
specific	11	1	1	4	0	4	0	1	0	0	0	0
human	11	0	0	0	0	0	0	0	6	3	1	1
be	11	0	4	0	0	0	0	0	4	1	1	1
is	10	2	3	1	0	0	0	3	0	0	1	0
analysis	10	0	0	1	9	0	0	0	0	0	0	0
we	10	1	2	2	4	0	0	1	0	0	0	0

Figure 5:6: Twenty most frequent collocates of *gene expression patterns* in *genecorp*

The most notable patterns discernible from the collocation data (figure 5:6) are *gene expression patterns in* (42) *of gene expression patterns* (38) *the gene expression patterns* (19) and *in gene expression patterns* (17).

5.3.2.1 *gene expression patterns + in*

The string *gene expression patterns in* is usually only of fairly trivial interest, simply serving to locate the site of the gene expression, either in terms of a species or part of a specimen, as in the following examples:

5:14. Figure 1. ***Gene-expression patterns in*** **human Th1 and Th2 cells**. (rog00_l)

5:15 Dynamic changes in ***gene expression patterns*** **in the forebrain** caused by ectopic overexpression of eng2a (and01_nt)

5:16. Each of these alleles permitted determination and internal comparisons of individual Pcdha ***gene expression patterns in*** **the embryonic or adult mouse**. (yin07_tr)

5:17. Because whole-mount in situ hybridization (WISH) currently is the standard technique for the visualization of ***gene-expression patterns in*** **the embryo**, our method relies on serially sectioned WISH preparations (Fig. 1a). (str00_nt)

Examples of *gene expression pattern in* + location of *gene expression pattern*

However *gene expression patterns in* is interesting in epistemic terms when such a *gene expression pattern* is explicitly linked to either a normal state of affairs or a disorder. In such cases it becomes clear that the attempt is being made to identify a causal role through *gene expression patterns*, either by explicitly linking these *patterns* to disorders or by differentiating between *gene expression patterns* in normal (healthy) specimens and those with disorders, as in the following examples:

5:18. Applications will include comparison of *gene expression patterns in* **the normal mouse** to those in **mice overexpressing selected genes** (transgenics) **or defective for selected genes** (point mutations, knockouts and so on) **as well as other animal models of human diseases**. (deb99_rev)

5:19. Similar to the analysis of cancer susceptibility genes, characterization of *gene expression patterns in* **primary cancers** has led to **the identification of genes whose expression levels are associated with specific cancer characteristics** such as metastasis potential, survival or response to therapy3, 11. (thr05_nav)

5:20. *Gene expression patterns in* **normal cells and tissues** (deb99_rev)

5:21 . Differential *gene expression patterns in* **disease** (deb99_rev)

5:22. Genome-wide view of methylation and *gene-expression patterns in* **breast cancer**. (yao05_l)

Examples of links between *gene expression patterns in* and *disorders*

In some cases the *gene expression patterns* are explicitly cited as evidence for or against a causal role for a given gene in a disorder, as in the following example:

5:23 Duodenal ***gene expression patterns in*** Hfe-deficient mice **argue against** a role for transcriptional activation of Slc11a2 (the apical iron transporter) or Slc39a1 (the basolateral iron exporter) **in the pathogenesis of iron overload** (muc03_l)

In this case there is again a link between *gene expression patterns*, in this case in *the pathogenesis of iron overload*, but here the data found is negative, and is said to *argue against a role for* the given gene.

5.3.2.2 of + gene expression patterns

Investigation of the string *of gene expression patterns* showed that this structure most often occurred as part of a noun phrase referring to a methodological procedure, as in the following examples:

5:24. We first applied hierarchical clustering **analysis** *of gene expression patterns* to assess the relative similarities among different mouse HCC models. (lee04_l2)

5:25. These relationships will be clarified by suitable **analysis** *of gene expression patterns* from intact as well as dissected tumours12, 14, 15, 41. (ros00_a)

These phrases were usually found to recount the methodology of the study rather than reporting findings or making epistemological claims and included *analysis of gene expression patterns* (8) *comparison of gene expression patterns* (6) *interpretation of gene expression patterns* (3) as well as similar L1 modifications such as *characterization*, and L2 and L1 patterns such as *models of* and *sets of*, also indicating the status of *gene expression patterns* as a body of data.

5.3.2.3 the + gene expression patterns

Though 19 concordance lines is few indeed to identify patterns, the semantic field of comparison can again be identified around *the gene expression patterns* with examples such as *to assess the similarilty of the gene expression patterns*, *the overall difference in the gene expression patterns*, *interspecies correlation in the gene expression patterns* and *compared the gene expression patterns* again indicating that the geneticists are seeking to create epistemological claims about the role of a gene through comparison of *gene expression patterns*.

5.3.2.4 in + gene expression patterns

The 17 lines of *in gene expression patterns* again showed a semantic field of comparison to the left side of the node, though in this case there were examples of change and variation. This semantic field was found to be represented by the strings *differences in gene expression patterns* (4), *variation in gene expression patterns* (4), *changes in gene expression patterns* (3) and the *lemma* SIMILARITY + *in gene expression patterns* (3)

5.3.3 *gene expression profiles*

Gene expression profiles was found to occur 158 times in *genecorp* according to the clusters function of *WordSmith Tools*. It is also the second most frequent cluster containing both three lexical elements and the keyword *expression,* as WordSmith Tools identifies 159 instances of *gene expression profiles* around the keyword *expression*. When entered into the concordance of *WordSmith Tools* 158 occurrences were also found. The following are identified as the twenty most frequent collocates of *gene expression profiles*:

word	no.	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
of	96	10	3	3	9	20	0	35	0	5	7	4
the	54	7	3	8	3	19	0	1	9	1	3	0
and	48	4	1	8	3	1	0	5	0	5	10	11
in	46	3	0	3	3	6	0	22	1	2	3	3
to	34	4	1	6	9	2	0	3	2	2	3	2
we	23	2	1	12	5	0	0	1	0	1	0	1
from	22	0	0	0	0	5	0	9	4	1	3	0
a	20	2	6	2	0	0	0	0	6	1	3	0
for	20	2	2	3	2	0	0	4	0	1	2	4
with	18	0	0	0	1	5	0	1	2	4	2	3
cells	14	0	0	0	2	0	0	0	1	6	3	2
by	13	2	1	0	4	0	0	1	2	0	0	3
using	13	2	0	0	2	1	0	3	0	2	3	0
human	12	0	1	1	0	0	0	0	5	1	1	3
compared	11	0	0	0	6	3	0	0	1	1	0	0
analysis	10	0	2	3	4	0	0	0	0	0	0	1
on	10	1	3	0	2	1	0	0	2	1	0	0
that	10	1	1	0	0	1	0	3	0	1	2	1
cell	9	2	0	0	1	0	0	0	0	2	2	2
global	9	0	0	0	0	9	0	0	0	0	0	0

Figure 5:7: Twenty most frequent collocates of *gene expression profiles* in *genecorp*

5.3.3.1 gene expression profiles + of

The string *gene expression profiles of* was found to participate in two major patterns of meaning based on the type of entity found to the right of the node phrase, as these were found to be either a location of the profiles in a particular place (eg: *gene-expression profiles of neoplastic cells*, *gene expression profiles of polyp regions*, *gene expression profiles of MCF7 cells*), or, more interestingly, of specific named diseases or syndromes. These latter represent the major epistemic findings connected with *gene expression profiles*, as illustrated in the following examples:

5:26. Molecular features of the transition from prostatic intraepithelial neoplasia (PIN) to prostate cancer: genome-wide *gene-expression profiles of* **prostate cancers** and PINs. (tom07_a)

5:27. **These findings support** the emerging notion that **the clinical outcome of individuals with cancer can be predicted** using the *gene-expression profiles of* **primary tumors at diagnosis**7, 26. (ram03_l)

5:28. These findings prompted us to examine *gene expression profiles of* **preleukemic PU.1-knockdown HSCs** in order to identify early transcriptional changes underlying the malignant transformation during the course of disease. (ste06_a)

5:29. Therefore, it is now possible to compare *gene-expression profiles* **of human and mouse cancer**. (swe05_a)

Examples of *gene expression profiles of* + aspects of disease

5.3.3.2 gene expression profiles + in

The string *gene expression profiles in* was found to be very similar to *gene expression profiles of* in terms of common patterns of meaning. Items to the right side of the node were again typically locations of the *gene expression profiles* (*gene expression profiles in the mouse and human metastasis sets*, *gene expression profiles in natural populations, gene expression profiles in cells)* or in either disorders (*gene expression profiles in prostate cancer)* or physical sites of such disorders (*gene expression profiles in tumors, gene expression profiles in tumor cells*).

5.3.3.3 of + gene expression profiles

The string *of gene expression patterns* was found to behave very similarly to *of gene expression patterns* with a semantic field of comparison and measurement present on the left side in collocates such as *comparison* (4), *analysis* (4) and *measurement* (2) in the L1 position. This is again consistent with the notion of *gene expression profiles* as a relatively unproblematic item of methodology, where geneticists simply recount that they

have carried out *analysis of gene expression profiles* or *comparison of gene expression profiles*.

5.3.3.4 the + gene expression profiles

The gene expression profiles also conformed to the patterning of the above examples, with the semantic field of comparison and analysis (*compared* (6) the lemma ANALYSE (4) and *using* (3) feature as left side collocates) and this again illustrates the use of *the gene expression profiles* within the recounting of methodology, as in the following examples:

5:30. We ***used*** the *gene expression profiles* from the mutant teratomas (kie02_a)

5:31 we ***analyzed*** *the gene-expression profiles* of 12 metastatic adenocarcinoma nodules of diverse origin (ram03_l)

5:32 we ***compared*** *the gene expression profiles* between each pair of twins (kak03_l)

Again what is present in such cases is not usually an epistemic claim but merely a recounting of the methods of the study.

5.3.4 *changes in gene expression*

Changes in gene expression occurs 122 times in *genecorp* according to the clusters function of *WordSmith Tools*. Precisely the same number of examples was also revealed by the concordancing tool of *WordSmith Tools*. The following are identified as the twenty most frequent collocates of *changes in gene expression*:

word	no.	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
in	47	1	1	0	3	2	0	18	6	7	5	4
to	45	1	6	11	9	4	0	0	2	7	3	2
the	36	3	3	3	4	9	0	0	3	5	2	4
of	36	7	3	4	4	3	0	2	1	2	5	5
that	24	0	0	4	3	4	0	10	0	0	0	3
are	21	1	1	1	0	0	0	10	3	1	4	0
and	18	2	2	1	6	0	0	3	0	0	3	1
by	13	1	0	0	3	1	0	0	4	2	1	1
a	13	1	2	1	1	0	0	0	4	0	3	1
be	11	0	3	0	0	0	0	0	4	2	1	1
with	10	2	0	1	0	0	0	1	2	1	2	1
after	9	1	1	0	0	0	0	6	0	0	1	0
specific	9	0	0	1	0	6	0	0	0	2	0	0
treatment	9	0	0	0	5	0	0	0	2	0	1	1
as	8	0	0	0	1	1	0	0	0	4	0	2
or	8	0	1	1	0	1	0	0	0	2	2	1
also	7	2	1	1	0	0	0	0	2	1	0	0
can	7	0	1	0	0	0	0	5	0	0	0	1
may	7	0	0	1	2	0	0	2	2	0	0	0
using	7	3	0	0	0	0	0	1	0	1	2	0

Figure 5:8: Twenty most frequent collocates of *changes in gene expression* in *genecorp*

Few, if any, patterns of note were found in either the collocation data or the extended contexts of changes in gene expression. Whilst there were a large number of interesting epistemological claims surrounding this node, little in the way of a formal or semantic pattern to these claims could be discerned. There were a few causal claims surrounding *changes in gene expression*, and indeed there are five instances of *caused* as a collocate of *changes in gene expression*, but these were not thought to be common enough to be worth discussing in detail.

5.4 Clusters containing *genes*

5.4.1 *X linked genes*

X-linked genes is the second most frequent cluster containing both three lexical elements and the keyword *genes*, occurring 111 times in *genecorp* according to the clusters function of *Wordsmith Tools*. When entered into the concordance of *WordSmithTools* the following are identified as the twenty most frequent collocates of *X-linked genes*.

word	no.	L5	L4	L3	L2	L1	centre	R1	R2	R3	R4	R5
of	88	3	3	11	15	45	0	0	0	4	3	4
in	53	4	3	0	2	0	0	20	4	5	13	2
the	43	6	4	8	3	8	0	0	7	3	3	1
expression	34	0	3	3	16	1	0	0	3	4	3	1
and	24	2	1	2	1	1	0	2	1	4	7	3
to	21	2	2	1	0	0	0	4	5	1	5	1
expressed	17	0	1	1	0	4	0	3	4	4	0	0
a	17	0	4	3	0	0	0	1	4	1	2	2
are	15	0	0	1	0	1	0	9	1	0	3	0
copy	13	0	1	1	0	10	0	0	0	1	0	0
single	13	1	1	0	10	0	0	0	1	0	0	0
we	12	5	0	3	2	0	0	1	1	0	0	0
with	11	0	2	0	0	0	0	4	0	1	2	2
is	11	1	0	1	0	1	0	6	0	1	1	0
that	11	1	2	0	2	1	0	2	2	0	1	0
at	11	0	0	2	0	0	0	2	2	2	2	1
for	11	1	2	0	1	3	0	1	1	1	1	0
brain	10	0	0	0	0	0	0	0	2	2	1	5
have	8	0	0	0	0	0	0	3	2	1	1	1
multicopy	8	0	0	0	0	6	0	0	1	1	0	0

Figure 5:9: Twenty most frequent collocates of *X-linked genes* in *genecorp*

As figure 5:15 illustrates *X-linked genes* has a number of frequent lexical collocates, most notably *expression* (34) and *expressed* (17) making 51 instances of the lemma EXPRESS in just over twice that number of concordance lines.

When entered into *WordSmith Tools* as a node, 112 instances were found. Epistemologically *X-linked genes* is very interesting in that it behaves very differently from the examples discussed previously. Whilst there are some examples of a range of lexical verbs expressing an epistemic gradation of commitment surrounding the role of *X-linked genes* (such as *show*, *associated*, *suggest the involvment of*, *have been implicated* and *have been proposed as candidates for*) the most common collocation by far is with the lemma EXPRESS where epistemic statements are usually simple statements of fact. Whilst these examples do not themselves constitute unusual or unexpected textualisations of epistemology they are extremely significant theoretically in constituting examples of scientific practice where little or no epistemic marking is required. The explanation for this is presumably that sufficient consensus exists within the scientific community as to the methodology surrounding *X-linked genes* such that the status of the geneticists' findings need not be couched in any hedging language; they can simply state what they have found.

5.4.1.1 *X-linked genes* + the lemma EXPRESS

The lemma EXPRESS occurs 83 times in the 129 concordance lines in which *X-linked genes* occurs. In epistemic terms the relationship between *X-linked genes* and the lemma EXPRESS appears to be straightforward, since the phrase *X-linked genes* occurs within a noun phrase which has a subject relationship with the verb realised by the lemma EXPRESS, with the fact given being that the genes are expressed, as the following examples show:

5:33. Thus, not only is the mouse X chromosome enriched for spermatogenesis genes functioning before meiosis, but in addition, approx18% of mouse ***X-linked genes*** are expressed in postmeiotic cells. (mue08_l)

5:34. The new study by Mueller et al.2 shows that many ***X-linked genes*** are specifically expressed in spermatids. (dis08_nav)

5:35. Here, we report that 33 multicopy gene families, representing approx273 mouse ***X-linked genes***, are expressed in the testis and that this expression is predominantly in postmeiotic cells.

5:36. We then determined whether the 33 multicopy ***X-linked genes*** are expressed in germ cells or somatic testis cells. (mue08_l)

5:37. ***X-linked genes*** are highly expressed in brain tissues, consistent with a role in cognitive functions. (ngu06_a)

Examples of *X-linked genes* + the lemma EXPRESS

To the right of the lemma EXPRESS a phrase locating the expression of the genes is often found, as illustrated by the following table:

quantification	X-linked genes	are (x) expressed	location
approx18% of mouse	***X-linked genes***	are expressed	in postmeiotic cells
many	***X-linked genes***	are specifically expressed	in spermatids.
approx273 mouse	***X-linked genes***,	are expressed	in the testis
the 33 multicopy	***X-linked genes***	are expressed	in germ cells or somatic testis cells
	X-linked genes	are highly expressed	in brain tissues

Figure 5:10: A common textual pattern surrounding *X-linked genes*

As can be seen in this particular pattern the fact being stated is the proportion of *X-linked genes* that are expressed in a given location. This is instantiated linguistically by the copula form, and no hedging or other epistemic nuancing is present in these statements.

5.4.1.2 *X-linked genes* + the lemma CHROMOSOME

Unlike the examples in figure 5:17 above, when *X-linked genes* is not in the vicinity of the lemma EXPRESS there is far more epistemic marking. This can be seen in the following examples, where chromosome occurs in the same sentence as *X-linked genes*. The actual semantic contribution of chromosome provides the location of the *X-linked genes*, as the following examples illustrate:

5:38. Nevertheless, amplification of ***X-linked genes*** may have evolved to compensate for the repressive chromatin environment affecting the X chromosome in postmeiotic cells (Fig. 5c). (mue08_l)

5:39. Thus, amplification of ***X-linked genes*** may have evolved as a way to restore gene expression from the meiosis-repressed X chromosome (Fig. 1). (dis08_nav)

5:40. Expression analysis of individual ***X-linked genes*** also indicated reactivation after male meiosis, and cytological studies showed that RNA polymerase II was

associated with the sex chromosomes after meiosis1, 9. Mueller et al.2 set out to examine the expression of multicopy genes. (dis08_nav)

5:41. In these latter conditions, genetic data suggest the involvement of at least 12 ***X-linked genes*** distributed along the X chromosome4, 5. (car99_a2)

5:42. Our finding of 10 ***X-linked genes*** is highly unlikely to have occurred by chance (P<10-8), and it indicates a roughly 15-fold enrichment on the X chromosome for male germ-cellâ^'specific, spermatogonially expressed genes. (wan01_l)

There are many examples of epistemic marking in these examples, some of which pertain to *X-linked genes* and some of which are relevant to the collocate *chromosome(s)*. The modal structure *may have evolved* appears twice, notably once in a letter and once in a news and views in the same year, and indeed the latter is a direct reference to, and partial rewording of, the former, and it is noteworthy that the exact form of the epistemic marking remains the same in the second text, constituting an interesting example of the phraseology of epistemic intertextuality.

5.4.1.3 *X-linked genes* + the lemma REACTIVATE

In the examples of *X-linked genes* and the lemma REACTIVATE there is again a noteworthy lack of any epistemic hedging. The relationship between *X-linked genes* and the lemma REACTIVATE is broadly a process participant relationship (in Hallidayan terms) though

this can appear either with *X-linked genes* as the subject (some *X-linked genes* are *reactivated*) or within a nominalisation (*reactivation* of *X-linked genes* occurs) or again with *X-linked genes* appearing as a qualifier within a noun phrase that functions as a subject (a subset of *X-linked genes* is *reactivated*), as in the following examples:

5:43. A new comprehensive study by Jacob Mueller and colleagues2 on page 794 of this issue addresses this question by demonstrating that a subset of ***X-linked genes*** is **reactivated** after meiosis. (dis08_nav)

5:44. fter meiotic sex chromosome inactivation, the X and Y chromosomes are turned off. In X-bearing spermatids, **reactivation** of ***X-linked genes*** occurs, mostly at loci with multiple copies arranged either in palindromes (head to head arrows) or in tandem (head to tail arrows). (dis08_nav)
reactivated

5:45. Some ***X-linked genes*** are **reactivated** in spermatids30, 31, 32, 33, but the global transcriptional output from the X chromosome was unknown in these haploid cells (X:A or Y:A). (ngu06_a)

5:46. It is possible that other compensatory mechanisms, aside from increased copy number, also counteract postmeiotic repression, because rare cases of robust

reactivation of single-copy ***X-linked genes*** (for example, Uba1) have also been identified8. (mue08_l)

Examples of *X-linked genes* in relation to the lemma REACTIVATE in the expanded contexts of *X-linked genes*

In epistemic terms the *reactivation of X-linked genes* again appears to be stated as a fact. Whilst this is often accompanied by quantification indicating that this is not common, in examples such as 'rare cases of robust *reactivation*' and 'some *X-linked genes* are *reactivated*') the expression of the fact that *reactivation* takes places is simply stated as fact. This is particularly intriguing in the light of the following example:

5:47. Are ***X-linked genes*** permanently silenced, or do they reactivate? (dis08_nav)

Here the paper encoded here as mue08_l is being reported in dis08_nav. From the manner of this reporting it would appear that the *reactivation* is at issue epistemically, or at the very least that it is a suprising finding. Nonetheless the reporting of this in the article appears to involve no hedging other than the admission that it is not a common feature.

5.4.1.4 *X-linked genes* + *repression*

There are 11 examples of ***repression*** occurring within the expanded context of *X-linked genes*, four of which appear below:

5:48. Retrogenes may also function after meiosis to compensate for partial ***repression*** of single-copy ***X-linked genes***. (dis08_nav)

5:49. Our ongoing differential screens have thus far not found any evidence for locus-specific ***repression*** of maternal ***X-linked genes*** in mice. (rae05_l)

5:50. We conclude that, in contrast to the complete meiotic silencing of ***X-linked genes*** during MSCI, postmeiotic ***repression*** of the X chromosome is incomplete. (mue08_l)

5:51. We identified a cluster of ***X-linked genes*** containing at least three genes that show transcriptional **repression** of paternal alleles. (rae05_l)

Examples of *X linked genes* and *repression* in the expanded contexts of *X-linked genes*

Epistemically speaking, the relationship between *repression* and *X-linked genes* appears to be an interesting one, with researchers stating that they have 'not found any *evidence* for locus-specific *repression* of maternal *X-linked genes*'. The word *evidence* may well constitute an important focal point for epistemic marking in *genecorp* and the frequency

of *evidence* in *genecorp* (4,312) appears to suggest that it is a very significant lexical marker of epistemology .

5.4.1.5 *X-linked genes* + the lemma INACTIVE

The lemma INACTIVE occurs 15 times in the expanded context of *X-linked genes*, encompassing the forms *inactive*, *inactivated* and *inactivation*. Whilst there are too few examples to discern any particular patterns it is worth noting that the *inactivation* refers to the X chromosome genes, as textualised by the forms *X inactivation* and *inactivated genes on the X chromosome*. In some examples a causal role is given to this *X inactivation*, such as in the examples '*X inactivation* equalizes the expression of *X-linked genes*' and '*X inactivation* provides a mechanism to protect the organism'.

5:52. X *inactivation* equalizes the expression of *X-linked genes* between XY males and XX females through transcriptional silencing of one of the two female X chromosomes in a random manner2, 3, 4. (wut02_a)

5:53. In all eutherian mammalian species (except X-monosomic mutants40), X chromosome *inactivation* (XCI) is used to achieve an equality of expression of *X-linked genes* between males and females41 (Fig. 3). (yan07_per)

5:54. Although chromosome-wide changes in histone modification are observed on the X chromosome in Eed mutants, more ***X-linked genes*** should be tested to rigorously prove that Eed is required to maintain all *inactivated* genes on the X chromosome. (fer03_nav)

5:55. Fourth, X *inactivation* also seems to be normal in the ICM, although aberrant expression of *X-linked genes* has been observed in the placenta. (yan07_per)

5:56. X *inactivation* provides a mechanism to protect the organism from functional 'tetrasomy' of upregulated *X-linked genes*. (ngu06_a)

Examples of *X-linked genes* and the lemma INACTIVATE in the expanded contexts of *X-linked genes*

5.4.2 *protein coding genes*

Protein coding genes is the third most frequent cluster containing both three lexical elements and the keyword *genes*, occurring 112 times in *genecorp* according to the clusters function of *Wordsmith Tools*. When entered into *Wordsmith Tools* as a node 112 instances were found. In epistemological terms *protein coding genes* has not proven to be particularly interesting. Most of the statements made in the immediately context are what Latour and Woolgar call ‘type four statements’: mere statements of fact that contain little

or no epistemological marking, either implicit or explicit. The only real exception to this is the lemma *know*, appearing 12 times in total, ten times as the L1 collocate *known*.

5.5 Clusters containing *expression*

Each of the clusters containing *expression* also contains *gene* and has thus been dealt with already in section 5.4.

5.6 Clusters containing *cell*

Overall the phrases surrounding the keyword *cell* were found to occur in the contexts of statements about methodological processes. Patterning around such phrases thus reveals the typical lexicon of processes and discourse objects from within the methodology of genetics. However, it is rare that any epistemic claims surround these phrases, which is perhaps unsurprising since such claims would be expected to occur within results or discussion sections. The strings *cancer cell lines* and *lymphoblastoid cell lines* will be discussed in some detail to demonstrate the typical epistemic contexts of phrases found around the keyword *cell*.

5.6.1 cancer cell lines

The string *cancer cell lines* occurs 207 times according to the clusters function of *WordSmith Tools*. When entered into the concordancer as a search term 210 instances were found. The twenty most frequent collocates of *cancer cell lines* are as follows:

word	no.	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
and	120	12	11	12	19	6	0	28	11	7	8	6
in	108	12	12	21	32	18	0	3	5	3	2	0
of	80	10	10	18	12	2	0	3	4	8	7	6
breast	51	0	0	1	0	47	0	0	2	0	0	1
human	43	0	4	1	19	16	0	0	0	2	1	0
the	43	3	6	6	9	1	0	2	6	2	5	3
ovarian	31	4	2	2	0	16	0	1	0	2	3	1
colon	27	0	1	2	0	23	0	0	0	0	1	0
a	26	2	6	2	0	0	0	2	4	4	4	2
with	25	0	0	0	1	0	0	12	6	3	2	1
that	19	2	1	1	1	0	0	6	3	1	2	2
normal	18	2	3	0	3	0	0	0	5	2	3	0
from	18	3	1	0	6	2	0	2	0	2	1	1
colorectal	16	1	0	2	0	12	0	1	0	0	0	0
for	16	3	0	2	2	0	0	3	2	2	1	1
expression	15	4	4	1	1	0	0	0	2	1	1	1
we	15	5	5	0	1	0	0	2	2	0	0	0
to	14	0	2	0	1	0	0	2	5	1	3	0
prostate	13	1	0	0	1	10	0	0	0	0	0	1
tumors	12	1	2	3	1	0	0	0	2	3	0	0

Figure 5:11:Twenty most frequent collocates of *cancer cell lines* in *genecorp*

Perhaps the most obvious lexical collocates are those that typically occur to the left of the node indicating position in the body (*breast*, *ovarian*, *colon* etc) and these will be discussed further below. *Normal* is also a noteworthy collocate, and occurs frequently in the 5:5 span of *cancer cell lines* since direct comparison of normal and cancerous *cell lines* is a common pattern of meaning surrounding this node, as will be discussed below. The most common patterns surrounding cancer cell lines will now be discussed in some detail.

5.6.1.1 breast + cancer cell lines

The most salient lexical collocate of *cancer cell lines* was *breast* occurring 51 times in total, with 47 of these occurrences being in the L1 position forming the string *breast cancer cell lines*. This collocate clearly falls into a semantic field locating the *cancer cell lines* in the body with the collocates *ovarian* (31) *colon* (27) *colorectal* (16) and *prostate* (13) all appearing in the list of twenty most frequent collocates and each usually appearing in the L1 position as a classifier forming part of a noun phrase with *cancer cell lines*. In the wider context of this phrase *breast cancer cell lines* was found to form part of a larger noun phrase that constituted the object of study within the methodology section, as in the following examples:

5:57. we first analysed a panel of *breast cancer cell lines* (jac00_a)

5:58. We examined a panel of tumors and *breast cancer cell lines* (yan02_bc2)

As such this phrase is often used in an epistemologically neutral setting where no claims are (as yet) being made about the *breast cancer cell lines*. Where this is the case there is often comparison taking place between normal and cancerous cell lines, and the claims involve statements of *amplification* or *expression* of a protein in the *cancer cell lines*, as in the following examples:

5:59. Some researchers have found that **PTP1B expression is higher** in human *breast cancer cell lines* than in a normal breast cell line20. (jul07_a)

5:60. Southern-blot analysis of genomic DNA after digestion with PvuII showed that **PPM1D was amplified** in the *breast-cancer cell lines* with elevated PPM1D mRNA levels (bul02_l)

5:61. we detected **overexpression of the PPM1D protein** in breast cancer cell lines harboring genomic **amplification** of PPM1D (Fig. 1e). (yan02_bc2)

Examples of claims concerning *amplification* and *expression* of a named protein in *breast cancer cell lines*

What is interesting epistemologically in such cases is that what is being textualised is the result of genetic differences in *breast cancer cell lines* as opposed to *non-cancerous cell lines*, with an underlying genetic cause for the cancer being sought. However the semantics of causation are not really present explicitly (albeit that *expression* and *amplification* are the names of certain genetic processes). Rather, all that is being claimed at this stage is that differences between cancer and non-cancerous *cell lines* have been identified.

5.6.1.2 cancer cell lines + and

Overall there were few patterns in the 28 examples of *cancer cell lines and*, with the lemma TUMOR (8) the most frequent collocate, appearing on the right side and revealing a relationship of comparison between the two objects, as exemplified in the following:

5:62. The finding of mutations in DNA samples derived from both *cancer cell lines* and tumors **implicates** ST7 as a TSG. (zen01_a)

5:63. We selected a panel of 29 tumor suppressor and candidate tumor suppressor genes that **are known to be** frequently hypermethylated in various *cancer cell lines* and primary tumor samples (right) from a review of the literature (ohm07_l)

5:64. Increased 20q13.2 copy number is observed in approximately 18% of primary breast tumors and 40% of breast *cancer cell lines* **and is associated with** aggressive

tumor behavior, poor prognosis, cellular immortalization and genomic instability. (ewa03_a)

Examples of *cancer cell lines and* + the lemma TUMOUR

In each case *cancer cell lines* and an aspect of the lemma TUMOUR are being discussed in the same way either by constituting the site of samples (examples 1 and 2) or as the location of a copy number that is associated with given physical disorders (*poor prognosis*, *cellular immortalization* and *genomic instability*). Explicit epistemic marking around these strings is marked in bold in examples above and of particular interest are the phrases *implicates* and *is associated with* which can be seen as examples of verb choice functioning as a lexical hedge around each of the respective claims.

5.6.1.3 in + X + cancer cell lines

The string *in + X + cancer lines* was found to collocate mainly with nouns with a classifying function naming the location of the cancer cell lines including *human* (3) *breast* (8) *colon* (4) *colorectal* (1) *prostate* (2) and *ovarian* (2). There were also some examples with quantifying functions in the X position at L1 from the node *cancer cell lines* including *various* (3) *all* (1) *both* (1) *30* (1) *and 60* (1) *several* (1) *many* (1). No common epistemic patterns were found within the extended contexts of *in + X + cancer cell lines*.

5.6.1.4 in + X + X + cancer cell lines

This string was found to collocate again with lexical items indicating number and location of *cancer cell lines* such as *human prostate cancer cell lines* (2) *human breast cancer cell lines* (2) *31 breast cancer cell lines* (2). There were again found to be no common patterns of wider epistemic marking around these strings.

5.6.2 colon cancer cell lines

Colon cancer cell lines has already been identified as being one example of a larger semantic field of position in the body identified by nouns classifying *cancer cell lines* on the left side of the node, usually in the L1 position. Other than the string *human colon cancer cell lines* (8) this phrase was not found to participate in any common wider patterns.

5.7 Clusters containing *DNA*

5.7.1 *DNA binding domain*

The most common cluster containing *DNA* was *DNA binding domain*, and the following table illustrates the twenty most frequent collocates of this phrase:

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
the	155	10	7	7	35	55	0	2	14	10	8	7
of	75	1	1	12	16	0	0	36	2	1	3	3
and	46	7	5	0	4	0	0	12	2	5	4	7
a	45	5	4	6	10	7	0	0	7	4	1	1
to	42	4	3	11	5	0	0	0	7	2	7	3
GAL4	42	0	1	0	1	30	0	0	4	3	3	0
in	32	3	2	3	12	0	0	5	4	1	1	1
with	22	4	0	4	3	0	0	1	4	3	1	2
domain	20	3	2	2	2	0	0	0	1	4	4	2
fused	20	2	6	2	1	0	0	7	0	1	0	1
that	11	2	3	1	1	0	0	1	0	2	1	0
amino	10	2	0	0	1	0	0	3	1	3	0	0
protein	10	3	2	0	1	0	0	0	0	1	1	2
fig	10	1	1	0	0	0	0	2	1	3	2	0
conserved	10	0	1	0	1	6	0	0	0	0	1	1
or	9	0	0	0	2	0	0	2	1	3	0	1
is	9	0	0	1	2	0	0	0	2	1	3	0
acids	9	1	2	0	0	0	0	0	3	0	3	0
dbd	9	0	0	2	0	2	0	3	2	0	0	0
as	8	1	1	1	0	0	0	1	1	1	1	1

Figure 5:12: Twenty most frequent collocates of *DNA binding domain* in genecorp

5.7.1.1 the + DNA binding domain

The L1 collocate *the* forms the most common phrasal pattern containing the node *DNA binding domain*, occurring 55 times in *genecorp*. The most common pattern around this node phrase is *the DNA binding domain* + *of* + protein which occurs in 29 of the 55 instances of *the DNA binding domain*, as illustrated in these examples:

5:65. analysis of the co-crystal structure of ***the DNA-binding domain of*** **p53** (ber06_a)

5:66. a cDNA fragment encoding ***the DNA-binding domain of Oct-3/4***, (niw00_l)

5:67. ***the DNA-binding domain of the GAL4 protein*** (say06_l)

5:68. mutation in exon 3, which encodes part of ***the DNA-binding domain of SF-1*** (ach99_cor)

5:69. ***the DNA binding domain of Irf6*** was amplified from E14 mouse cDNA (ric06_l)

Examples of *the DNA binding domain* + *of* + protein in genecorp

In each of the five examples here *the DNA binding domain* is followed by *of* and then a specific, named protein. Very little explicit epistemic marking can be found in the immediate cotext of this phrase, which may reflect its use in the recounting of methodology as in example 5:70 below:

5:70. A human fetal brain yeast two-hybrid expression library was screened with a partial NPHP6 clone (residues 1â€“684) fused with ***the DNA-binding domain of the GAL4 protein*** (say06_l)

Statements such as this merely recount the techniques used in the study and unsurprisingly display little in terms of explicit epistemic marking since the results of the study and any knowledge claims will occur in a separate section. However, such knowledge claims can be identified within concordance lines containing *the DNA binding domain*, as in the following examples:

5:71. However, our original conclusion, that the Y153H variation in *the DNA binding domain of the winged helix protein* encoded by STOX1 **is involved in** the etiology of preeclampsia (dij07_cor)

5:72. Based on analysis of the co-crystal structure of *the DNA-binding domain of p53* and the SH3 domain of ASPP2, **it was shown that** this change allows ASPP2 to have higher binding affinity to the DNA-binding domain of p53 (ber06_a)

In example 5:71 above the verb phrase *is involved in* is the linguistic expression of the current state of knowledge surrounding the given *DNA binding domain*. This choice of verb can be seen as a qualification of the claim that the named variation causes preeclampsia; rather, what it is being claimed is that it is merely *involved in* the causative process (*etiology*). In example 5:72 explicit epistemic signalling is again seen, this time by the choice of the reporting verb in the phrase *it was shown that*, which signals that what follows is to be taken as an established fact.

5.7.1.2 DNA binding domain + of

The string *DNA binding domain of* occurs 36 times in *genecorp* and is the second most common pattern containing *DNA binding domain*. As has already been seen above (5.7.1.1) where *DNA binding domain* is followed by *of* it is also usually preceded by the (since this has already been identified as happening 29 times) and is then usually followed by a specific protein name or more general reference to a protein or protein family.

5.7.1.3 GAL4 + DNA binding domain

The third most common string containing *DNA binding domain* in *genecorp* is *GAL4 DNA binding domain*, which occurs 31 times, though in one of these instances it appears

as *GAL-4 DNA binding domain*. GAL4 modifies the node phrase by specifying a particular protein as the location of the *DNA binding domain*, and indeed this is also the case in the strings *CTCF DNA binding domain* (3) *LexA DNA binding domain* (2) *ATFT DNA binding domain* (1). In the wider context of this phrase the lemma FUSE occurs as both a left side and right side collocate as the process of fusion between the *DNA binding domain* and other objects is described, with *GAL4 binding domain* in either the subject or object position within the clause, as in the following examples:

5:73. Co-transfection of a construct encoding MBD2a ***fused*** to the ***Gal4 DNA-binding domain*** together with the DNA pol beta (zha99_l2)

5:74. the ***Gal4 DNA-binding domain fused*** with the wild-type Mef2c transactivation domain (yan00_l)

5.7.1.4 of + X + DNA binding domain

The pattern *of + X + DNA binding domain* occurs 16 times in *genecorp* and is instantiated by the string *of the DNA binding domain* 14 out of these 16 times. No salient patterns were noted around this phrase

5.7.2 DNA copy number

The string DNA copy number was somewhat unusual in having a relatively high number of lexical collocates within the twenty most frequent collocates surrounding it (ten in total), as can be seen from the following table:

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
of	64	3	5	4	6	26	0	4	4	1	4	7
the	56	5	3	1	5	8	0	0	6	22	2	4
in	54	3	1	0	1	21	0	6	15	2	3	2
and	41	2	6	5	0	1	0	11	5	2	5	4
for	34	2	0	4	6	4	0	2	10	0	2	4
changes	32	0	0	0	5	0	0	27	0	0	0	0
to	31	1	5	3	5	7	0	0	1	5	0	4
variation	24	1	0	0	1	0	0	22	0	0	0	0
genome	18	0	7	0	2	1	0	1	0	0	3	4
expression	15	4	0	0	2	0	0	0	2	6	1	0
analysis	14	0	1	1	11	0	0	0	1	0	0	0
a	14	5	1	0	1	2	0	0	1	2	1	1
by	14	1	2	1	0	3	0	1	3	2	0	1
gene	14	0	0	3	0	0	0	1	6	3	0	1
genomic	11	1	1	0	0	5	0	0	0	1	1	2
that	10	2	2	2	0	1	0	2	1	0	0	0
profiles	10	0	0	0	0	1	0	9	0	0	0	0
alterations	9	0	0	1	6	0	0	2	0	0	0	0
from	9	1	0	0	0	1	0	0	0	1	2	4
cdna	9	2	2	0	0	0	0	0	0	4	1	0

Figure 5:13: Twenty most frequent collocates of *DNA copy number* in *genecorp*

The relationship between each of these lexical collocates and the node string was examined and the following are those examples which yielded the most relevant data from the viewpoint of examining epistemic relationships.

5.7.2.1 DNA copy number changes

One salient pattern found in the immediate cotext of *DNA copy number changes* is a semantic preference for causation, as illustrated in the examples below:

5:75. 82 regions of correlation entirely or partly **accounted for** by ***DNA copy number changes*** (str06_a)

5:76. The correlated expression in these regions **was therefore due to *DNA copy number changes***. (str06_a)

5:77. These 82 regions that were mostly not conserved after recalculation, in which the correlation **was essentially due to *DNA copy number changes*** (str06_a)

5:78. we identified regions retained on transcriptome correlation maps after recalculation in which the correlation **could not be accounted for** by ***DNA copy number changes*** (str06_a)

5:79. The identification of these regions of correlation **not due to *DNA copy number changes*** could not be accounted (str06_a)

Examples of *DNA copy number changes* with *accounted for* or *due to*.

In a third of the instances of *DNA copy number changes*, the forms *due to* or *accounted for* can be found in the immediate cotext, with negation also present in some but not all of these examples. In epistemic terms the significance of this is that the claims surrounding the node phrase involve the ascription or rejection of a causal role to *DNA copy number changes*. This causal role is in each case related to the lemma CORRELATE (underlined in the examples 5:75-5:79 above). Thus in examples 5:75, 5:76 and 5:77 above *correlation* or a *correlated expression* is claimed to be *accounted for* or *due to DNA copy number changes*. In examples 5:78 and 5:79 the same epistemic process is occurring, but in these cases negation is present since the claim is that the *correlation* is not caused by the *DNA copy number changes*. As can also be seen from figure 5:12 this is also an example of a 'bursty' phenomenon since this pattern of *accounted for* or *due to + DNA copy number changes* is found in only one text, that coded as str06_a.

5.7.2.2 of +DNA copy number

The most common items to the right of the string *of DNA copy number* were *variation* (9) and *changes* (5). To the left of this node phrase were a number of nominalized processes forming longer phrases generally referring back to techniques undertaken in the reported studies such as *analysis of DNA copy number* (11) and *measurement of DNA copy number* (5).

5.7.2.3 DNA copy number + variation

Close inspection of the concordance lines for *DNA copy number variation* revealed that nine of the 22 instances of this phrase came from references sections and are thus not discussed here. As is discussed above (4.4) a small number (approximately 10%) of the references remained in the corpus after all extraction techniques had been attempted; in order to investigate the corpus in a consistent way these examples are therefore not discussed here, but rather they are discarded manually in the way described in section 4.4. Amongst the remaining instances, words or lematized words pertaining to experimental procedures were again present, including the lemma MEASURE and the phrases *analysis of* and *assessment of*. The lemma MAP also occurs twice to the left of the phrase *DNA copy number variation*, instantiating a further use of the map metaphor in genetics as the writers state that they 'would like not only to **map *DNA copy-number variation*** at high resolution, but also to measure changes in DNA copy number gene by gene, for every human gene' and then illustrate an example of this process in the same paper with this described as 'Genome-wide **mapping** of ***DNA copy-number variation*** for breast cancer cell line BT474' (pol99_l).

5.7.2.4 in + DNA copy number

The phrase *in DNA copy number* occurs 21 times in *genecorp* and in all 21 occurrences the semantic field of *difference* can be identified to the left of the phrase. This is instantiated in each case by the word in the L2 position relative to the node *DNA copy number*, giving strings such as *changes in DNA copy number*, *variation in DNA copy number* and *alterations in DNA copy number*, as illustrated below:

5:80. array CGH has been used to localize changes ***in DNA copy number*** that underlie the progression of mouse islet carcinoma (pol02_rev)

5:81. Identification of chromosomal imbalances and variation ***in DNA copy-number*** is essential to our understanding of disease mechanisms and pathogenesis. (ish04_tr)

5:82. large segmental differences ***in DNA copy number*** are often the product of recurrent mutation (ega07_l)

5:83. We observed no substantial alterations ***in DNA copy number*** in six independent nodules (gup05_a)

5:84. Alteration ***in DNA copy number*** is one of the many ways in which gene expression and function may be modified. (pin05_per)

Examples of the semantic field of *difference* to the left of the phrase *in DNA copy number* in *genecorp*

5.8 Clusters containing *protein*

5.8.1 *green fluorescent protein*

A number of lexical collocates appear within the twenty most frequent collocates of *green fluorescent protein* including *enhanced* (34 occurrences) *tagged* (23 occurrences) and the lemma EXPRESS (40 occurrences). Notably *enhanced* only appears in the L1 position as an adjective modifier of *green fluorescent protein*, whilst *tagged* and the lemma EXPRESS show more widely dispersed collocation profiles with *tagged* in particular occurring on the left and the right side of the node in almost equal numbers.

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
GFP	134	0	1	2	0	0	0	123	4	1	2	1
the	73	8	5	6	2	15	0	0	3	14	12	8
a	67	6	12	6	4	22	0	1	2	5	2	7
of	66	6	12	7	14	15	0	0	1	3	3	5
and	52	4	4	3	4	7	0	3	14	5	5	3
with	46	9	3	5	6	16	0	0	1	2	3	1
in	40	6	0	2	2	0	0	3	8	4	8	7
to	38	1	6	1	5	7	0	0	3	6	5	4
enhanced	34	0	0	0	0	34	0	0	0	0	0	0
EGFP	33	0	0	2	3	0	0	23	3	1	0	1
by	29	2	5	5	3	4	0	1	1	3	4	1
or	23	1	1	0	1	3	0	4	7	2	2	2
tagged	23	0	0	4	8	0	0	0	9	2	0	0
reporter	21	2	1	3	1	1	0	1	10	0	2	0
cells	20	2	3	5	3	0	0	0	1	0	6	0
expression	20	2	2	5	7	0	0	1	0	1	1	1
expressing	20	0	1	3	8	7	0	0	0	0	0	1
encoding	19	1	0	0	5	11	0	0	0	1	0	1
we	19	7	4	3	2	0	0	0	0	1	1	1
gene	18	0	1	1	5	0	0	4	4	2	0	1

Figure 5:14: Twenty most frequent collocates of *green fluorescent protein* in *genecorp*

5.8.1.2 enhanced green fluorescent protein

The L1 collocate *enhanced* is only present where *GFP* is not present in the R1 position and thus appears to represent a separate entity identified as such by the qualifying adjective *enhanced*, though it might also be the case that these are in fact two competing terms for the same entity, or even that the terms can be used interchangeably within the same text.

5.8.1.3 green fluorescent protein EGFP

EGFP is only present where *green fluorescent protein* is preceded by the L1 collocate *enhanced* and is therefore an acronym of *enhanced green fluorescent protein*, occurring in complementary distribution with *green fluorescent protein gfp*.

5.8.1.4 a + green fluorescent protein

The string *a green fluorescent protein* occurs 22 times in *genecorp*. To the left of this phrase a wide range of verb choices can be seen, often with *a green fluorescent protein* in an object relation to the verb as in the case of *constructed* (1), *inserted* (1), *introduced* (1), *placed* (1) and *using* (1). To the right of this phrase there are 16 occurrences of *GFP* in

the R1 position and in the R2 position *reporter* occurs 5 times. The lemmas FUSE and EXPRESS are again present to the right of this node.

5.8.2 *protein protein interactions*

The concordance function of *WordSmith Tools* returned only 76 instances of *protein protein interactions* despite 151 instances being found by the clusters tool, as can be seen in the following table:

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
of	32	2	1	2	1	14	0	1	2	2	5	2
and	26	6	1	2	1	1	0	12	0	0	0	3
the	23	4	4	5	0	2	0	0	4	2	1	1
in	21	1	1	0	1	9	0	5	1	1	1	1
to	20	2	6	2	4	1	0	1	0	2	0	2
with	11	0	2	0	1	2	0	3	0	0	1	2
that	10	1	1	0	0	3	0	3	0	0	1	1
for	10	1	0	0	2	4	0	0	2	1	0	0
number	9	0	0	0	9	0	0	0	0	0	0	0
be	9	1	2	4	0	0	0	0	0	2	0	0
expression	7	1	1	2	0	0	0	0	1	2	0	0
known	7	3	0	2	1	1	0	0	0	0	0	0
domains	6	2	1	1	2	0	0	0	0	0	0	0
are	6	0	0	2	1	0	0	3	0	0	0	0
a	6	2	2	0	0	0	0	0	1	1	0	0
is	6	0	1	3	1	0	0	0	1	0	0	0
which	6	0	3	1	1	0	0	0	0	1	0	0
gene	6	1	2	0	0	0	0	0	2	1	0	0
through	5	0	0	0	0	4	0	1	0	0	0	0
these	5	0	0	2	0	1	0	0	0	0	1	1

Figure 5:15: Twenty most frequent collocates of *protein protein interactions* in *genecorp*

As a result of this discrepancy *protein protein interactions* will not be investigated in any details since 76 concordance lines is below the minimum number set out in section 5.1.1 above.

5.8.3 *green fluorescent protein GFP*

In 123 of the 195 instances of *green fluorescent protein*, *GFP* appears as an R1 collocate of *green fluorescent protein*, since *GFP* is an acronym for the node phrase. The concordance lines for *green fluorescent protein GFP* unsurprisingly match the global collocation data for *green fluorescent protein* in containing patterns including the lemma EXPRESS, the lemma ENCODE and the lemma TAG.

5.9 Clusters containing *mutations*

The phrases containing *mutations* proved to be the most epistemically interesting and are discussed in detail in chapter six below.

5.10 Clusters containing *genome*

The phrases containing *genome* tended to occur within 'noise' within the text not removed during the clean-up process. Whilst a number of examples of these were found

in references not fully removed from the corpus there were also other sources of these such as acknowledgements to funding bodies and the addresses of institutions, as the following examples illustrate:

5:85. Cancer Genetics Branch, National Human ***Genome*** Research Institute, NIH, Bethesda, Maryland 20892, USA. (car02_l)

5:86. D.W.-V. was supported by a grant from the National Human ***Genome*** Research Institute (T32 HG02536). (pet07_l)

5:87. BISP and WUGSC were supported by grants from the National Human ***Genome*** Research Institute (zod08_a)

As such the clusters containing *genome* were deemed not to be worthy of in-depth investigation or discussion. Indeed, even when the collocation data appeared to suggest the presence of an epistemically significant pattern this transpired to take place within the references section rather than the main body of the text and was therefore not deemed to be a valid item of study within the present thesis. The R2 collocate *identifies* in figure 5:15 below provides a typical example of this.

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
of	325	12	15	27	22	63	0	3	138	20	13	12
a	282	18	13	12	10	168	0	0	4	38	11	8
studies	260	2	0	2	1	0	0	236	11	1	5	2
study	257	1	2	0	1	0	0	244	6	1	1	1
in	190	5	16	17	35	31	0	6	46	8	9	17
for	190	3	8	34	13	58	0	0	40	5	23	6
the	190	23	19	32	23	35	0	1	5	25	18	9
identifies	123	0	0	0	0	0	0	0	84	0	2	37
analysis	112	22	17	7	10	0	0	47	8	0	1	0
and	110	12	7	8	6	8	0	4	20	10	14	21
we	80	7	22	33	7	0	0	1	7	2	0	1
to	73	7	5	6	7	11	0	1	11	9	7	9
scan	69	1	0	0	0	0	0	67	1	0	0	0
by	55	0	3	9	4	16	0	0	12	6	2	3
SNPS	54	7	2	7	1	0	0	0	0	0	26	11
with	49	8	1	3	0	5	0	1	12	3	7	9
from	48	1	3	8	15	9	0	0	5	2	2	3
cancer	46	3	2	0	0	2	0	0	0	0	22	17
loci	45	1	2	0	1	0	0	0	0	18	4	19
type	41	2	0	0	1	0	0	0	0	20	0	18

Figure 5:16: The twenty most frequent collocates of *genome wide association* in *genecorp*

The pattern *genome wide association* + X + *identifies* seems *prima facie* to be a pattern well worthy of further study but inspection of concordance lines revealed that 76 of these come from the references not fully removed from the corpus, with only 8 coming from the main body of the text. As such the phrases containing *genome* were not investigated in detail and discussion of them will be limited to an outline of the data to be presented in the appendices below.

5.11 Clusters containing *analysis*

The main epistemic patterns found in the concordance lines and collocation data for the clusters containing *analysis* involved reporting verbs. Whilst the use of reporting verbs for epistemic signalling is already fairly well understood in linguistic studies of scientific texts it is interesting to note that these investigations appear to indicate that they may be more common around particular types of discourse item. Whilst reporting verbs can be found in the concordance lines of most of the 63 phrases studies these tended to initiate a *that*- clause in which the main claim was then made. In this sense the reporting verbs did not have a close relationship that encoded the epistemic claim but rather signalled it. What is different in some of the examples containing *analysis* is that the reporting verb forms part of the same clause as the node phrase. These phrases reveal the connecting locus between the method and the resulting claims regarding findings. An outline of the

use of reporting verbs in concordance lines and collocation data featuring *northern blot analysis* will make this point clearer.

5.11.1 northern blot analysis

Overall there were found to be 599 occurrences of *northern blot analysis* in *genecorp*, a number that is somewhat unmanageable for detailed concordance investigation. In such a case collocation data is particularly useful in identifying patterns and the collocates list for *northern blot analysis* reveals that *showed* features as one of the most common lexical collocates, with 35 occurrences:

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
of	291	12	12	6	9	1	0	208	0	14	23	14
and	133	11	10	21	2	33	0	19	1	11	15	10
a	109	3	6	5	3	36	0	1	32	12	5	6
in	96	9	9	5	5	6	0	10	1	9	28	14
RNA	87	3	2	9	6	0	0	0	20	21	19	7
the	85	13	9	9	0	2	0	0	28	5	8	11
by	84	1	2	0	2	74	0	0	0	1	3	1
expression	76	2	7	3	4	0	0	0	6	35	10	9
from	67	0	8	0	0	1	0	1	1	19	16	21
we	60	7	4	25	3	0	0	12	0	4	4	1
that	52	0	2	0	0	1	0	3	34	3	5	4
as	48	3	6	16	0	0	0	10	0	5	4	4
figure	48	3	40	5	0	0	0	0	0	0	0	0
MRNA	44	5	1	0	6	0	0	0	5	15	10	2
using	44	0	0	0	0	8	0	24	0	4	3	5
C	42	0	0	1	2	34	0	1	0	3	1	0
thumbnail	41	0	4	1	31	5	0	0	0	0	0	0
for	41	0	4	3	2	24	0	3	0	2	2	1
was	39	2	1	8	0	0	0	13	0	2	9	4
showed	35	1	0	0	0	0	0	27	1	4	1	1

Figure 5:17: Twenty most frequent collocates of *northern blot analysis* in *genecorp*

It is of particular interest that *showed* generally appears in the R1 position since this will usually instantiate a subject-predicate relationship between the node phrase and this frequent lexical collocate. Moreover further inspection of the list produces reporting verbs such as *revealed* (21), *detected* (16), *indicated* (15), *confirmed* (14), *demonstrated* (9) and *identified* (6) in addition to the 35 instances of *showed*.

5.12 Conclusion

In this chapter the final list of clusters for investigation was presented and a summary was provided of the main findings surrounding each of these clusters, ordered according to the keyword around which they were discovered. The most frequent collocates surrounding a number of these node phrases were shown and the epistemic significance of the patterns revealed by the collocation data was considered. However at this stage the phrases that were found to be of most interest have not yet been discussed since these will form the main considerations of chapters six and seven, to which I now turn.

Chapter 6: Causation in *genecorp*

6.1 Introduction

The phrases containing *mutations* were *mutations in the gene encoding*, *loss-of-function mutations* and *disease-causing mutations*. These phrases were found to be unique amongst the 63 studied in that they were typically involved in encoding causal meaning. This is most obviously evident in the string *disease-causing mutations* which of course already carries an epistemic claim and the wordform *causing* but the other two phrases were also found to be typically involved in constructing causal claims. Indeeed, *loss-of-function mutations* can also be seen as a nominalisation of a causal claim in that it apparently contains the notion of *mutations* that cause a *loss-of-function*. In what follows the findings related to each of these three phrases will be discussed in detail and supported by an indicative corpus-based search of the lemma CAUSE that apparently reveals both that causative language is very common in *genecorp* and also the close relationship between *mutations* and causation that exists in genetics as exemplified by these three phrases.

6.2 mutations in the gene encoding

Mutations in the gene encoding is the third most frequent cluster containing both three lexical elements and the keyword *gene*, occurring 175 times in *genecorp* according to the clusters function of *Wordsmith Tools*. When entered into the concordance of *WordSmithTools* 174 occurrences were found. The following are identified as the twenty most frequent collocates of *mutations in the gene encoding*:

word	no.	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
the	71	3	0	5	0	0	0	54	2	1	3	3
a	33	7	3	1	0	0	0	5	8	1	3	5
of	32	3	2	1	11	3	0	0	0	4	5	3
protein	23	0	0	0	0	0	0	0	1	7	11	4
in	23	1	1	1	1	0	0	0	2	8	4	5
and	19	2	4	2	4	3	0	0	0	1	1	2
by	19	0	3	1	1	13	0	0	0	0	0	1
that	19	2	2	0	1	13	0	0	0	0	1	0
cause	19	1	0	0	0	0	0	0	6	1	6	5
2	19	0	0	1	0	0	0	1	6	3	4	4
with	17	4	4	1	0	3	0	0	0	0	2	3
caused	14	3	0	1	10	0	0	0	0	0	0	0
to	14	2	2	1	1	7	0	0	0	1	0	0
have	13	0	1	2	1	4	0	0	1	1	2	1
is	13	0	2	9	0	0	0	0	0	1	0	1
receptor	12	0	0	0	1	0	0	0	2	3	4	2
function	11	0	0	0	0	11	0	0	0	0	0	0
we	11	0	6	2	3	0	0	0	0	0	0	0
loss	10	0	0	10	0	0	0	0	0	0	0	0
for	9	0	0	1	0	4	0	0	0	1	1	2

Figure 6:1: Top twenty most frequent collocates within a 5:5 span of the node *mutations in the gene encoding*

Figure 6:1 gives the 10 most common collocates of *mutations in the gene encoding*. Whilst the list is perhaps useful in a heuristic way, drawing attention to meanings of causation surrounding the node, with 19 instances of *cause* and 14 of *caused*, it certainly does not accurately characterise the typical frequency of relationship between either the node and the lemma CAUSE or between the node and the semantic field of causation. Detailed investigation of the expanded contexts of the node revealed that the lemma CAUSE is present in 62 of the 174 concordance lines, and whilst CAUSE does not appear within the 5:5 span of the node in 29 of these examples, the grammatical and indeed semantic relationship between the two is by no means weakened by this distance. Consideration of several typical examples of this makes the point clearer:

6:1. Mutations in the gene encoding 3bold beta-hydroxysteroid-Delta 8,Delta7-isomerase **cause** X-linked dominant Conradi-HÃ¼nermann syndrome (bra99_l)

6:2. We show that mutations in the gene encoding giant-muscle filament titin (TTN) **cause** autosomal dominant DCM linked to chromosome 2q31 (CMD1G; MIM 604145). (ger02_l)

6:3. Mutations in the gene encoding peroxisomal alpha-methylacyl-CoA racemase **cause** adult-onset sensory motor neuropathy (fer00_l)

6:4. Mutations in the gene encoding the latency-associated peptide of TGF-bold beta1 **cause** Camurati-Engelmann disease (jan00_bc)

6:5. Mutations in the gene encoding the human matrix Gla protein **cause** Keutel syndrome (mun99_l)

Examples of *mutations in the gene encoding* and *cause* in the expanded contexts of *mutations in the gene encoding*

In each of the examples *cause* occurs outside of the 5:5 span of the node, and yet the relationship between *mutations in the gene encoding* and *cause* is that of the head of a noun phrase forming the subject of a clause of which *cause* is the predicator. Indeed, this relationship forms part of a semantic sequence which can be described as follows:

mutations in the gene encoding	**protein encoded by named gene**	***cause***	**named human syndrome**
mutations in the gene encoding	3bold beta-hydroxysteroid-Delta 8,Delta7-isomerase	*cause*	X-linked dominant Conradi-HÃ¼nermann syndrome
mutations in the gene encoding	giant-muscle filament titin (TTN)	*cause*	autosomal dominant DCM linked to chromosome 2q31
mutations in the gene encoding	peroxisomal alpha-methylacyl-CoA racemase	*cause*	adult-onset sensory motor neuropathy
mutations in the gene encoding	the latency-associated peptide of TGF-bold beta1	*cause*	Camurati-Engelmann disease
mutations in the gene encoding	the human matrix Gla protein	*cause*	Keutel syndrome

Figure 6:2: Illustration of the semantic sequence *mutations in the gene encoding* + named protein + *cause* + named syndrome in the expanded contexts of *mutations in the gene encoding*

Figure 6:2 demonstrates a typical semantic sequence involving *mutations in the gene encoding*. In each case *mutations in the gene encoding* forms part of the subject of a

clause, with *mutations* being the head of the noun phrase forming that subject. This is significant because it is the *mutations* that are being designated as having the causal role in each case. The remainder of that subject is formed by the element encoded by the gene, which is always a named protein. The naming of a given protein encoded by the gene appears to function as a way of locating the specific *mutations* that are being attributed with a causal role in each case. From a methodological viewpoint what is crucial here is that the names of these proteins are usually long noun phrases separating the node phrase from the predicator. This has the effect of preventing the concordancer from identifying *cause* as a collocate in many cases, since it is outside of the 5:5 window, but the underlying pattern, of which figure 6:2 illustrates one type, is usually *mutations in the gene encoding* + named protein + predicator + named syndrome. What is particularly striking in this instance is that the epistemological information in that clause appears to be expressed by the choice of verb. What varies in the choice of verb is whether it encodes a claim about a causative role, or falls short of making such a claim. This is highly significant in terms of identifying the linguistic processes of epistemic encoding in genetics since there is little or no epistemic hedging in the environs of *mutations in the gene encoding* other than this choice of verb. As such, geneticists appear to encode knowledge surrounding this node phrase lexically, and almost exclusively through this choice (or omission) of a verb that expresses the apparent causative role of the *mutations*. The principal ways in which these claims are made are as follows.

6.2.1 the lemma CAUSE

The lemma CAUSE occurs within the expanded context of *mutations in the gene encoding* and is usually found within a very close grammatical relationship with the node phrase, of the type demonstrated in figure 6:2. In all cases *mutations in the gene encoding* is forming part of an agent that is being given a causative role in the clause, and this is manifested principally by the forms *cause* and *caused by* as in the examples 6:1-6:5 above. These examples demonstrate the way in which claims surrounding *mutations in the gene encoding* are typically expressed when the word *cause* is present in the surrounding sentence. In 34 of the 56 instances *mutations in the gene encoding* forms part of the subject related to the verb *cause*, and almost always forms an unhedged claim about a genetic feature causing a symptom or syndrome observable at a non-genetic level. This is an incredibly striking finding, particularly when one considers the fact that previous studies into the popularisation of genetics (Carver et al. 2008) have explicitly criticised media coverage of such findings for using the word *cause* in an unhedged form to express genetic knowledge, which was deemed to be an overly deterministic transformation of the original research. The present findings in the expanded contexts of *mutations in the gene encoding* would appear to indicate that unhedged causative claims using the lemma CAUSE are in fact often made by geneticists in research articles and similar reporting of original research. Indeed, of these 34 examples only two contain hedging of any sort, and are as follows:

6:6. We then evaluated whether mutations in the gene encoding the noncatalytic subunit at 1q41 **might also cause** Micro syndrome. (ali05_bc)

6:7. Mutations in the gene encoding nephrocystin-4, NPHP4, **can cause** nephronophthisis or a combination of nephronophthisis and progressive retinal degeneration known as SLSN1, 19. (art07_l)

Examples of *mutations in the gene encoding* + the lemma CAUSE where hedging is present.

In example 6:6 we can see a rare example of grammatical modality as a hedging device in the form of the verb phrase *might also cause* rather than the much more common and unhedged *cause*. Example 6:7 is also of interest in that there is again the use of a modal auxilliary which can again be seen as a form of hedging since the writer is encoding possibility within the claim rather than presenting a straightforward causal claim as is found in the other 32 examples.

6:8. Tangier disease **is caused by** mutations in the gene encoding ATP-binding cassette transporter 1 (rus99_l)

6:9. Charcot-Marie-Tooth type 4B **is caused by** mutations in the gene encoding myotubularin-related protein-2 (bol00_bc)

6:10. Here we show that CPX **is caused by** mutations in the gene encoding the recently described T-box transcription factor TBX22 (ref. 14). (bra01_l)

6:11. The disease **is caused by** mutations in the gene encoding the pyrin protein2, 3, 4. (sch01_l)

Examples of *mutations in the gene encoding* and *is caused by*.

In 15 of the examples containing *mutations in the gene encoding* the lemma CAUSE is present in the form *caused by* with *mutations in the gene encoding* again constituting the causal agent but with the phenomenon present now being encoded as the subject of the clause. Interestingly these almost all again represent unhedged claims about the causative role of a genetic agent. Indeed in only two of the 15 examples is there any form of hedging, this time expressed by the modal auxilliary *can* (example 6:12. below), indicating possibility within the claim, and by the adverb *typically* (example 6:13. below) indicating typicality within the claim and therefore constituting a relation that falls short of a causal relationship, which is always the case.

6:12. Genetic studies have demonstrated that HHT **can be caused by** loss-of-function mutations in the gene encoding activin receptor-like kinase-1 (ACVRL1; ref. 5). (urn00_l)

6:13. Hyperekplexia is a human neurological disorder characterized by an excessive startle response and **is typically caused by** missense and nonsense mutations in the gene encoding the inhibitory glycine receptor (GlyR) alpha1 subunit (GLRA1)1, 2, 3. (ree06_l)

Examples of *mutations in the gene encoding* and *is caused by* with hedging.

6.2.2 other verbs expressing causation

A number of other verbs were found that also expressed causation other than the lemma cause, and these were as follows:

(are) due to (6), RESULT + in (5), ARISE + from (3), reduce (3), RESULT + from (3), underlie (3), account for (2), alter (2), contribute to (1) lead to (2), predispose (2), affect (1), attributable to (1), disrupt (1), increase (1), produce (1)

Mutations in the gene encoding also exhibits a semantic preference for causation in forming part of the subject or object of 37 other examples of verbs expressing causation. These are again highly notable in constituting examples of unhedged claims about causation, as in the following examples:

6:14. It is of interest that HSN1 **arises** from mutations in the gene encoding a subunit of the enzyme SPT, (bej01_bc)

6:15. Mutations in the gene encoding B, a novel transporter protein, **reduce** melanin content in medaka (fuk01_l)

6:16. Mutations in the gene encoding ABCR **are responsible for** Stargardt macular dystrophy. (mol00_bc)

6:17. Taken together, our data **demonstrate** that mutations in the gene encoding RANKL **lead to** an osteoclast-poor form of osteopetrosis in humans. (sob07_bc)

6:18. Osteoclast-poor human osteopetrosis **due to** mutations in the gene encoding RANKL (sob07_bc)

Examples of other verbs expressing causation in relation to *mutations in the gene encoding*

In each of these cases *mutations in the gene encoding* forms part of a causative agent and indeed it would appear that any of the verb phrases chosen here could have been replaced with the lemma CAUSE with no apparent loss of meaning whatsoever, again indicating that these are outright causatative claims. Indeed, in only one example of this set of 37 is there any hedging whatsoever present, and this is again expressed through a modal auxilliary (in this case *may*) as follows:

6:19. Overlapping syndromes such as acrocallosal syndrome (OMIM 200990) may belong to a family of 'megalinopathies' **due to** mutations in the gene encoding megalin or an interacting gene in the same pathway. (kan07_bc)

6.2.3 verbs falling short of expressing causation

A number of verb choices were also found that were judged to fall short of constituting a causal relationship but that nonetheless construed some form of relationship between the *mutations* and a named phenomenon, as illustrated by the following list:

to be + associated with (10), *implicated in*, (*may also*) *segregate* (1), *define* (1)

A further semantic grouping that can be identified amongst the remaining examples is that of verbs expressing some sort of link or role for an agent at least partly constituted by *mutations in the gene encoding* but falling short of a causative claim. This is highly significant in terms of the encoding of knowledge since it would appear that what is being expressed is that this agent may play some form of causative role but that the exact nature of this either falls short of a fully causative role (ontological hedging) or is as yet not established (epistemological hedging). The geneticists are therefore signalling possibility around the epistemological status of their claim not through grammatical modality but through verb choice, as in the following examples:

6:20. Human mitochondrial DNA deletions **associated with** mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria (spe01_a)

6:21. In support of this hypothesis is the observation of a similar phenotype **associated with** mutations in the gene encoding minK (Kcne1; 4). (del99_l)

Examples of verbs falling short of an expression of a causative relationship in relation to *mutations in the gene encoding*

6.2.4 absence of a verb expressing epistemological status

A further common pattern surrounding the node is to express the existence of such mutations in subjects with a given syndrome or disease whilst avoiding providing a linguistic characterisation of the nature of the connection between these phenomena. The following example illustrates this strategy:

6:22. Here we report mutations in the gene encoding RANKL (receptor activator of nuclear factorâ€“KB ligand) in six individuals with autosomal recessive osteopetrosis whose bone biopsy specimens lacked osteoclasts. (sob07_bc)

In this example the verb *report* is a reporting verb, which indicates that what follows is to be taken as a fact, but gives no further information as to the relationship between the node and the syndrome. Similarly, in eight instances the verb chosen is *have*, expressing as a fact that the given subjects again have both *mutations in the gene encoding* a given protein and have a specified disease or syndrome but again providing no linguistic characterisation of the nature of the connection between the two phenomena, as in the following examples:

6:23. Mice carrying mutations in the gene encoding aggrecan **have** herniation of intervertebral discs and deformation of vertebral bodies8. (sek05_l)

6:24. Although AS patients **have** infrequent mutations in the gene encoding an E6-AP ubiquitin ligase required for long-term synaptic potentiation (LTP), most cases are attributed to de novo maternal deletions of 15q11âˆ’q13 (ref. 3). (meg01_bc)

This strategy can again be seen as hedging in the sense that the geneticists are clearly implying that there is a connection between the located mutations and the observed syndrome, but fall short of labelling this as a causative connection.

6.2.5 Summary

Verb choice is the principal means of epistemic signalling around the node *mutations in the gene encoding*. The most common choice of verb is the lemma CAUSE (56) and verbs expressing causation in ways other than by using the lemma CAUSE are also common (31). As such in approximately half of the examples of *mutations in the gene encoding* a causative role is being conferred upon an agent represented by a noun phrase the head of which is *mutations*. What is epistemically interesting is that where geneticists wish to fall short of conferring a causative role to this node they usually do so not by using hedging devices such as modal auxiliaries or modal adjectives but through choice of verb, using verbs that express a connection of some sort or even by avoiding using a verb that characterises the relationship completely. This is particularly significant because a corpus-based study of hedging in genetics would completely fail to identify such a strategy if the object of study was the (known) hedging devices themselves: rather, the analyst would need to know from the outset that verb choices such as *cause*, *are due to*, *is*

associated with and even the choice to provide no linguistic characterisation of the relationship at all are how geneticists encode the epistemic status of given *mutations in the gene encoding + X*, and this would appear to be a very implausible assumption indeed.

6.3 loss of function mutations

The string *loss of function mutations* is the most common phrase containing at least three lexical elements surrounding the keyword *mutations* in *genecorp*, occurring 219 times. Concordancing and collocation information for this phrase identified a plethora of propositions constituting knowledge claims containing this phrase and in common with *mutations in the gene encoding* above, these claims were broadly found to be unhedged and causative in nature. Indeed *loss of function mutations* itself would appear already to contain causative meaning in that *loss of function* appears to classify given *mutations* as having the effecting of removing a given function from a gene. As such this would appear to be a significant epistemic node within the discourse of genetics, occurring 219 times in the corpus and expressing a causative function for a given entity. The salient ways in which this relationship is expressed in the corpus are as follows.

6.3.1 *loss-of-function mutations* and the lemma CAUSE

The lemma CAUSE forms a similar relationship with *loss of function mutations* as that found with *mutations in the gene encoding* described above. These examples fall into two main groups: cases where the *loss-of-function mutations* are in the subject position and are found with the wordform *cause* and cases where the *loss-of-function mutations* are in the object position where the verb phrase *are caused by* is found. As such longer phrases containing the string *loss-of-function mutations* often form the subject of a clause where *cause* is the main verb, as in the following examples:

6:25. Loss-of-function mutations in Tub **cause** late-onset obesity, retinal degeneration and hearing loss in tubby mice4, 5, 6. (mak06_l)

6:26. Here we report that quivering mice **carry** loss-of-function mutations in the mouse beta-spectrin 4 gene (Spnb4) that **cause** alterations in ion channel localization in myelinated nerves (par01_l)

6:27. Loss-of-function mutations in the TGF-beta type II receptor **cause** type II Marfan syndrome (OMIM 154705)25, and the LAP domain mutation of TGF-beta1 is responsible for Camurati-Engelmann disease (OMIM 131300)26. (sek05_l)

6:28. Loss-of-function mutations in RELN (encoding reelin) or PAFAH1B1 (encoding LIS1) **cause** lissencephaly, a human neuronal migration disorder1. (ass03_l)

6:29. GDF8 loss-of-function mutations **cause** double-muscling in mice, cattle and humans6, 7, making it an obvious candidate. (clo06_l)

Loss-of-functions mutation as all or part of the subject of *cause* in *genecorp*

The concordance lines surrounding *loss-of-function mutations* again contain examples where the node phrase is embedded in a complex nominal group. These can also involve a prepositional phrase locating the *loss-of-function mutations*; in such cases *loss-of-function mutations* and *cause* are separated by too many wordforms to be identified as collocates by WordSmith Tools. Where *loss-of-function mutations* is found in such clauses there is usually an unhedged causative claim as in the examples above. There can also be a coordination of claims as in example 6:25 above where *loss-of-function mutations* in Tub **cause** late-onset obesity, retinal degeneration and hearing loss in tubby l) mice.

In the examples where *loss-of-function mutations* is found in the object position as agent of the phrase *caused by* there seemed to be a slightly different epistemic process. Whilst the claim that:

X is/are *caused by loss-of-function mutations*

was still present in each case, this claim was also found embedded in a further proposition that was the main epistemic claim in the sentence, as illustrated by examples 6:30- 6:32 below:

6:30. Next, we tested **whether** reducing the expression of the same six genes **could enhance** the 'dumpy' phenotype **caused by** loss-of-function mutations in the gene dpy-20 (ref. 27). (leh06_a)

6:31. Furthermore, the PHAII phenotype of hypertension, hyperkalemia and hypercalciuria **is the virtual mirror image of** the low blood pressure, hypokalemia and hypocalciuria that are the major features of Gitelman syndrome, which **is caused by** loss-of-function mutations in the gene encoding NCC (SLC12A3)2. (cof06_nav)

6:32. We **propose** that haploinsufficiency of ATP1A2, **caused by** loss-of-function mutations of a single allele, **leads to** FHM2 by two synergistic events: an increase in extracellular K+ owing to impaired clearance of brain K+ by neurons and glial cells, producing a wide cortical depolarization21, and a local boost in intracellular Na+, which promotes an increase in intracellular Ca2+ through the Na+/Ca2+ exchanger. (fus03_l)

Examples of *loss of function mutations* occurring with *is caused by* in *genecorp*

In each of these cases the proposition that something is *caused by loss-of-function mutations* can be seen as a supporting fact. Indeed, the fact constructed by the lemma CAUSE + *loss-of-function mutations* seems almost incidental or parenthetic in these examples. In example 6:31 the main claim (bold) is as follows:

6:31 the PHAII phenotype of hypertension, hyperkalemia and hypercalciuria **is the virtual mirror image of** the low blood pressure, hypokalemia and hypocalciuria that are the major features of Gitelman syndrome

The *that*-clause at the end of this example then presents a second fact, that:

the three features listed at the end (low blood pressure, hypokalemia and hypocalciuria) are the major features of Gitelman syndrome.

Finally the fact that a syndrome is *caused by loss-of-function mutations* appears as a subordinate *which*-clause apparently providing additional information about Gitelman syndrome:

6:32 which **is caused by** loss-of-function mutations in the gene encoding NCC (SLC12A3)2.

In such cases this final fact is clearly not being reported for the first time but is rather being used to support some further finding. This would appear to illustrate Latour and

Woolgar's (1979) type four claims, since it is apparently a claim which no longer requires any modality to be attached when it is used. Though there are several examples of such claims being made in supporting roles in conjunction with *is caused by* there is not enough evidence to suggest a relationship with this form of the lemma. Indeed, there are a number of examples where *loss-of-function mutations* collocates with *is caused by* to form the main claim in a clause, as in the following cases:

6:33. Notably, lipoid proteinosis, a rare mucocutaneous autosomal recessive disorder, **is caused by** loss-of-function mutations in ECM1 (ref. 9). (dub08_nav)

6:34. Genetic studies **have demonstrated that** HHT **can be caused by** loss-of-function mutations in the gene encoding activin receptor-like kinase-1 (ACVRL1; ref. 5). (urn00_l)

Finally there are examples where the lemma CAUSE is present as part of a nominal group, where the verb *found* is used instead of the copula. Such cases are again unhedged causal claims:

6:35. Loss-of-function mutations in ATP6B1 **are a main cause** of this syndrome3. (smi00_l2)

6:36. We **have shown** that MYO5B loss-of-function mutations **are** a major cause of MVID. (mul08_bc)

6.3.2 *loss of function mutations* and other verbs expressing causation

There were a large number of verb phrases expressing causation in the concordance lines for *loss of function mutations* and the following list illustrates these:

is completely **abrogated** by, would be primarily **affected** by, **affecting**, do not **alter** (2), can be **attributed to**, completely **block**, **compromise**, **confer**, **demonstrated**, **diminished**, **disrupts**, are **due to**, may be **due to**, **eliminate**, **encodes** (2), **encoding**, **enhanced** (2) is **enhanced by, impair** (2) would **impair**, **involve**, is **knocked down through**, **lead to** (3), can **lead to**, **leading to**, **lower**, **predispose to**, can also **predispose to**, **produce** (3) would **produce**, **recapitulated**, strongly **reduce**, **reducing**, **restored**, **result in** (8) can **result in**, **results in** (3), **results from** (5), **stop**, **trigger**, **underlie** (2)

Verb phrases carrying causal meaning in clauses or clause complexes containing *loss-of-function mutations* in *genecorp*

In addition to the forty instances of the lemma CAUSE discussed above (6.3.1) there were also 62 instances of other verb phrases carrying causative meaning. Whilst it is not an entirely straightforward matter to categorise verb phrases as being causative the examples above can be seen as conferring some causative role to *loss-of-function mutations*. The verb phrases *block*, *due to*, *eliminate*, *produce*, *reduce* and *result in/from* seem to be clearly causative, as can be seen in the following examples:

6:37. ***Loss-of-function mutations*** in the cathepsin C gene result in periodontal disease and palmoplantar keratosis (too99_l)

6:38. ***Loss-of-function mutations*** in TYROBP (DAP12) result in a presenile dementia with bone cysts (pal00_l)

However, examples such as *underlie*, *compromise* and *predispose to* are perhaps more nuanced examples and may reflect a more complex etiological description where the role of the *loss-of-function mutations* is less simple, as the following examples illustrate:

6:39. Two common ***loss-of-function mutations*** within the filaggrin gene predispose for early onset of atopic dermatitis. (san07_l)

6:40. Filaggrin ***loss-of-function mutations*** predispose to phenotypes involved in the atopic march. (san07_l)

There were also examples where clause complexes involving *loss-of-function mutations* expressed more than one fact directly related to the node phrase. In the following example the nominal group containing *loss-of-function mutations* is the causative agent of two verb phrases:

6:41. Each of these genes governs renal salt handling, and homozygous loss of function mutations in them **lower** blood pressure by **reducing** salt reabsorption; heterozygous mutations in SLC12A3 have also been shown to increase renal salt loss29. (jia08_a)

In this example there is not only a causative claim (homozygous loss of function mutations in them **lower** blood pressure) but also a further claim explaining how this is done (by **reducing** salt reabsoption). In such cases the *loss-of-function mutations* form part of the subject of two clauses; the second of these is of course outside of the 5:5 span of the node phrase. It should be noted that in such cases where there are a number of claims made in separate clauses only the verb phrase present in the first clause will appear as a collocate of the node phrase.

6.3.3 *Loss of function mutations* with verbs falling short of expressing causation

In common with the string *mutations in the gene encoding* discussed above (6.2) the main process of epistemic nuancing found around this node phrase is manifested through verb choice. Whilst the most common choices are the lemma CAUSE and then a range of other apparently causative verb phrases, a third clear choice found was the use of the verb *associated*, as in the following examples:

6:42. Loss-of-function mutations in TRPM6 **are associated with** hypomagnesemia with secondary hypocalcemia, a rare autosomal-recessive disorder8. (gud05_nav)

6:43. First, haplo-insufficiency probably cannot explain ddRTA, because heterozygous loss-of-function mutations in the longer isoform of SLC4A1 expressed in erythrocytes (which cause hereditary spherocytosis) **are usually not associated with** a defect in renal acidification. (dev03_bc)

6:44. Loss-of-function mutations in LRP5 **are associated with** osteoporosis-pseudoglioma syndrome, an autosomal recessive disorder6, and hypermorphic alleles of this gene are associated with high bone mass (HBM) phenotypes7, 8, 9. (liu05_a2)

Given that there were few examples of hedging around causal claims containing *loss-of-function mutations* this appears to be one of the principal ways that geneticists writing in *Nature Genetics* nuance claims surrounding this node. Rather than stating that such *mutations might* or *may cause* a disorder or syndrome these writers fall short of a claim of that strength through verb choice, in this case by claiming that the *loss-of-function mutations* are *associated* with these deleterious effects.

6.3.4 *loss of function mutations* and the copula

Another pattern identified around *loss-of-function mutations* was the use of the copula to make claims. These claims tended to be ontological rather than causative in nature with various *mutations* being identified as *loss-of-function mutations*, as in the following examples:

6:45. These data, taken together with the observation that overexpression of PCSK9 in mice leads to elevated LDL levels5, 6, 7, **indicate that** the Y142X and C679X mutations of PCSK9 **are** loss-of-function mutations. (coh05_l)

6:46. The best news to come from the identification of this disease gene is that the mutations **are** loss-of-function mutations, so delivery of 4'-phosphopantothenate or coenzyme A might prevent neurodegeneration. (rou01_nav)

6:47. These data **indicate that** R201Q and R166Q **are likely to be** loss-of-function mutations **that would have** deleterious effects on DNA binding by CBFA2. (son99_a)

6:48. As such, the three different protein-truncating mutations **are probably** complete loss-of-function mutations of UPF3B. (tar07_l2)

6:49. We **predict** five of these mutations **to be** complete loss-of-function mutations as a consequence of NMD degradation of their respective PTC-containing mRNAs (Fig. 2d and data not shown). (dib08_l)

Examples of *loss of function mutations* and the copula in *genecorp*

In each of these examples the claim being made is that a given set of *mutations* are *loss-of-function mutations*. Notably there is some hedging around such examples with *are likely to be* (example 6:47) and *are probably* (example 6:48). The reporting verbs found

in these concordance lines are also of interest and *indicate that* appears twice in the set of examples above. In both cases it is *data* that *indicate that* X *are loss-of-function mutations*. It is difficult to judge the strength of the claim that is being made when geneticists report that 'data *indicate X'* and it would be interesting to compare the competing claims made around data to see whether there is a nuanced range of such claims.

6.3.5 *Loss of function mutations* + named disorder without verb expressing epistemic relationship

In common with the examples above in the concordance lines surrounding *mutations in the gene encoding* (6.2) a further strategy identified was the juxtaposition of *loss-of-function mutations* with a named disorder or syndrome without any linguistic encoding of the epistemic relationship between those two entities. In such cases the verb form found is usually *identified* and as such the new knowledge being reported concerns the locating of *loss-of-function mutations* in subjects with a given disorder, as in the following examples:

> 6:50. In summary, we **identified** loss-of-function mutations in 3 of 68 multiplex and consanguineous families with JBTS, **indicating that** RPGRIP1L accounts for approx5% of JBTS cases in our cohort. (art07_l)

6:51. We **identified** loss-of-function mutations in ATP6V0A2, encoding the a2 subunit of the V-type H+ ATPase, in several families **with autosomal recessive cutis laxa type II or wrinkly skin syndrome**. (kor08_bc)

6:52. Sequencing of LEMD3 **identified** loss-of-function mutations in **all** affected individuals of the three families and in three unrelated individuals with osteopoikilosis (Table 1 and Fig. 4e). (hel04_l)

6:53. Therefore, we analyzed RPGRIP1L as a candidate gene for JBTS and **identified** loss-of-function mutations in three families with typical JBTS, including the characteristic mid-hindbrain malformation. (art07_l)

Examples of *loss of function mutations* with a named disorder but without a verb expressing the epistemic relationship

These cases would appear to be significant in constituting an early stage in identifying the causal role of *loss-of-function mutations*; the identification of such *mutations* as being present in subjects with certain defined disorders. Presumably once certain *mutations* have been *identified* further investigations take place to see if a causal relationship can be established between the *mutations* and the disorder, leading to the presence of structures such as those outlined in sections 6.3.1-6.3.3.

6.3.6 *Loss-of-function mutations* and the verb *to have*

Another variation in verb choice of epistemic significance was the use of *have* to make claims about *loss-of-function mutations*. In such cases the issue of causation is still present though the type of claim made varies, as the following examples illustrate:

6:54. Two families with juvenile hemochromatosis not linked to 1q **were recently found to have** loss-of-function mutations in the gene encoding hepcidin10. (pap04_l)

6:55. Loss-of-function mutations for single miRNA family members **have little** phenotypic effect12. (chi07_nav)

In example 6:54 we find a strategy similar to that described in 6.3.5 above, where the presence of *loss-of-function mutations* is being identified in a subject with a named disorder. This example is particularly interesting in that this fact is being cited as a finding in a previous paper. This appears to show the epistemic process at work within the wider community, since the original report has found that families with juvenile hemochromatosis *have loss-of-function mutations* in a named gene, and the present paper uses this finding as the starting point for further research. In such cases there is no

explicit causal claim being made at this stage but it seems fairly clear that the epistemic value of identifying such *mutations* is the possibility of further work that sets out the etiological process through which these *mutations* CAUSE the named disorder.

In example 6:55 above the causal meaning is even more clear since the *loss-of-function mutations* are reported as having *little phenotypic effect*. In such cases there is a different epistemic process present, since the identification of the *loss-of-function mutations* has apparently already been established. In cases such as this the causal meaning is carried not through verb choice but through the nominalisation of the causal process as an *effect*. There were a number of such examples and these are discussed below in section 6.3.7 as a final common linguistic representation of causation.

6.3.7 *Loss of function mutations* with causation expressed through consequences and effects

A final linguistic strategy identified was the encoding of causative meaning through wordforms such as *consequences* and *effects*. Such nominalisation typically represents not the reporting of the (new) fact that *loss-of-function mutations cause* a disorder but rather a further proposition based on this previously established fact, as in the following examples:

6:56. As long as their functions overlap to some extent, **phenotypic effects of** loss-of-function mutations **will be weak**, which appears to be the case for the many partially redundant genes involved in vertebrate development1, 2, 3, 4, 5, 6. (wag00_a)

6:57. By contrast, **the effect of** loss-of-function mutations in yeast **indicates a minor role** for MLH3 (ref. 5). (ber01_bc)

6:58. From our screening, we found that reducing the activity of these chromatin-modifier hub genes enhances the phenotypic **consequences** of loss-of-function mutations in many diverse genes **acting** in different pathways. (leh06_a)

Examples of *loss of function mutations* with *EFFECT* or *consequences* in *genecorp*

In each case the fact that *loss-of-function mutations cause* something has apparently been previously established and the writer is moving on to use this to construct further knowledge. In example 6:56 the *phenotypic effects of loss-of-function mutations* are assumed by the writer though she goes on to say that they will be weak; this structure appears to exemplify the Hallidayan given/new pattern, but interestingly it is embedded in an if/then logical structure realised here by the clause structure. There is a more straightforward case in example 6:57 where the apparently established effect of *loss-of-function mutations in yeast* is used for the further claim that this *indicates a minor role for MLH3.*

6.3.8 Summary

The string *loss-of-function mutations* is already a nominalisation carrying a causal meaning, since the underlying claim is that the *mutations* cause a *loss-of-function*. However despite this nominalised causal meaning, *loss-of-function mutations* appears in further causal claims in *genecorp*. Indeed, the patterns identified in the concordance lines are remarkably similar to those surrounding the string *mutations in the gene encoding*.

Verb choice is the principal means of epistemic signalling around the node phrase and the most common choice of verb is again represented by the lemma CAUSE (40). Verbs such as *block*, *compromise*, *eliminate*, *impair* (2), *lead to* (3), *lower*, *predispose to* and *result in*, which express causal meaning in ways other than by using the lemma CAUSE are also again common (62) and a range of these was identified. The means of varying these statements for epistemic effect was also similar to those containing the string *mutations in the gene encoding*; where the geneticists writing in *Nature Genetics* wish to fall short of conferring a causative role to this node they again do so not by using hedging devices such as modal auxiliaries or modal adjectives but through choice of verb, using verbs that express a connection of some sort or perhaps even by avoiding any linguistic expression of the relationship between *loss-of-function mutations* and the associated disorder. A final form of causal meaning was found in nominal groups encoding *effects* or *consequences* connected to the *loss-of-function mutations*.

6.4 disease causing mutations

The string *disease-causing mutations* occurs 92 times in *genecorp* according to the clusters function of *WordSmith Tools*. The concordance tool found 89 occurrences of *disease-causing mutations*; figure 6:5 below is the collocation profile of the twenty most frequent collocates.

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
in	57	0	5	1	1	1	0	30	4	6	6	3
the	34	1	3	6	2	6	0	1	5	2	7	1
of	32	3	3	1	7	7	0	2	0	3	4	2
and	28	5	3	2	5	0	0	3	0	6	1	3
to	22	1	6	0	9	3	0	0	0	1	0	2
for	13	1	0	2	3	6	0	1	0	0	0	0
we	11	4	0	0	1	0	0	2	1	1	0	2
identified	10	1	0	1	2	2	0	0	1	2	1	0
been	10	0	2	0	0	0	0	0	3	4	0	1
are	10	2	1	0	1	0	0	4	0	0	2	0
genes	9	1	0	3	1	0	0	0	0	1	1	2
not	8	0	0	4	1	0	0	0	1	1	0	1
were	7	0	1	0	1	0	0	5	0	0	0	0
human	7	2	0	0	0	2	0	0	1	1	1	0
that	6	1	0	0	0	2	0	2	0	0	0	1
a	6	1	1	0	0	0	0	1	2	0	0	1
with	5	0	0	1	0	1	0	0	0	1	0	2
which	5	1	0	0	0	3	0	0	0	0	0	1
this	5	0	0	2	0	0	0	0	1	0	0	2
have	5	0	0	1	0	0	0	2	0	1	1	0

Figure 6:3 Twenty most frequent collocates of *disease causing mutations* in *genecorp*

6.4.1 *disease causing mutations* + the lemma IDENTIFY

The wordform *identified* is the most common lexical collocate of *disease-causing mutations*, with ten occurrences according to Wordsmith Tools. In total 15 occurrences of the lemma IDENTIFY were found, as illustrated in these examples:

6:59. Our analysis of DNA samples from Alexander disease patients **has identified** putative *disease-causing mutations* in four amino acids in the rod and tail domains of GFAP (Fig. 3). (bre01_l)

6:60. Next, we sequenced MKS1 in 22 non-Finnish MKS families available to us and **identified** *disease-causing mutations* in four of them (Table 1). (kyt06_bc)

6:61. We were unable to **identify** *disease-causing mutations* in four families. (hur99_l)

6:62. We carried out whole-genome homozygosity mapping, gene expression analysis and DNA sequencing in individuals with isolated mitochondrial ATP synthase deficiency and **identified** *disease-causing mutations* in TMEM70. (ciz08_bc)

6:63. No *disease-causing mutations* **have been identified** in non-syndromic clefting. (mue02_nav)

Examples of *disease-causing mutations* + the lemma IDENTIFY in *genecorp*

These examples broadly fall into two categories: the geneticists are either reporting that they have successfully identified *disease-causing mutations* or they are reporting the negative finding that *disease-causing mutations* have not been identified. Linguistically there is variation in the structure of the negative form with the adjective *unable* in 'we were *unable* to identify' performing this function in example 6:61 whilst the negation in '*no disease-causing mutations* have been identified' expresses the same meaning in example 6:63. Epistemically both examples constitute a nuanced expression since the language does not express the stronger claim that

There are no *disease-causing mutations* in X

But rather falls short of this by expressing the negative result that the procedures used have *not identified* such mutations.

6.4.2 *disease-causing mutations* + the lemma RESULT + *in*

The string *disease-causing mutations* already contains the wordform *causing* and is an explicit nominalisation of a causal claim. However, further causal claims are found in the concordance lines containing *disease-causing mutations*. Whilst few examples of these were discovered in the list of collocates (*result* only appears as a collocate for times) inspection of the extended contexts of the concordance lines revealed more examples of this form. The most common form of these was *disease-causing mutations* + RESULT + *in*, as illustrated by these examples:

6:64. *Disease-causing mutations* in BRCA1 and BRCA2 **result in** inactivation of the encoded proteins, generally by causing premature protein truncation or nonsense-mediated RNA decay. (str08_nav)

6:65. All *disease-causing mutations* of FOXC1 **have resulted in** autosomal dominant transmission. (ben01_bc)

6:66. Because Char syndrome is an autosomal dominant disorder and both *disease-causing mutations* **resulted in** missense changes. (sat00_a)

6:67 In CHEK2, ATM, BRIP1 and PALB2, most of the *disease-causing mutations* **result in** premature protein truncation or nonsense-mediated RNA decay through nonsense codons or translational frameshifts. (str08_nav)

6:68 As *disease-causing mutations* in these genes **do not generally result in** large pedigrees with multiple breast cancer cases, further susceptibility genes of this class will not easily be mapped by genetic linkage analysis. (str08_nav)

Examples of *disease-causing mutations* with the lemma RESULT + *in* in *genecorp*

Whilst the examples 6:64-6:68 are remarkably similar to the causal claims found in the concordance lines containing *mutations in the gene encoding* and *loss-of-function mutations*, there are far fewer examples of these in the concordance lines containing

disease-causing mutations. This presumably reflects the fact that *disease-causing mutations* already contains an explicit causal claim. Whilst in roughly half of the examples studied *mutations in the gene encoding* and *loss-of-function mutations* formed part of a nominal group functioning as an agent in a causal claim, just a few of such examples were found surrounding *disease causing mutations*: those in the following examples where the *disease causing mutations* are being given a causal role where X *may result from disease-causing mutations*, or where *disease causing mutations* either *prematurely truncate* […] or *cause* X:

6:69. we searched for genes with absent or very low expression **that may result from** two allelic **disease-causing mutations**. (har08_l)

6:70. many **disease-causing mutations** either prematurely **truncate** the protein or **cause** splicing defects that eliminate key domains that might mediate its interactions with BiP. (zog05_nav)

6:71. We first looked for well documented **disease-causing mutations that had been shown biochemically to cause** loss of function of the encoded cotransporter or channel18, 19, 20, 21, 22. (jia08_a)

Finally, it might be assumed that *disease-causing mutations* would occur at a later stage in the epistemic process once the causal claim had been established, but things are alas not this straightforward. This string can also be used not only to nominalise a previously

established claim but to prospect forward in anticipation of a future claim, as in the following example:

6:72. To **identify disease-causing mutations**, we sequenced exons of all known genes and mRNA species in the critical region in affected individuals from K2685 and K4233 (Supplementary Table 1 online). (pad06_a)

6.4.3 *disease-causing mutations* + the lemma FIND

A common lemma found in the concordance lines surrounding *disease-causing mutations* was FIND, which occurred eight times in the 5:5 span of the node, but with four occurrences being the wordform *found* and four being the wordform *find* these do not occur in the list of twenty most frequent collocates listed above. Interestingly five of these 8 examples reported negative results, as illustrated below:

6:73. We **did not find** any *disease-causing mutations* in any of the known genes within the interval. (del06_a)

6:74. We analyzed 18 genes (Fig. 1) in SPG11 index patients by direct sequencing of all exons and their splicing sites but **did not find any** *disease-causing mutations* in 17 of them. (ste07_l)

6:75. We then analyzed one proband from each of ten 'GEFS+ like' families for GABRG2 mutations, but **did not find additional** *disease-causing mutations*, underlining the well-established genetic heterogeneity of human idiopathic epilepsies. (bau01_l)

6:76. We **did not find any** *disease-causing mutations*. (hai00_l)

6:77. **No potentially** *disease-causing mutations* **were found** in the coding region, intron-exon boundaries, the 5' and 3' UTR, or intron 1 of KERA in Cuban CNA1 patients. (pel00_l)

Examples of *disease-causing mutations* and the lemma FIND in *genecorp*

These examples reporting a negative result again exemplify a form of epistemic nuancing in that they fall short of the claim that such *disease-causing mutations* exist, but rather make the more modest claim that their own work has not uncovered any such *mutations*, in much the same way as that described above (6.4.1) where negation is found in clauses involving *disease-causing mutations* and the lemma IDENTIFY.

6.4.4 *Disease causing mutations + in*

The collocate *in* occurs 30 times in the R1 position of the node phrase *disease-causing mutations* accounting for almost a third of the total number of 89. Whilst this proportion

occurring in one specific position might be expected to reveal a semantic sequence or epistemically significant pattern there was found to be a great deal of variation in the type of entity occurring in the prepositional phrase, as the following examples illustrate:

6:78 Taken together, these results strongly suggest that the three protein-truncating mutations and the mutation resulting in the missense Y160D change are the *disease-causing mutations* in these families. (tar07_l2)

6:79 We introduced *disease-causing mutations* in the wild-type construct through recombinant PCR. (ram06_l2)

6:80 We sequenced CEL in 38 probands from a Norwegian diabetes registry16, **known to be negative for** *disease-causing mutations* in the MODY1â€“6 genes. (rae06_a)

6:81 Thus, for genes that are represented by many ESTs in public databases, our method **can actually detect** *disease-causing mutations* in the human population. (iri00_l)

Examples of the string *disease-causing mutations* + *in* from *genecorp*

Whilst a semantic sequence of *disease-causing mutations* + location could be argued to be constructed through this pattern, the semantic variation illustrated in examples 6:78-81 is rather broad and epistemically speaking it is unlikely that geneticists are doing the

same thing when they talk about *disease-causing mutations* 'in the human population' (example 6:81) as they do when they declare that a specific gene is 'negative for' *disease-causing mutations* (example 6:80). It would also seem amiss not to note that example 6:80 would appear to be rather unusual in declaring that a Norwegian diabetes registry is '*known* to be negative for *disease-causing mutations*', and this would appear to be a rare example of a negative result being treated as definitive knowledge that such *mutations* are not present in a given entity.

6.4.5 Other epistemic signalling surrounding *disease-causing mutations*

Whilst the patterns discussed above (6.4.1-6.4.4) were not the only means of constructing semantic sequences or epistemic signalling around the phrase *disease-causing mutations*, the other means identified were not common enough to constitute a pattern and indeed many of them occurred only once within this set of concordance lines. These included the use of adjectives found to the left hand of the node such as the predictable *possibly* and the less predictable *apparent* and *putative* as in the following examples:

6:82. Using microarray techniques that simultaneously assay mRNA levels from tens of thousands of transcripts in individuals, Hartong et al. searched for genes with absent or very low expression **possibly due to** biallelic *disease-causing mutations*. (mun08_nav)

6:83. To distinguish between **putative** *disease-causing mutations* and SNPs, we studied a pair of monozygotic twins discordant for the VWS phenotype and whose parents were unaffected. (kon02_l)

6:84. Assay of AIPL1 in 14 families of European descent with LCA that had not been tested previously for linkage to 17p identified **apparent** *disease-causing mutations* in three additional families, as follows. (soh00_l)

There were also a number of multi-word unit type phrases expressing epistemic meaning, though these were few indeed amongst the 89 concordance lines of *disease-causing mutations*; as can be seen from the following examples, various 4-gram strings (bold) clearly represent explicit epistemic signalling in the concordance lines containing *disease-causing mutations*:

6:85. **It remains to be determined** whether these changes are benign polymorphisms or *disease-causing mutations*. (sun05_l)

6:86. The different haplotypes found in the ethnically unrelated HIBM patients, who were all heterozygotes with respect to the HIBM interval, **suggest the existence of** different *disease-causing mutations*. (eis01_l)

6:87. We did not observe those alleles in the population, **consistent with them being** *disease-causing mutations*. (pan07_l)

In addition to these examples there were also occasional examples of grammatical modality as an explicit epistemic signalling device, as in the following example which would typically be described as a hedge:

6:88. Using microarray techniques that simultaneously assay mRNA levels from tens of thousands of transcripts in affected individuals, we searched for genes with absent or very low expression that **may result from** two allelic *disease-causing mutations*. (har08_l)

There was also an example of the use of *candidate* to signal modal meaning, as in the following example:

6:89. As nine of the identified RP genes encode photoreceptor-specific proteins2, other photoreceptor-specific genes **are candidates for** the remaining *disease-causing mutations*. (pie99_a)

In example 6:89 the labelling of 'other photoreceptor-specific genes' as being '*candidates for* the remaining *disease-causing mutations*' expresses modal meaning in the sense that it expresses the proposition that the photoreceptor-specific genes might be *disease-causing mutations*. However it is interesting to note that instead of expressing this using grammatical modality as might typically be expected, the writers instead encode this linguistically in what is at the surface level an unhedged statement- that X are Y rather than X might be Y. However, the right side of this clause of course contains modal

meaning and thus the claim that is made is one of possibility rather than a definite knowledge claim.

6.4.6 Summary

To sum up: whilst *disease-causing mutations* already contains causal meaning in a nominalised form, investigation of the 89 concordance lines shows that it shares a number of the features of *mutations in the gene encoding* and *loss-of-function mutations* in forming part of further causal claims. The main patterns found around this node demonstrate that the main epistemic processes involving *disease-causing mutations* are the attempts by geneticists to identify or find such mutations and their use of this phrase is often to announce that they have either identified or found or failed to identify or not found such *mutations* in a given entity or set of entities. As such the claims surrounding *disease-causing mutations* rarely contain any epistemic signalling; instead they make outright knowledge claims such as these from the following two examples already seen above:

6:60. Next, we sequenced MKS1 in 22 non-Finnish MKS families available to us and **identified** *disease-causing mutations* in four of them (Table 1). (kyt06_bc)

or

6:65. All *disease-causing mutations* of FOXC1 **have resulted in** autosomal dominant transmission. (ben01_bc)

Such examples once again appear to exemplify the marginal role that predictable epistemic signalling devices such as grammatical modality or modal adjectives play in reporting findings around these node phrases.

6.5 Discussion: Causation in *genecorp*

Sections 6.2-6.4 have presented the results of the investigation of collocation data and concordance lines for the strings *mutations in the gene encoding*, *loss-of-function mutations* and *disease-causing mutations*. Whilst collocation data for *mutations* across the whole corpus might have revealed a relationship between *mutations* and CAUSE, detailed investigation of the concordance lines has shown the extent of this relationship and the linguistic means of making and nuancing causal statements in the vicinity of these three phrases. The close relationship between these strings and causal meaning is realised not just in the construction of claims involving the lemma CAUSE but through a wide variety of verbs that also express causal meaning and some that fall just short of expressing causal meaning, and these were particularly prevalent in the concordance lines featuring *mutations in the gene encoding* and *loss-of-function mutations*. The string *disease-causing mutations* already constitutes a nominalisation of a causal claim but was also found to form causal claims particularly in the form *result in* X, where X is a

syndrome or disorder. What was most revealing was the way in which the texts in *genecorp* construct claims that fall short of causal claims when they use a phrase involving *mutations*. Whilst such structures were of course rarely present surrounding the string *disease-causing mutations* there were a range of strategies used with *mutations in the gene encoding* and *loss-of-function mutations* and these rarely involved predictable elements such as modal adjectives or grammatical modality. However it could be legitimately objected that whilst the concordance lines investigated constitute over five hundred examples of these phrases in total, these may not be representative of the other uses of *mutations* that occur in the corpus, or indeed of the wider existence of causative meaning or use of the lemma CAUSE. Whilst it is simply not possible to examine many more than the six or seven thousand concordance lines used in this thesis in a project of this size, some sense of the scope of causative meaning in *genecorp* can be gained from a concordancing of the lemma CAUSE in *genecorp*. This revealed 6, 214 instances of the lemma CAUSE in the 2,979 texts in the corpus, and the collocation data for CAUSE is shown in figure 6:21:

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
of	3142	295	301	299	246	5	0	703	222	469	387	215
the	2788	319	342	349	287	215	0	265	338	92	315	266
cause	2659	2	1	0	1	0	2652	0	0	0	2	1
in	2234	210	222	212	249	1	0	12	131	558	403	236
caused	2115	0	1	0	0	0	2113	0	0	0	0	1
by	1842	40	30	22	6	1	0	1614	42	35	29	23
a	1489	105	105	84	152	67	0	460	238	55	109	114
causes	1440	0	0	1	0	0	1437	0	1	1	0	0
and	1293	111	140	144	104	132	0	21	94	186	189	172
that	1055	139	190	172	80	336	0	2	8	36	46	46
mutations	983	86	115	188	93	148	0	1	224	66	27	35
to	967	79	84	83	103	358	0	1	49	74	72	64
is	845	71	73	166	76	355	0	4	9	29	34	28
are	401	19	33	114	50	116	0	6	5	19	20	19
mutation	365	28	30	44	44	90	0	0	24	45	42	18
be	325	15	14	49	36	175	0	0	3	6	11	16
gene	318	41	34	43	59	44	0	4	7	7	21	58
disease	318	13	11	17	22	48	0	54	52	43	38	20
or	315	22	28	30	27	13	0	21	33	63	39	39
an	314	25	29	13	40	0	0	102	57	12	15	21

Figure 6:4: The twenty most frequent collocates of the lemma CAUSE in *genecorp*

The collocation data in figure 6:6 illustrates the three wordforms *cause*, *caused* and *causes* that appear in *genecorp*. That there are over six thousand instances of some form of CAUSE in a corpus of 2,979 texts suggested that causative language is indeed common in *Nature Genetics*, dispelling the fear that the examples found in the concordance lines containing *mutations in the gene encoding*, *loss-of-function mutations* and *disease causing mutations* might not be typical of the wider corpus. Whilst there is not space or time within the present study to extend this investigation to the presence of the other verb phrases that were found to encode causal meaning such as *due to*, *result in* etc. it seems likely given the proportions found around these node phrases that these would also be present in considerable number in the wider corpus. We can also see from figure 6:6 that the relationship between *mutations* and causative language extends well beyond the strings investigated in this corpus since both *mutation* and *mutations* occur as frequent collocates of the lemma CAUSE, with 1,348 instances of these in the 5:5 span of the node lemma.

In many cases the tri-lexical phrase containing *mutations* is described as being the cause of a named phemenon or phenomena, usually a syndrome such as *Keutel syndrome* or a feature or set of features such as *late-onset obesity, retinal degeneration and hearing loss* Such examples of routinely unhedged findings would appear to be strong candidates for what text would look like in what Kuhn (1962) called 'Normal Science'; the stage of scientific progress where scientists are engaged in 'puzzle solving' within a widely accepted paradigm where knowledge accretion is a routine and relatively uncontroversial

process. The typically unhedged causative statements surrounding phrases containing *mutations* in genecorp therefore can be proposed as an answer to the question of what Normal Science actually looks like as text.

The conception of what is more or less 'scientific' writing in genetics set out by Carver et al. (2008) is also clearly not consistent with approximately half of the examined statements containing *mutations* in *genecorp*. Careful concordance analysis of hundreds of examples of *mutations in the gene encoding*, *loss-of-function mutations* and *disease causing mutations* reveals that the 'uncertain' or 'relativistic' frame is not definitive of the scientific view of genetic causation. Rather, the form of these claims emerges as just one of a range of options; and they are by no means the most common or preferred way of encoding causation in *Nature Genetics*. Indeed, it seems plausible to hypothesise that in many cases structures such as *is linked to* or *is associated with* are in fact epistemic markers pointing to the need for further work in a particular area that may result in findings of the more deterministic type discussed above; rather than representing an uncertain metaphysical state they may represent an uncertain epistemological state. In the former of these what is being claimed is that genes may or may not play a given role depending on a complex interaction with environmental factors. However, in the latter possibility it is the epistemology that is uncertain; in other words, it is not yet established whether or not a given gene does in fact cause a given phenomenon. Where this is the case the 'relativistic frame' serves not to express the complexity of genetic causation but rather to express the need for further work on a given gene. Moreover, the role of such

claims may be to signal to the rest of the genetic science community that such further work is needed.

It should also be noted that in *genecorp* the occurrences of the lemma ASSOCIATE in these contexts reflects the terminology and discourse of statistical analysis being employed by geneticists. Thus this phraseology is not merely one that seeks to in some sense hedge a claim about causative relations a gene participates in; it is also a technical expression of a particular type of finding from the field of genetics. So, in the following example at least part of the meaning is a purely statistical claim about an *association*:

6:66. Loss-of-function mutations in TRPM6 **are associated with** hypomagnesemia with secondary hypocalcemia, a rare autosomal-recessive disorder8. (gud05_nav)

However, this is not to say that this expression is a final statement encoding the scientific understanding of genetic causation. This form is not chosen because it reflects the complex entanglement of genetic etiology; as the concordance lines make clear, it is perfectly possible for geneticists to express straightforward causal relationships. Rather, it may in many instances be that what is expressed here is that this is all that is known about this relationship at present. Further work may reveal, for example that *loss-of-function mutations* in TRPM6 *cause* hypomagnesemia with secondary hypocalcemia. What emerges from the concordance lines is a much more nuanced and interesting picture of the linguistic encoding of causation in genetics. It is not at all correct to say that geneticists do not talk in terms of genetic causes of disorders and similar phenomena. However, when they do this they identify *mutations* as the agent of this *cause*. Whilst

these *mutations* are located in specific *genes* by devices such as the gene name *TRPM6* or by phrases such as *the gene encoding X* they are not identical to the genes themselves. In fact they are of course a process that has occurred within a *gene*, which has in turn become nominalised and is now designated as the agent for a further process. Thus where the popular reporting of findings in genetics is misleading is not in the mere use of words such *cause* or in the attempt to describe a deterministic relationship but rather in the failure to make the distinction between *mutations* and *genes*.

Geneticists were also found to connect *mutations* to various disorders or syndromes whilst avoiding any linguistic characterisation of any causative relationship between the two phenomena. This can be seen as a further strategy for communicating the state of knowledge surrounding a *mutation*. Thus in addition to a geneticist being able to say that a *mutation causes* a phenomenon, or that it is *associated with* it she can also say that a number of individuals who display a certain phenomenon also have certain *mutations*, as in the following examples:

> 6:67. Here we report *mutations in the gene encoding* RANKL (receptor activator of nuclear factor KB ligand) in six individuals with **autosomal recessive osteopetrosis whose bone biopsy specimens lacked osteoclasts**. (sob07_bc)

6:68. We **identified** *loss-of-function mutations* in ATP6V0A2, encoding the a2 subunit of the V-type H+ ATPase, in several families **with autosomal recessive cutis laxa type II or wrinkly skin syndrome**. (kor08_bc)

6:69. Sequencing of LEMD3 **identified** *loss-of-function mutations* in **all** affected individuals of the three families and in three unrelated individuals with **osteopoikilosis** (Table 1 and Fig. 4e). (hel04_l)

In each of these examples there is an underlying pattern of *mutations* + *location* + *disorder*. Clearly the underlying implication is that the discovered *mutations* may be a causal factor in the disorder, but this is not stated explicitly. Rather the geneticist simply identifies a common *mutation* in a number of individuals who have a given disorder. Thus in example 6:67 *mutations* are *identified* in six individuals with *autosomal recessive osteopetrosis*. Similarly in example 6:68 several families with *wrinkly skin syndrome* are identified as having common *loss-of-function mutations* in ATP6V0A2 and in example 6:69 *all affected individuals* from three families and *three unrelated individuals with osteopoikilosis* are identified as having *common loss-of-function mutations*. Thus whilst the implication is that the *mutations* may be a causal factor in these disorders this possibility is not actually expressed as a hedged proposition. It may well be that this is due to a preference for a finding that is definite. Thus whilst what is really of interest scientifically is that the *mutations* may cause the disorder this is unknown to the writer at this point; what they have established is the commonality of the mutations amongst those affected, and this is what they state. The epistemic implications are expressed not through

a predictable form such as a modal adjective of grammatical modality but through the juxtaposition of a phrase that is usually associated with causal claims (*mutations in the gene encoding*, *loss-of-function mutations*, *disease causing mutations*) and a phenomenon that is typically described as being *caused* in such claims, such as a *syndrome* or *disorder*.

Chapter 7: Ontological categorisation

A second major area of knowledge construction that emerged from this study was ontological categorisation: the labelling of an entity according to its properties as established by scientific processes. The history of focus on the existence or otherwise of particular entities has long been a focus of epistemology and of the philosophy of science (as discussed above 2.1-2.3) and the emergence of this category of knowledge construction in *genecorp* is hardly surprising. What is significant for this study however is the way in which this process is realised linguistically. Whilst this process was observed in a number of the phrases studied the strings *tumor suppressor gene* and *tumor suppressor genes* will be discussed in detail since ontological categorisation emerged as the key concern in the concordance lines containing these phrases.

7.1 tumor suppressor gene

Tumor suppressor gene is the fifth most frequent cluster containing both three lexical elements and the keyword *gene*, occurring 154 times in *genecorp* according to the clusters function of *Wordsmith Tools*. When entered into the concordance of *WordSmith Tools* 154 occurrences were found. The following are identified as the twenty most frequent collocates of *tumor suppressor gene*:

word	no.	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
the	88	11	6	6	27	19	0	1	6	4	7	1
of	86	10	8	17	23	8	0	0	0	6	9	5
a	85	3	3	5	27	39	0	0	5	1	1	1
in	61	1	3	4	3	1	0	18	13	2	7	9
is	45	1	2	7	7	0	0	6	12	4	2	4
that	34	12	5	2	1	0	0	9	2	0	2	1
and	28	2	1	3	1	1	0	5	7	3	3	2
as	26	2	3	10	8	1	0	2	0	0	0	0
to	26	4	2	2	1	0	0	1	2	8	3	3
putative	16	0	0	1	0	15	0	0	0	0	0	0
for	15	2	0	3	1	2	0	3	3	0	0	1
cancer	13	0	0	1	0	1	0	0	1	5	2	3
candidate	11	0	1	0	1	9	0	0	0	0	0	0
which	11	1	2	0	0	0	0	2	4	0	1	1
with	10	1	2	0	2	0	0	1	0	1	1	2
mutations	10	1	6	2	0	0	0	0	1	0	0	0
IDB4	9	1	5	1	0	0	0	2	0	0	0	0
1	9	1	0	2	1	1	0	1	0	1	1	1
inactivation	8	0	1	6	0	0	0	0	1	0	0	0
function	8	0	1	0	0	1	0	5	0	0	0	1

Figure 7:1: The twenty most frequent collocates of *tumor suppressor gene* in *genecorp*

The list of the twenty most frequent collocates of *tumor suppressor gene* is striking in containing a number of lexical words that may have an epistemic function, most notably *candidate* and *putative* which would appear to both mark possibility within the span of the node. Examination of the expanded contexts made this relationship much clearer and indeed identified a much more common and epistemically significant strategy: the labelling of a given *tumor suppressor gene* indicating extant knowledge. The principal linguistic means of epistemic signalling within the expanded contexts of *tumor suppressor gene* were as follows.

7.1.1 named *tumor suppressor gene*

The most common feature found in the concordance data surrounding *tumor suppressor gene* was the name of the gene being refererred to, and hence being labelled as a *tumor suppressor gene*, as can be seen in the following examples:

7:1. Inactivation of **the *tumor-suppressor gene*** PTEN and lack of p27KIP1 expression have been detected in most advanced prostate cancers1, 2. (cri01_l)

7:2. This cell line also lacks the von Hippelâ’’Lindau (VHL) ***tumor suppressor gene*** (der01_pro)

7:3. We considered that inactivation of the Trp53 ***tumor-suppressor gene*** might rescue this cell lethal phenotype, as Trp53 is activated in response to a wide variety of signals of DNA damage20. (jac01_a)

7:4. The protein RB1CC1 (retinoblastoma 1 (RB1)-inducible coiled-coil 1) has been identified as a key regulator of the **tumor-suppressor gene** RB1 (ref. 1). (cha02_l)

7:5. Mutations in the TP53 **tumor-suppressor gene** are found in 70âˆ'80% of BRCA1-mutated breast cancer but only 30% of those with wildtype BRCA1 (ref. 3). (har02_l)

Examples of named tumor suppressor genes from *GENECORP*

What is epistemically significant in each of these cases is that the ontological status of the gene as a *tumor suppressor gene* is apparently already known. Whilst it is undoubtedly significant that it is deemed necessary to mention in each case that the gene is a *tumor suppressor gene*, the new knowledge in each sentence is something in addition to this. For example, in example 7:1 above it is presented as a given that PTEN can be accorded the status of *tumor suppressor gene* (**the *tumor-suppressor gene* PTEN)** and the new knowledge being presented is that inactivation of this has been detected in most advanced prostate cancers. This use appears to correspond to what Latour and Woolgar (1979) call a 'Type 4' statement, in that whilst what is being expressed is apparently a certainty, it is still apparently worth mentioning in a research article, rather than having reached the stage of being such an established fact that it can go unsaid.

7.1.2 *putative tumor suppressor gene*

The most frequent lexical collocate of *tumor suppressor gene* is *putative*, which can be found 16 times in the 5:5 span of the node *tumor supressor gene*, with 15 of these occurrences being in the L1 position, forming the cluster *putative tumor suppressor gene* in examples such as the following:

7:6. Results of transfection studies in experimental animal systems support of the idea that Idb4 is a ***putative tumor-suppressor gene*** in hematologic malignancies (liu05_a)

7:7. Global assessment of promoter methylation in a mouse model of cancer identifies ID4 as a ***putative tumor-suppressor gene*** in human leukemia (liu05_a)

The adjective *putative* would appear to be acting as an epistemic marker here, expressing possibility. As such, *putative* can be seen as a further example of the lexical expression of modality present in *genecorp*. Given that *putative* occurs as a collocate of *tumor suppresor gene* in approximately 10% of its instances and has an overall frequency of 1, 756 in *genecorp*, this is likely to be a highly significant epistemic signalling device both in *genecorp* and arguably in the wider discourse of genetics. In addition to this, further evidence of epistemic signalling appears to be present in the surrounding context of *putative tumor suppressor gene*, as in the following examples:

7:8. Results of transfection studies in experimental animal systems support of the idea that Idb4 is a ***putative tumor-suppressor gene*** in hematologic malignancies (liu05_a)

7:9. evidence for Idb4 as a ***putative tumor-suppressor gene*** in the pathogenesis of cancer, such as shown here for both murine and human leukemia, has, to our knowledge, not been previously reported. (liu05_a)

7:10. We used this system to identify a new ***putative tumor-suppressor gene***, Idb4, (liu05_a)

7:11. Our results did not, however, imply the existence of another common ***putative tumor-suppressor gene*** (blu02_l)

7:12. the role of the ***putative tumor-suppressor gene*** H19 is uncertain3, 4. (spa04_bc)

7:13. Although 17p deletions occur in 25â˜’50% of cases, a ***putative tumor suppressor gene*** remains unidentified7. (mac01_a)

Examples *of putative tumor suppressor gene*

In each of these examples further epistemic devices can be observed in the expanded context of *tumor suppressor gene*. In examples 7:8, 7:9 and 7:11 the notion of evidence is evoked in the surrounding context.

Further examples include that in 7:10 above, where the authors say that they have identified a ***new*** *putative tumor suppressor gene*, rather than merely expressing that they have identified a *putative suppressor gene*. There is also an example of the expression of uncertainty in examples 7:12 and 7:13 which express the propositions that the role of the relevant *tumor suppressor gene* is uncertain and even that a *putative tumor suppressor gene* cannot yet be found:

7.12 the role of the ***putative tumor-suppressor gene*** H19 is uncertain3, 4.

7.13 Although 17p deletions occur in 25â˜’50% of cases, a ***putative tumor suppressor gene*** remains unidentified7.

Each of these examples surrounding the initial cluster *tumor-suppressor gene* appears to exemplify a tendency of epistemic talk to cluster around contested epistemic nodes. Uncertainty surrounding the causal role of *tumor suppressor gene* is initially signalled through the adjective *putative*, and the writers also go on to express this further in explicitly stating that it is new, unknown or that its role is uncertain. Example 7:13 above is particular interesting in that it points to an epistemic stage prior to the identification of a *putative tumor suppressor gene* where some conditions are fulfilled (17p deletions occur in 25â˜’50% of cases) and yet this is not enough to warrant the identification of a *tumor suppressor gene*. Finally the adjective *uncertain* is of interest in this context, being a clear epistemic marker. Interestingly concordance data for *uncertain* in *genecorp*

reveals only 125 occurrences in just 98 of the 2,797 texts, suggesting that this is not a common way of epistemic signalling in genetics.

7.1.3 *candidate tumor suppressor gene*

The second most frequent multi-word unit in which *tumor suppressor gene* participates is *candidate tumor suppressor gene*. In the present example *candidate* appears as a collocate of *gene* indicating the possibility of a particular named *gene* being a *tumor suppressor gene*. This phrase then again appears to exemplify the use of *candidate* as a lexical expression of modality in *genecorp*, and indeed *candidate* occurs in *genecorp* 2,754 times suggesting that this is indeed a common means of expressing epistemic status in genetics. Interestingly the concordance data suggests that a *candidate tumor-suppressor gene* can be both a starting hypothesis for a piece of research and the conclusion of that research, as in the following examples:

7:14. Reactivation of the gene's promoter resulted in reexpression of SLIT2 and suppressed colony growth, defining it as a **candidate tumor-suppressor gene** for breast and lung cancer. (liu05_a)

7:15. We conclude that HIC1 is a ***candidate tumor-suppressor gene*** for which loss of function in both mouse and human cancers is associated only with epigenetic modifications. (che03_l)

In example 7:14 above an initial hypothesis is cited as having been the identification of a *candidate tumor-suppressor gene*, whilst in example 7:15 the conclusion is that Idb4 is a *candidate tumor-suppressor gene* of a modified form, where 'loss of function in both mouse and human cancers is associated only with epigenetic modifications. (che03_l)'.

7.1.4 X is a *tumor suppressor gene*

A second though much less frequent means of textualising the status of a gene as a *tumor suppressor gene* was the use of the copula, and this occurred five times, as in the following examples:

> 7:16**.** TSLC1 is a ***tumor-suppressor gene*** in human non-small-cell lung cancer (kur01_l)

> 7:17. SUFU is a newly identified ***tumor-suppressor gene*** that predisposes individuals to medulloblastoma by modulating the SHH signaling pathway through a newly identified mechanism. (tay02_a)

These examples again label a given gene as being a *tumor suppressor gene*, though through a slightly different form. Though there are not enough examples here to be able to make any confident generalisations, it would appear that the use of the copula is found

at or around the point of discovery; in example 7:16 above *TSLC1* being a *tumor suppresor gene* is the main finding of the paper, whilst in example 7:17 the status of *SUFU* as a *tumor suppressor gene* is explicitly marked as being *newly identified*. However, the following further example of this type indicates that the use of the copula can still be associated with a more hedged claim:

7:18. The observation of bi-allelic alterations in TCF1 in human liver tumors meets the criteria of the classical two-hit recessive model of oncogenesis23, 24 and supports the hypothesis that TCF1 is a ***tumor-suppressor gene*** that is altered early in carcinogenesis, leading to adenoma formation. (blu02_l)

In this example *tumor suppressor gene* occurs in the copula construction *TCF1 is a tumor-suppressor gene* but in this case that construction itself occurs in a *that*-clause within *supports the hypothesis that TCF1 is a tumor suppressor gene*. Whilst this still appears to be contributing to a claim of the type that *X is a tumor suppressor gene* this positioning within a *that*-clause constitutes a modification and slight hedging of the claim, indicating that the copula form may still be positioned within a hedged claim.

7.1.5 *classic/classical tumor suppressor gene*

The use of the label *classic* or *classical* was found to be a further linguistic strategy relating to the ontological status of a gene as a *tumor suppressor gene*. In this case this appears to have a strengthening effect on the claim, and seems to constitute an even stronger claim than either of the previous forms discussed above since the use of CLASSIC

can be understood in the sense that what has been found is a prototypical example where the evidence is exactly and ideally in accordance with the ontological criteria. The following examples illustrate this phenomenon:

7:19. Thus, VHL acts as a classic ***tumor-suppressor gene*** that is inactivated according to Knudson's two-hit hypothesis1. (cor03_a)

7:20. This classical ***tumor-suppressor gene*** is completely inactivated in HCT116 cells by a frameshift mutation of one unmethylated allele and hypermethylation of the other allele7. (tin04_bc)

Indeed, this connection is explicitly made in example 7:19 above, where the writer states that *VHL* meets *Knudson's two-hit hypothesis*. However this decription is complicated by the writers' use of *acts as* as the process in this clause, rather than, for example, the copula. It would appear that *VHL acts as a classic tumor-suppressor gene* falls somewhat short of the proposition 'VHL is a classic tumor suppressor gene'; and yet the use of the word *classic* appears to indicate that the classification criteria have been (ideally) met.

7.1.6 The frame X the X of + *tumor suppressor gene*

Another context for *tumor suppressor gene* is the frame *X the X of + a tumor suppressor gene*, which occurred four times. The similarity in meaning expressed by strings instantiating this pattern can be seen in figure 7:2 below:

X	**the**	**X**	**of**	**a tumor suppressor gene**
implying	the	existence	of	a tumor suppressor gene
indicating	the	presence	of	a tumor suppressor gene
suggesting	the	presence	of	a tumor suppressor gene
in agreement with	the	inactivation	of	a tumor suppressor gene

Figure 7:2: Table illustrating the use of the frame X *the* X *of* + *a tumor suppressor gene* in *genecorp*

In each of the examples *tumor suppressor gene* appears to occur in the context of a hedged claim. Each of these four examples connects some evidence with the possibility that the conclusion to be drawn is that there is a *tumor suppressor gene* present. The words in the first X position appear to have a shared meaning of 'suggests' whilst the words in the second X slot seems to have a shared meaning of 'presence'. Whilst no one word is always present in either of these two slots, the frame itself can be seen as carrying the meaning of 'suggests the presence of' a *tumor suppressor gene*.

7.1.7 functions as a *tumor suppressor gene*

Finally the form *functions as a tumor suppressor gene* can be found four times in *genecorp*. This again appears to be a further example of a lexical expression of a hedged ontological status, since it again fall short of constituting a form such as '*X is a tumor suppressor gene'*, as in the folowing examples:

7:21. Later, it was uncovered that it is in fact the wildtype copy of the gene that functions as a ***tumor suppressor gene*** and is capable of reducing cell proliferation. (pfe01_nav)

7:22. We conclude that SUFU functions as a ***tumor-suppressor gene*** in a subset of desmoplastic medulloblastomas. (tay02_a)

7:23. NF1 functions as a ***tumor-suppressor gene***, and loss of heterozygosity in somatic tissues has been associated with tumor formation3. (git03_l)

7:24. Our results indicate that Notch1 functions as a ***tumor-suppressor gene*** in mammalian skin. (nic03_l)

Examples of functions as a *tumor suppressor gene* in *genecorp*

7.1.8 *tumor suppressor gene* and the lemma *KNOW*

The lemma KNOW occurs only five times within the examples of *tumor suppressor gene*, and only once is it relevant to an epistemic claim about a *tumor suppressor gene*, in the following example:

7:25. can act as a tumor-suppressor gene in paraganglioma genesis but is not known to be a breast *tumorâˆ'suppressor gene* (kur02_bc3)

Whilst the lemma KNOW would therefore not appear to be a significant strategy for signalling epistemic status around the string *tumor supressor gene* it is perhaps worth noting precisely because it is so infrequent in comparison to the naming of a *tumor supressor gene*, or the use of *candidate* and *putative*. What is striking about this is that once again the (perhaps) predictable epistemic strategy (the use of the lemma KNOW) is shown to be of only very marginal relevance in comparison to the unpredictable linguistic markers revealed by the corpus-driven investigation.

7.1.9 *may*

Finally grammatical modality emerged as a further means of epistemic signalling around the categorisation of a named gene as a *tumor suppressor gene* with the wordform *may* occurring five times as a hedging device, as in the following examples:

7:26. suggests that PTPRJ **may** be a *tumor suppressor gene* acting in human colorectal cancer. (rui02_l)

7:27. Together with sequence and functional analyses suggesting that RB1CC1 is an upstream regulator of RB1 expression1, 2, our findings indicate that RB1CC1 **may** be a *tumor-suppressor gene* in breast cancer. (cha02_l)

7:28. The clinical significance of hypermethylation across chromosome 2q14.2 is unclear, but the fact that it is a common event suggests that regions within the cytogenetic band **may** encode **possible** *tumor suppressor gene(s)*. (fri06_a)

7:29. we hypothesized that Idb4 **may** be a **candidate** *tumor-suppressor gene* in cancer, especially in leukemia. (liu05_a)

7:30. A new study indicates that wildtype Kras2 **has properties of** a *tumor suppressor gene* and **may** have the capacity to reduce the transforming potential of oncogenically activated ras. (pfe01_nav)

Examples of *tumor suppressor gene* with the wordform *may* as a hedging device in genecorp

7.1.10 Summary

The epistemic status of phenomena labelled as a *tumor suppressor gene* is often marked epistemically either by being reified as a given (known) *tumor supressor* name with an established label or by modification in the L1 position by a classifier apparently intended to express the possibility that a given gene either is a *tumor suppressor gene* or perhaps that it is able to act as one whilst having other functions. It is unlikely that a linguist could predict such a strategy or would be likely to identify it from a wordlist of thousands of words from *genecorp* and indeed in the case of the strategy labelled above as named *tumor suppressor gene* it would not even in principal be possible to identify this from a wordlist since the epistemic status is merely expressed by the gene name.

7.2 *tumor suppressor genes*

Tumor suppressor genes is the most frequent cluster containing both three lexical elements and the keyword *genes*, occurring 138 times in *genecorp* according to the clusters function of *WordSmith Tools*. When entered into the concordance of *WordSmith Tools* 133 occurrences were found. The following are identified as the twenty most frequent collocates of *tumor suppressor genes*:

word	freq	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
of	88	5	6	7	15	38	0	2	0	3	7	5
in	45	4	4	1	1	11	0	13	3	2	2	4
and	42	3	9	1	1	2	0	14	4	1	4	3
the	34	3	6	9	0	5	0	0	3	4	3	1
to	32	8	4	4	3	1	0	2	3	3	3	1
by	18	1	2	2	1	0	0	3	3	1	4	1
oncogenes	15	1	2	2	6	1	0	0	2	1	0	0
cancer	15	0	0	0	0	0	0	0	3	8	2	2
for	14	0	1	3	3	1	0	1	1	3	0	1
or	14	0	0	2	1	5	0	1	0	1	2	2
are	14	0	1	1	1	1	0	5	1	0	2	2
a	14	2	3	0	0	0	0	0	3	3	1	2
may	13	3	0	0	3	0	0	5	1	0	0	1
mutations	13	1	0	3	8	0	0	0	0	1	0	0
as	12	0	1	0	3	1	0	0	6	0	1	0
that	12	2	1	1	3	0	0	3	1	0	1	0
silencing	10	0	0	2	8	0	0	0	0	0	0	0
with	9	1	1	2	0	0	0	0	1	3	1	0
these	9	1	0	2	2	2	0	0	1	1	0	0
inactivation	9	0	1	0	7	0	0	0	1	0	0	0

Figure 7:3: Twenty most frequent collocates of *tumor suppressor genes*

Whilst all of the 63 phrases identified for further study were treated separately rather than being lemmatised, in practice *tumor suppressor gene* and *tumor suppressor genes* were not found to be operating in very different ways epistemically. It is interesting to note that the wordform *may* occurs 13 times in these concordance lines, demonstrating that examples of grammatical modality acting as an epistemic marker are also present here: it is important to note that such examples can be found in the corpus, and that grammatical modality is present as an epistemic signalling device. However what is really of interest in this thesis is the process of identifying other such devices. The collocation profile for *tumor suppressor genes* was found to be broadly similar to that for *tumor suppressor gene*, though neither *oncogenes* nor *silencing* were found in the previous list. One striking difference was that neither *putative* nor candidate were present in the twenty most frequent collocates of *tumor suppressor genes*, which is surprising since these were identified as two common epistemic signalling strategies for *tumor suppressor gene*. It is difficult to assess the significance of such differences but it seems plausible to suggest that the epistemic processes surrounding *tumor suppressor gene* and *tumor suppressor genes* may be somewhat different. This in turn suggests that treating each separate linguistic form of a lemma, in this case the singular and the plural, may be a worthwhile distinction. Given that the process surrounding *tumor suppressor gene* was identified as being one of ontological categorisation where a specific individual gene was being identified as being a *tumor suppressor gene* it is perhaps not such a surprise that a particular epistemic process might attach to the singular rather than the plural form. Investigation of the concordance lines and extended contexts of *tumor suppressor genes*

found a number of examples of more generalised statements around the node phrase rather than attempts to identify a specific *tumor suppressor gene*, as in the following examples:

7:31. The resultant increased expression of oncogenes and decreased expression of ***tumor suppressor genes*** provide a selective growth advantage to tumor cells that retain such aberrations.(pol02_rev)

7:32. This is consistent with data on several ***tumor-suppressor genes***, of which a single nonmutated allele is retained in tumors. (rui02_l)

7:33. ***Tumor-suppressor genes*** have been identified by retroviral tagging, although they are rare6, 7. (suz02_l2)

In these examples general statements applying to a number of *tumor-suppressor genes* are being made. In example 7:33 the identification of *tumor-suppressor genes* is the issue under discussion, but this is again a general statement about identification methods (in this case retroviral tagging) rather than the naming of actual or potential *tumor suppressor genes*. Where this process of identifying *tumor suppressor genes* was present *candidate* and *putative* were again seen as modifiers of *tumor-suppressor genes*, as in the following examples:

7:34. Our analyses of human genome sequences syntenic to these regions suggest that CYP24, PFDN4, STMN1, CDKN1B, PPP2R3 and FSTL1 are **candidate** oncogenes or *tumor-suppressor genes*. (hod01_l)

7:35. In a search for **putative** *tumor-suppressor genes*, we genotyped DNA from a series of ten adenomas (blu02_l)

Tumor suppressor genes also proved similar to *tumor suppressor gene* in encoding knowledge through naming specific *tumor suppressor genes* (17) 'For example, deletions are important in the inactivation of *tumor suppressor genes*, such as **PTEN37** and **CDKN2A38**' (alb03_rev) and through use of the lemma KNOW (5) to indicate those genes whose status as a *tumor suppressor gene* is to be taken as given. There were also a number of verbs indicating causation including the lemma AFFECT (6) and the lemma CAUSE (6) but these were only present in a few of the 133 concordance lines. The process of ontological categorisation present in the concordance lines surrounding *tumor suppressor gene* was again apparent in the concordance lines featuring *tumor suppressor genes*, with epistemic signalling devices that were previously identified again being apparent, as in the following examples:

7:36. Our analyses of human genome sequences syntenic to these regions suggest that CYP24, PFDN4, STMN1, CDKN1B, PPP2R3 and FSTL1 **are candidate** oncogenes or *tumor-suppressor genes*. (hod01_l)

7:37. In theory, the deleted chromosome regions **may contain** *tumor-suppressor genes*. (tho02_pro)

Example 7:37 here is of particular interest since it apparently shows an earlier stage of this epistemic process, prior to the identification of specific genes as *tumor suppressor genes*, instead identifying a more general region which may contain *tumor suppressor genes*.

7.3 *candidate*

Whilst candidate has emerged as an important linguistic signalling device describing a gene as potentially being a *tumor suppressor gene*, it is again of course unclear at this stage how significant this usage is in terms of the wider corpus and of the discourse of genetics. As *candidate* had been discovered as an epistemic signalling device surrounding *tumor suppressor gene*, I therefore attempted to make an assessment of the significance of this in the corpus as a whole. Once the concordance function of *WordSmith Tools* had identified 2,754 occurrences, it seemed worthwhile to explore the use of *candidate* in *genecorp* and once again the method of exploring collocation data and concordance lines was employed. Since *candidate* usually appears as a modifier in a noun phrase the right side lexical collocates give some indication of the kinds of things that can be described as *candidate* and these included *region* and *SNPs* as can be seen in figure 7:4, below:

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
the	1345	97	115	132	179	316	0	0	94	189	109	114
a	990	51	52	79	252	373	0	1	33	64	44	41
of	909	70	71	104	207	239	0	1	44	30	58	85
genes	880	10	13	16	2	1	0	727	69	23	8	11
for	738	26	28	20	43	41	0	241	241	47	31	20
gene	660	10	30	13	5	4	0	500	31	29	22	16
in	601	42	39	29	64	16	0	10	167	91	60	83
and	554	62	57	66	59	42	0	6	98	52	47	65
to	487	55	65	72	76	11	0	12	68	42	50	36
as	330	18	18	62	132	31	0	9	9	21	17	13
we	292	60	42	63	31	0	0	3	48	20	9	16
is	291	19	48	86	49	0	0	14	29	16	14	16
region	279	7	14	13	0	0	0	179	12	4	31	19
that	217	34	26	13	30	3	0	4	53	21	17	16
identified	172	19	9	23	28	24	0	0	24	17	20	8
by	162	13	13	6	12	5	0	0	30	35	27	21
with	143	8	12	7	12	13	0	1	26	28	18	18
this	139	18	17	10	8	12	0	0	12	42	16	4
be	137	11	21	36	17	4	0	0	5	19	13	11
SNPS	129	7	3	12	6	0	0	83	11	0	6	1

Figure 7:4: The twenty most frequent collocates of *candidate* in *genecorp*

Most notable was the presence of *genes* and *gene* as right side collocates with *genes* occurring 727 in the R1 position, 69 times in the R2 position and 23 times in the R3 position whilst *gene* occurs 500 times in the R1 position, 31 times in the R2 position and 29 times in the R3 position. The phrases *candidate genes* and *candidate gene* were investigated in some detail since these are by far the most common patterns surrounding the node *candidate*, accounting for 1,227 of the 2,754 occurrences. The salient patterns around these node phrases were explored and the semantic fields of location and disease were identified. The phrases *candidate genes in* and *candidate gene in* were typically found to have right-side collocates such as *region* and *interval*, locating the *candidate genes* under discussion, whilst *candidate gene for* and *candidate genes for* had right-side collocates linking a *candidate gene* or *candidate genes* to a specific disorder, as can be seen in the examples below.

7:38. we have identified KCNQ1 as a previously unreported susceptibility gene as well as several other *candidate genes for* **type 2 diabetes mellitus**. (yas08_l)

7:39. The central role of these proteins in the innate immune system of the skin suggested that beta-defensin genes **could be** *candidate genes for* **psoriasis susceptibility**. (hol08_bc)

7:40. Genetic mapping studies **identified** ATR as a *candidate gene for* **Seckel syndrome**, but its location on the physical map had not been defined. (odr03_l)

7:41. Thus, CHEK2 presents a biologically plausible *candidate gene for* breast cancer providing one of the pillars of risk factor assessment. (bro02_nav)

Examples of candidate gene for + disorder and candidate genes for + disorder in genecorp.

It can be seen from these examples and from the overall frequency of *candidate* that it has a much wider role in the corpus, and presumably in the discourse of genetics generally, rather than just a local role realising epistemic signalling around *tumor suppressor gene* and *tumor suppressor genes*. As such this could be argued to be a highly significant finding since it apparently indicates that epistemic signalling within a scientific community may be highly lexical and local to that particular discourse- rather than being part of a small group of items typically studied by linguists working in this area such as *possibly* or *probably* or *may* or *might*.

7.4 *Putative*

Putative was also discovered to have wider use as a marker of modal meaning in *genecorp*, though it was not found to be as frequent as *candidate*, with 1,756 occurrences being identified. Whilst a detailed investigation of the two thousand concordance lines containing *putative* is again outside of the scope of this study, the collocation data for this wordform is useful as an indication of the scope and broad patterns of its use in the wider corpus:

word	no.	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
the	997	97	120	62	46	441	0	0	3	65	90	73
of	688	61	51	51	176	141	0	0	76	60	36	36
a	553	25	41	24	17	358	0	0	4	16	34	34
and	479	42	53	43	82	62	0	1	54	61	44	37
in	372	44	23	18	50	8	0	0	50	71	61	47
to	211	29	25	41	57	3	0	0	9	13	12	22
for	176	14	13	8	33	27	0	0	24	33	11	13
genes	144	18	10	16	9	0	0	38	24	13	8	8
with	143	8	11	8	38	22	0	0	11	13	19	13
protein	139	5	7	16	2	0	0	34	34	21	10	10
that	139	11	16	22	36	5	0	0	14	10	14	11
we	138	24	26	42	16	0	0	0	4	12	11	3
is	126	13	11	8	27	0	0	0	7	26	14	20
binding	117	3	4	4	3	0	0	18	59	22	3	1
by	114	8	4	3	9	3	0	0	3	28	41	15
gene	113	6	13	15	2	0	0	26	13	25	6	7
identified	95	4	4	14	31	12	0	0	7	10	7	6
sites	95	0	1	3	6	2	0	3	29	29	17	5
as	87	7	10	4	19	10	0	0	10	6	10	11
are	82	6	8	3	4	3	0	0	16	18	14	10

Figure 7:5: Twenty most frequent collocates of *putative* in *genecorp*

Unlike *candidate* the words *gene* (77) and *genes* (91) were not quite as salient amongst the right-side collocates of *putative* suggesting that it has a more dispersed role as an epistemic modifier, whilst *candidate* tends more often to refer to a *gene* or *genes*. Other discourse objects modified by *putative* included *protein*, *mutations*, *region* and *domain*. The appearance of *identified* in the twenty most frequent collocates again suggests that the process of identifying a *putative X* is a common linguistic structure in the corpus.

7.5 Discussion: Ontological categorisation in *genecorp*

The material presented in chapter 7 has focused in particular on two phrases: *tumor suppressor gene* and *tumor suppressor genes*. Investigation of the patterns surrounding these strings revealed that the main epistemic issue surrounding these node phrases was one of ontological categorisation; a process where what is at issue scientifically is whether or not a given entity is to be classed in a particular way or given a specific label. What has proven particularly interesting about this process is that in *genecorp* the linguistic means of nuancing claims around this process is not a good fit with the current linguistic understanding of scientific hedging as represented, for example, by Hyland (1998). Whilst the concordance lines featuring these phrases provide plentiful examples of geneticists falling short of making outright claims such as X **is** a *tumor suppressor gene*, rare indeed are the examples of modal adjectives, grammatical modality or what Hyland calls 'epistemic lexical verbs' (Hyland 1998) in these concordance lines. Rather, geneticists use linguistic devices such as *candidate* and *putative* to signal modal meaning. Indeed, not only do these two wordforms emerge as the main ways of nuancing a claim

around the node *tumor suppressor gene*, concordance data for these two wordforms also reveals that they are prevalent in *genecorp* as a whole and are used to create modal meaning around a range of entities.

Interestingly, once the concordance lines were further examined the most common epistemically significant pattern found in the concordance data surrounding *tumor suppressor gene* was neither *candidate* or *putative* but the name of the gene under discussion, as the following examples illustrate:

7:42. The protein RB1CC1 (retinoblastoma 1 (RB1)-inducible coiled-coil 1) has been identified as a key regulator of the **tumor-suppressor gene** RB1 (ref. 1). (cha02_l)

7:43. Mutations in the TP53 **tumor-suppressor gene** are found in 70 to 80% of BRCA1-mutated breast cancer but only 30% of those with wildtype BRCA1 (ref. 3). (har02_l)

The process of identifying the underlying similarity between these examples is instructive. In examples 7:42 and 7:43 the underlying similarity is the collocation of *tumor-suppressor gene* with a gene name, in these cases *RB1* and *TP53*. For the computer it is of course not possible to recognise this similarity automatically and it requires the input of the analyst to recognise this epistemic process. Instead it was the patterns with the formal similarity, such as *putative tumor suppressor gene* and *candidate tumor suppressor gene* that were identified first. Moreover given that *putative* and *candidate* occur in the L1

position these patterns can be recognised simply from the collocation data. Contrary to the discussion above of the relationship between *mutations* and causative language, in this example collocation data is both useful and accurate in assessing the frequency of the pattern: it will identify every example. However further investigation of these concordance lines revealed a more frequent lexical similarity that collocation could not identify: that of the gene name. So in this instance whilst collocation is useful in identifying a site of epistemic pattern (L1 modification) it is not able to identify the most common underlying semantic pattern of an epistemic nature.

The second most frequent strategy identified was realised by the most frequent lexical collocate of *tumor suppressor gene*: the aforementioned *putative*. This was found 16 times in the 5:5 span of the node *tumor supressor gene*, with 15 of these occurrences being in the L1 position, forming the cluster *putative tumor suppressor gene* in examples such as the following:

7:44. **Results** of transfection studies in experimental animal systems **support the idea** that Idb4 is a ***putative tumor-suppressor gene*** in hematologic malignancies (liu05_a)

7:45. Global assessment of promoter methylation in a mouse model of cancer **identifies** ID4 as a ***putative tumor-suppressor gene*** in human leukemia (liu05_a)

Whilst only 15 occurrences might not seem that significant, this constitutes approximately ten per cent of the 154 instances of *tumor suppressor gene*. Some further sense of the significance of *putative* in the discourse of genetics can also be gained by concordancing of *putative* as a node phrase and indeed this reveals that *putative* occurs in *genecorp* 1,756 times suggesting that this is indeed a common means of expressing epistemic status in genetics.

Candidate was also argued above to constitute an epistemic marker, expressing the possibility that a given gene is a *tumor suppressor gene*, in examples such as the following:

7:46. Reactivation of the gene's promoter resulted in reexpression of SLIT2 and suppressed colony growth, defining it as a ***candidate tumor-suppressor gene*** for breast and lung cancer. (liu05_a)

7:47. We conclude that HIC1 is a ***candidate tumor-suppressor gene*** for which loss of function in both mouse and human cancers is associated only with epigenetic modifications. (che03_l)

As such, *candidate* can also be seen as a further example of the lexical expression of modality present in *genecorp*. Given that *candidate* has an overall frequency of 2,754 in *genecorp*, this also seems likely to be a significant epistemic signalling device both in *genecorp* and arguably in the wider discourse of genetics.

In the other strategies identified, the epistemic issue at hand was made more explicit. Five instances of the use of the copula were identified as in the following examples:

7:48. TSLC1 is a ***tumor-suppressor gene*** in human non-small-cell lung cancer (kur01_l)

7:49. SUFU is a newly identified ***tumor-suppressor gene*** that predisposes individuals to medulloblastoma by modulating the SHH signaling pathway through a newly identified mechanism. (tay02_a)

Indeed this is a relatively obvious linguistic means of epistemic signalling and Teubert (2000) has already used search terms of the type 'X *is*' in order to locate claims about a particular discourse item in a corpus. However it is notable that the copula only occurs 5 times in the 154 occurrences of *tumor suppressor gene*; whilst it is present as an epistemic signalling strategy, it is less frequent than less predictable strategies discovered such as *putative* or *candidate*. There were also five instances of the modal auxilliary *may* in the collocation data for *tumor suppressor gene*, confirming that this predicatable epistemic signalling device is used to hedge the status of an object as a *tumor suppressor gene*, as in the following examples:

7:50. suggests that PTPRJ **may** be a *tumor suppressor gene* acting in human colorectal cancer. (rui02_l)

7:51. we hypothesized that Idb4 **may** be a **candidate** *tumor-suppressor gene* in cancer, especially in leukemia. (liu05_a)

Example 7:50. here is unique in combining a further form of hedging with the phrase *candidate suppressor gene*, and this may be due to this forming a hypothesis rather than a finding. However, since this form only occurs once, it is of course very difficult to draw any conclusions about this use. Another infrequent form identified was the use of the label *classic* or *classical*. This was also found to be a linguistic strategy relating to the ontological status of a gene as *tumor suppressor gene*, with three such examples being identified. In such cases the modifying adjective appears to be functioning as an intensifier marking the given object as an archetypal example of a *tumor suppressor gene*, as in the following example:

7:52. This classical ***tumor-suppressor gene*** is completely inactivated in HCT116 cells by a frameshift mutation of one unmethylated allele and hypermethylation of the other allele7. (tin04_bc)

In addition to these examples, the form *functions as a tumor suppressor gene* was found four times in *genecorp*. This again appears to be a further example of a lexical expression

of a hedged ontological status, since it again fall short of constituting a form such as '*X is a tumor suppressor gene'*.

Finally the frame 'X the X of + *tumor suppresor gene*' was identified as a further pattern of epistemic significance, as illustrated by figure 7:2 above. This frame is similar to the time + distance + journey one discussed above in that it appears to constitute an underlying semantic regularity which would again be difficult to identify automatically, especially when the words instantiating both X positions can vary. All four examples appear to function similarly in appraising that the available evidence supports the involvement of a *tumor suppressor gene*, whilst falling short of an outright assertion that a *tumor suppressor gene* is present. As such this frame seems to provide a yet further means of expressing the possibility of the presence of a given object.

The full list of collocates indicates that *candidate* is present five times as a collocate of *tumor suppressor genes* whilst *putative* appears twice. *Tumor suppressor genes* proved similar to *tumor suppressor gene* in encoding knowledge through naming specific *tumor suppressor genes* 'For example, deletions are important in the inactivation of *tumor suppressor genes*, such as **PTEN37** and **CDKN2A38**' (alb03_rev) and through use of the lemma KNOW to indicate those genes whose status as a *tumor suppressor gene* has been established. There were also a number of verbs indicating causation including the lemma AFFECT (6) and the lemma CAUSE (6) but these were only present in a few of the 133 concordance lines.

The phrases *candidate genes* and *candidate gene* were also investigated since these are by far the most common patterns surrounding the node *candidate*, accounting for 1,227 of the 2,754 occurrences. The salient patterns around these node phrases were explored and the semantic fields of location and disease were identified. The phrases *candidate genes in* and *candidate gene in* were typically found to have right-side collocates such as *region* and *interval*, locating the *candidate genes* under discussion, whilst *candidate gene for* and *candidate genes for* had right-side collocates linking a *candidate gene* or *candidate genes* to a specific disorder, as can be seen in the examples below.

7:53. we have identified KCNQ1 as a previously unreported susceptibility gene as well as several other *candidate genes for* **type 2 diabetes mellitus**. (yas08_l)

7:54. The central role of these proteins in the innate immune system of the skin suggested that beta-defensin genes **could be** *candidate genes for* **psoriasis susceptibility**. (hol08_bc)

7:55. Genetic mapping studies **identified** ATR as a *candidate gene for* **Seckel syndrome**, but its location on the physical map had not been defined. (odr03_l)

7:56. Thus, CHEK2 presents a biologically plausible *candidate gene for* breast cancer providing one of the pillars of risk factor assessment. (bro02_nav)

This phraseology was of particular interest in again exemplifying specific linguistic features that have previously been identified as forming part of a 'deterministic frame' that misrepresents geneticists when compared to 'more scientific' relativist descriptions of relationships between genes and human disorders, with this criticism explicitly focussing on talk of a gene '*for*' cancer, in the following way:

> **Instead of saying that scientists have found the genes 'for' breast cancer, one article in The Guardian stated that "Scientists have identified the genes BRCA1 and BRCA2 which can significantly increase the risk [of breast cancer] if they mutate, but these only account for 5% of all cases" (Jha, 2005). (Carver, Waldahl and Breiter, 2008, p.946)**

Whilst it may well be the case that structures such as '*a candidate gene for diabetes*' and '*a candidate gene for Seckel syndrome*' do not in themselves fully express the intricacies of contemporary genetic theory it hardly seems fair to criticise journalism as unscientific for using precisely the same forms that are present in *Nature Genetics*. What is perhaps a more reasonable claim is that these forms will be interpreted differently by geneticists as opposed to non-geneticist members of the general public and that these forms ought therefore to be avoided not because they are not used by geneticists but because they will only be understood in the way in which they were originally intended by members of that particular discourse community. Indeed it was also noted above that these forms tend to occur in the *letter*, *review* and *news and views* text types rather than in articles in *Nature Genetics* and indeed where they do occur in articles they appear in the form '*candidate susceptibility gene for*', with the adjective *susceptibility* seemingly distancing the phrase semantically from one that might be interpreted in a deterministic way. Though there are

too few examples to draw any firm conclusions it seems plausible to suggest that geneticists use the form '*gene for X*' as a more informal phrasing of this relationship; nonetheless given that they do so, it can hardly be surprising that this form is to be found in reports on findings in genetics in the wider media.

Putative was also discovered to have wider use as a marker of modal meaning in *genecorp*, though it was not found to be as frequent as *candidate* with 1,756 occurrences being identified. Unlike *candidate* the words *gene* (77) and *genes* (91) were not quite as salient amongst the right-side collocates of *putative* suggesting that it has a more dispersed role as an epistemic modifier, whilst *candidate* tends more often to refer to a *gene* or *genes*. Other discourse objects modified by *putative* included *protein*, *mutations*, *region* and *domain*.

Chapter 8: Conclusions

8.1 Introduction

This chapter will assess the principal achievements and implications of this thesis. I will begin by summarising the key empirical findings alongside a discussion of these (8.2). I will then consider the success of the thesis in answering the initial research questions (8.3) the main strengths of the study (8.4) and the limitations of the findings (8.5) as well as making some recommendations for further studies (8.6).

8.2 Summary of research findings

8.2.1 Using clusters to investigate epistemic signalling

The findings from chapter five indicated that overall the use of clusters to identify epistemic marking yields many results that do not contain the type of linguistic signalling that is of interest to a study like this one. The clusters often tended to report the application of methodology rather than construct new knowledge claims, such as those relating to *cells*, *cell*, *gene and expression*. Since these tended to be found to occur mainly in methodological contexts where there is little or no explicit epistemic signalling of note (since what is being done is regarded within the discipline as relatively unproblematic) there was often little to be said about these strings. It was argued that the study of these is still of significance in a study of the discourse surrounding geneticists'

publication of their findings (since it reveals common objects of study and sites of analysis within the discipline), but what is most illuminated by these is that the geneticists are often not explicitly constructing knowledge claims. Rather, they are typically reporting on routine procedures that will eventually support such claims. The material that most clearly related to this was presented in chapters 6 and 7.

8.2.2 Causation in genetics

The presence of causative and deterministic language was the most surprising feature of *genecorp*. Given what has been said elsewhere (Carver, Waldahl and Breivik 2008) about the falsely 'deterministic' media representations of findings in genetics it seems extraordinary that the causative language identified in *genecorp* could be present in the most prestigious journal in the discipline. Indeed, Carver, Waldahl and Breivik even identified the use of the word *cause* as a keyword in a 'deterministic frame' that misrepresents genetics in comparison to what they label the 'more scientific' relativist frame that 'uses phrases such as 'genetic link', 'predisposition', 'increased risk' and 'might lead to' thereby indicating that the genetic contribution is uncertain' (2008, p. 945). Yet the lemma *cause* is present thousands of times in *genecorp* and is commonly used to make unhedged (certain) claims about a causal relationship between genes and observable human phenomena including 'disease' which is again a keyword of the 'deterministic frame' according to Carver, Waldahl and Breivik (2008, p.945). Where the claims in *genecorp* differ from the criticised media reporting is in the identification of *mutations* within genes as being the causative agent. Indeed, claiming that a *mutation* has

a particular causative role appears to be a relatively unproblematic epistemological notion in genetics. In this thesis the phrases *mutations in the gene encoding*, *loss of function mutations* and *gene causing mutations* were identified as epistemic node phrases constructing causative claims in *genecorp*, and further corpus-based study of the lemma CAUSE appears to confirm the relationship between *mutations* and *cause* in *Nature Genetics*.

8.2.3 Ontological categorisation in genetics

The phrase *tumor suppresor gene*(s) was identified as a phrase that was of great interest in terms of the linguistic encoding of scientific knowledge. What distinguished this phrase as such was the presence of a pattern of variation around the node phrase that was of epistemic significance; in this case, the presence of adjectives such as *putative*, *candidate* and a number of other strategies apparently hedging the ontological status of the given object. The list of the twenty most frequent collocates of *tumor suppressor gene* was unusual in that it included a number of lexical words that appeared likely to have an epistemic function, most notably *candidate* and *putative*. Whilst it was relatively easy to identify these words as having this function once they appeared on a collocation list they are not typical or obvious examples of epistemic marking; indeed, an analyst taking a corpus-based approach that used specific linguistic features as the starting point for analysis would have been very unlikely to have identified these techniques of expressing modal meaning.

8.2.4 Lexis and epistemology: a summary

The broad similarity between the studies of causation and ontological categorisation in genetics that are detailed above is a crucial one: in both cases the linguistic nature of epistemic signalling is primarily *lexical*. Whether a geneticist chooses to say that a mutation *causes* a disorder or is merely *associated with* it is clearly of great significance in epistemic terms but the lexical processes used to express these claims are equally interesting. What is fascinating about these examples is that the variation around the node phrases that constitutes the linguistic signalling of knowledge is of a kind that is not likely to be accessible to a non-specialist analyst through intuition alone. A naïve view of such epistemic marking might have predicted that propositions such as:

Mutations in the gene BRAC1 ***may cause*** breast cancer.

Would appear in the discourse and slowly go through a process of 'de-modalisation' of such as that suggested by Latour and Woolgar (1979) where this proposition, if it is eventually accepted as knowledge, will reappear shorn of the modal auxilliary in a form such as this:

Mutations in the gene BRAC1 **causes** breast cancer.

However this is not what actually happens. The nuancing of epistemic status in propositions around these nodes is not principally expressed through such predictable means as modal adjectives, modal auxilliaries or reporting verbs but instead is expressed

lexically through a wide range of devices. When geneticists report findings relating to *mutations in the gene encoding*, *loss-of-function mutations* and *disease causing mutations* they are usually expressing causal relationships between *mutations* and various human disorders. When they do so they often produce unhedged claims about the causative role of *mutations*, using the lemma CAUSE and various other verb phrases that appear to be functioning as synonyms of CAUSE. What they do when they attempt to nuance such claims is equally fascinating. The two principal means of doing this both involve not the addition of modal auxiliaries to the lexical verb but variation of the verb phrase itself. They typically either use verbs that fall short of causation, in particular the lemma ASSOCIATE, or they elide the verb altogether, providing a structure which juxtaposes some form of *mutations* with a human disorder but does nothing to characterise the relationship between the two.

The phrases *tumor suppressor genes* and particularly *tumor suppressor gene* illustrate a different type of scientific process: that of ontological categorisation. However what is again seen from the concordance lines featuring these phrases is that the naïve view of epistemic signalling will again not describe adequately what actually takes place. Such a view might expect hedged propositions such as:

TRG1 **may** be a *tumor suppressor gene*

or

It is **probable** that TRG1 is a *tumor suppressor gene*

and that these propositions, if they are eventually accepted into the discourse community, will again 'drop' this modality. This is again rarely seen in the concordance lines. What happens instead is that nuancing of claims about *tumor suppressor genes* is realised by a wide range of mostly lexical processes. Moreover, these are again often processes that would by no means have been obvious or predictable to the intuition of an analyst prior to inspecting the data. Pre-modification through *putative* and *candidate* as well as less common examples such as *classic* was found to be the principal means of expressing modal meaning around these nodes in examples such as 8:1 and 8:2 below:

8:1. Global assessment of promoter methylation in a mouse model of cancer **identifies** ID4 as a ***putative tumor-suppressor gene*** in human leukemia (liu05_a)

8:2. We conclude that HIC1 is a ***candidate tumor-suppressor gene*** for which loss of function in both mouse and human cancers is associated only with epigenetic modifications. (che03_l)

Where such nuancing was not present the most common form of expression that indicates a known *tumor suppressor gene* was that described above as 'named *tumor suppressor gene*'; in such cases the proposition '*X* is a *tumor suppressor gene*' has apparently undergone a process of nominalisation indicating that the status of *X* as a *tumor suppressor gene* is now established in the discourse community and no longer

requires any form of nuancing, hedging, or indeed stating as a piece of knowledge in its own right. For 'RB1 is a *tumor suppressor gene*' is apparently no longer the live epistemological issue. Rather this piece of knowledge forms part of the web of background understanding that supports the new finding that 'the protein RB1CC1 [is] a key regulator of the *tumor suppressor gene* RB1. In the few examples where the naming of a *tumor suppressor gene* is the new finding expressed this was in the form of the copular. However, there were only five such examples. Indeed whilst more predictable means of expressing epistemic status (such as the lemma KNOW, or the modal auxiliary *may*) were present in the concordance lines these were by no means the most common strategies for marking epistemic status, or even for expressing modal meaning. Rather, they featured amongst a number of strategies that were revealed by inspecting collocation patterns, concordance lines and the extended contexts of the phrases studied.

8.3 Research questions

8.3.1 What method can be proposed to achieve findings about the linguistic nature of epistemic signalling in genetics?

Corpus studies provide an excellent basis for the empirical study of a lexical item once that item has been selected. However, it is a far from straightforward process to identify which of the thousands of items in a corpus should be chosen for further study. In the research leading to this thesis various methods were explored, and clusters containing highly frequent lexical items were finally chosen. This method had the advantage of isolating a number of phrases that were both present in a number of texts across a number of years and frequent enough for a study of the collocation, concordance and syntactic features present.

The pilot study of this technique revealed that the highly prevalent use of long nominal and prepositional phrases that typically surround node words suggests that collocation data for such nodes may be of limited use when approaching texts in the field of genetics. It was found that even a common syntactic relationship can occur outside of the 5:5 window of analysis that is achievable using collocation software. In order to address this difficulty collocation data was retained only as a heuristic device, useful in identifying some common lexical and grammatical patterns surrounding the node but with the findings and particularly the frequency of these patterns being regarded as only of limited use. Inspection of concordance lines and the wider textual context was undertaken in order to discover the nature and frequency of common syntactical patterns where these occurred outside of the 5:5 span of the node. Where this method is successful as an application of corpus linguistic is in the identification of the elements that are typically surrounded by knowledge claims in genetics, labelled in this thesis as *epistemic nodes*.

A number of these items were studied in great detail in order to explore knowledge construction in genetics. Items such as *mutations in the gene encoding*, *loss-of-function mutations* and *tumor suppressor gene* were found to be constitutive of new knowledge claims being proposed and thus can be seen as elements that are commonly present in the construction of new facts in genetics. In this sense these phrases have been suggested as 'preferred ways' of talking about *mutations* and *genes* when geneticists are reporting new findings and have been revealed as important lexical items in the discourse. Through this technique the thesis has been able to bridge the apparent gap between the highly concrete

'bottom up' approach of corpus linguistics and the theoretical 'top down' approach of philosophy social epistemology.

8.3.2 Can the methodology employed produce findings about the linguistic nature of epistemic marking in genetics that are not wholly predictable?

Most of the work in the study of epistemic signalling in scientific texts fails to tackle the problem of the pre-selection of the *linguistic* aspects of epistemic signalling. A key aim for this thesis was to face this problem: to look for a way to study epistemic signalling in a specialised discipline that is open to the possibility of finding something entirely new. A number of instances of such linguistic processes have been discovered including the use of adjectives such as *putative* and *candidate* in ontological hedging as a concise form of a proposition such as 'X might be a gene for breast cancer' and the use of lexical verb choice to moderate the strength of causative claims surrounding epistemic nodes containing the word *mutations*. What is notable about these findings is that the use of these adjectives is of comparable frequency to those predictable linguistic devices that are frequently discussed in the context of hedging. Thus whilst *possible* (3,275) and *might* (3,850) are certainly frequent in genecorp, *candidate* (2,754) appears to be of comparable frequency, whilst *putative* (1,756) if not quite so common, is certainly noteworthy. Indeed, where epistemic node phrases were investigated in detail such predictable hedging devices were found to be less frequent than other (unpredictable) strategies that were discovered through concordance lines and extended contexts. Thus whilst *may* (6) and *might* (6) are the most frequent of the previously identified lexical hedges found in all

219 sentences including the phrase *loss of function mutations* they are relatively infrequent, and certainly less frequent than a number of other devices. Such devices included choosing a lexical word that expresses a relationship between the node and a phenomenon that falls short of causation (such as *associated*) or indeed by just stating that both *loss of function mutations* and a phenomenon are present in a subject without making any attempt to characterise this relationship.

This strategy of juxtaposing elements that are often found in a causal relationship is of particular interest. Where this is found, the epistemic expression has in effect been elided, and one strong advantage of taking a corpus-driven approach to epistemology is that this type of writing strategy can be revealed through concordance and wider context investigation, whereas a corpus-based study of, for instance, modal verbs would presumably not be able to identify such techniques, spotting only those which are most obvious and predictable. Moreover, it seems reasonable to advance the hypothesis that linguistic devices such as *putative* and *candidate*, and indeed verb choices such as *is linked to* or *is associated with* have roles that are far more significant than merely hedging a claim; they would appear to act as signifiers to the discourse community of the next relevant epistemological step that is required in the scientific process. Thus if a gene is named as a *candidate gene* for (ultimately) the cause of a particular syndrome this would indicate not only the possibility that the cause is to be discovered in a particular location but also that *a specific task within the discipline is now required* in order to further develop understanding of this particular object. So whilst some researchers within the discipline identify statistical *associations* between *genes* and a phenomenon, others

then investigate these *associations* to attempt to discover *candidate genes* and finally the research into a *candidate gene* may lead to the declaration that *mutations* therein *cause* the given phenomenon. These remarks must of course remain as no more than hypotheses at this stage however; as is discussed below (8.5.4) one of the limitations with the vertical approach is that whilst it is a highly effective tool for revealing an ontology of possible statements or predicates surrounding a discourse object in the corpus, it is unfortunately a completely separate and additional process to then attempt to view these claims diachronically, tracing the epistemological processes which a discourse object passes. The possibilities for further work of this type are discussed below in the recommendations for further research (8.6).

8.4 Strengths of the research

8.4.1 Corpus construction

The corpus created for the study constitutes every text published in *Nature Genetics* over a ten year period from January 1999 to December 2008. This amounts to a specialised corpus of almost 10 million tokens. *Nature Genetics* was chosen due to its status as the most prestigious journal in the field of genetics; a status that has been established through and is reflected by the journal being the most cited in the field according to the Thomson Reuters *Journal Citation Reports* as well as being one of the ten most cited journals in all scientific disciplines. Whilst it has often been the practice in corpus linguistics to sample and compare journals from a number of scientific fields this was not undertaken here. The

reason for this lies in the attempt to achieve findings that are relevant to the focus of social epistemology and the philosophy of science by investigating discourse objects and purported entities in science. The aim of the current study was to identify a number of discourse objects and explore the epistemic signalling typically found in the environment of such objects. Since it is unlikely that scientific disciplines will share discourse objects this variable will of course not be comparable in this way. Instead the aim was to create a corpus that would be large enough to study lexical items and patterns and to in some sense be regarded as 'representative' of the discipline. *Genecorp* fulfils these criteria by capturing 2,979 texts from *Nature Genetics*, giving the analyst access to every single occurrence of a node phrase over a ten-year period in the most highly regarded journal in the discipline.

8.4.2 Inductive methodology

The key advantage of the corpus-driven approach is the potential to discover hitherto unknown and unsuspected linguistic features. Thus the methodology chosen in this study is squarely aimed at approaching the problem of preselection in linguistic studies of epistemology. Groom (2007) draws attention to the current trend of work that tends to 'gravitate towards the same small cluster of language features' (p.40) and a key research aim for this thesis was to apply a methodology in such a way as to extend the range of linguistic features of epistemic signalling that are currently of interest. This has resulted in the identification of linguistic variables such as the choice of lexical verb (*cause* vs *predisposes to* or *is associated with*) and noun phrase modification (*putative*, *candidate*,

classic etc.). The methodological implications of these findings are considerable. If the most common epistemic strategies surrounding a phrase may be the choice of a lexical verb a corpus-based approach that begins by analysing given preselected linguistic features would be extremely unlikely to identify them. In propositions such as

8:3 *Mutations in the gene encoding* the latency-associated peptide of TGF-bold beta1 **cause** Camurati-Engelmann disease (jan00_bc)

and

8:4 *Mutations in the gene encoding* lecithin retinol acyltransferase **are associated with** early-onset severe retinal dystrophy (tho01_bc)

there does not seem to be any feature present that has previously been identified as being a linguistic device expressing epistemic status. There is no grammatical modality, no reporting verbs and no hedging; yet it is clear that they make very different claims. This difference is expressed by verb choice alone.

8.4.3 Relationship between data and theory

The corpus-driven methods employed in this study also enable the analyst to develop more robust claims about epistemology in genetics; both in terms of the role of certain

words in the discourse (such as *mutations*) and the epistemic signalling typically found around such words (such as CAUSE). The exploration of hundreds of concordance lines reveals (for instance) the explanatory role given to *mutations* in the discourse and the consistent relationship to causation found in *mutations in the gene encoding*, *loss-of-function mutations* and *disease causing mutations* creates a compelling picture of the textual nature of this role, providing many examples of the causal relationship between *mutations* and various human disorders. One potential weakness in this methodology is the difficulty in assessing the overall significance of these findings in the wider discourse. In order to address this, the corpus has also been used as a test bed for any claims or hypotheses advanced. Where these claims have clear empirical consequences *genecorp* has been used to assess whether these consequences are apparent in the wider corpus and indeed whether the features discovered are present in sizable numbers outside of the concordance lines studied. The presence of CAUSE elsewhere in *genecorp* (6,221 occurrences) and the collocation of this lemma with *mutations* (983) and *mutation* (365) lends some support to the description above of genetics as primarily allocating the causative role to *mutation(s)* within *gene(s)*. Indeed, given what is reported above about the reliability of collocation data when working with highly nominalized language it is quite plausible that this relationship is even stronger than these figures suggest. Similarly the frequency of *candidate* (2,754) and *putative* (1,756) elsewhere in *genecorp* suggests a wider role for these words as lexical markers of modal meaning in genetics.

8.5 Limitations of the research

8.5.1 Reliance on intuition and the friendly geneticist

One serious doubt at the outset of this thesis related to the reliance on analyst intuition in noticing and interpreting patterns surrounding node phrases. Studying what is in effect the semantics of contemporary research in genetics is by no means a straightforward task for an analyst with no specialist training and there was a real fear that this might not prove possible. In practice this did not prove to be a serious impediment to identifying linguistic patterns of epistemic significance but the interpretation of these patterns was certainly enhanced by conversations with friendly geneticists. It was often a frustration not to be able to put the insights afforded by such experts into a more scholarly framework and I would recommend that where possible collaborative work between linguists working in this area and practising researchers in the field would greatly enhance and secure the interpretations of any findings. One cautionary note regarding such collaborations should be made however; the intuitions of experts in genetics are of course just as fallible as those of linguists. Those researchers I was able to discuss my findings with surprised me by opining variously that geneticists 'don't talk about genes causing things' and even when pushed further that mutations 'shouldn't be described as the cause of something but just a factor'. Such intuitions, which appear to be palpably false on the basis of the findings from *genecorp*, would be well worth exploring in a

sustained and scholarly way elsewhere, but it has been beyond the scope and particularly the funding of this thesis to do so here. A methodology combining a structured interview of a number of specialists in the field with some of the key findings from corpus investigation could prove to be a very illuminating piece of research.

8.5.2 Corpus

The corpus constructed for this thesis has completely comprehensive coverage of *Nature Genetics* over the period 1999-2008. Nonetheless and to my consternation the thorny issue of representativeness appears again when we consider the implications of this thesis. This is because, as Paul Thompson first pointed out to me, *Nature Genetics* may be a bit unusual. Constituting as it does the pre-eminent, most highly reputed and most cited journal in genetics, *Nature Genetics* may, by that very fact, not be typical of writing in genetics. Rather, the texts in *genecorp* might represent the language of excellent or cutting edge research in the field. Now, I would certainly argue that the findings regarded as being the very best in the discipline would be worth studying *a fortiori* since these are presumably the standards to which the field aspires. However, the observation that *Nature Genetics* might be a bit odd weakens any claims this thesis can make about what geneticists typically or usually do. This might explain, for example, why causational language is prevalent in *genecorp* (because only the most significant and explanatory findings will be published in *Nature Genetics*) and perhaps why findings about genes reported in *Nature Genetics* seem rarely if ever to be disputed in later publications (because the findings reported there are authoritative and therefore treated as 'final' by

the discourse community (this latter possibility was again suggested to me by Paul Thompson). Whilst these possibilities are mere speculations they of course suggest a further line of future research: the creation of a larger corpus comprised of a number of subcorpora each representing a different journal in the field. This would allow for the techniques used in this thesis to be employed on a number of comparable corpora in order to discover whether the features identified in *Nature Genetics* are typical only of a particular type of 'cutting edge' research or found in the wider discipline.

8.5.3 Software

Whenever a suite of pre-designed software is used for a corpus study there are limitations placed upon the type of study that can be done and this thesis is, of course, no different in this regard. Whilst *WordSmith Tools 4* functions adequately as a concordancer the kinds of multi-word unit analysis possible has greatly improved in the intervening years since this thesis began. Even amongst such suites of corpus software, significant advances have now been made and the c-grams function in *Wmatrix* and the (enhanced) cluster function available on *WordSmith Tools* are considerably more sophisticated than earlier versions, allowing an analyst to see immediately what trees of related multi-word units are present and to notice the presence of such units even when intervening words interrupt them. This is to say nothing of the possibilities that custom made programs can offer. Indeed, this latter option is of particular interest given the argument outlined above to the effect that collocation data is unreliable when dealing with highly nominalised data; one clear implication of this is that the types of software currently being used to explore highly

nominalised text types are probably inadequate. A study of the syntactic patterns that surround tri-lexical clusters that could solve this problem would potentially provide fascinating (and quick) insights into the lexical associations of such keywords and indeed these might well prove to be of epistemic significance in the way that the relationship between *mutations* and verb choice was shown to be in this thesis. It has been suggested to me (by Oliver Mason) that a shallow parsing program could achieve just such results and this would appear to be a potentially very fruitful line of further enquiry stemming from this study.

8.5.4 The vertical approach

One methodological decision taken very early in this study was that the corpus would be comprised of whole texts rather than being divided into text segments. This was done in order to keep an open mind as to where precisely in a text epistemically significant variation might occur. However one drawback of this approach is that the analyst cannot automatically identify the textual position of a proposition, beyond the '%' figure shown by *WordSmith Tools*. This figure in only really useful in giving a rough idea of the position of a phrase within a text and it would have been worthwhile to get a precise sense of which text segments contained the most epistemically significant examples. In light of work such as that recently undertaken by O'Donnell, Brook, Scott and Mahlberg (2012) it seems likely that an approach that subdivided texts into their constituent sections might help to isolate and focus findings on the most epistemically relevant phrases. Indeed, the results of this study found that the phrases that predominantly seemed to occur within the methods sections tended to be somewhat mundane and

epistemically uninteresting; this is perhaps because phraseological items from these sections are likely to appear as common tri-lexical phrases because the methodology is standard practice and will commonly contain phrases such as 'southern blot analysis, western blot analysis, gene expression patterns and gene expression profiles'. Future studies might therefore profitably focus on the title, abstract, results and discussion sections in order to enhance the possibility of identifying epistemically significant chunks.

8.5.5 Disciplinary specificity

A more wide-reaching implication of this study is the finding that epistemic signalling may be local to a particular discipline and even to specific kinds of discourse objects within that discipline (such as *mutations* vs *genes*). One clear implication of this is that the detailed findings of this study are not likely to be generalisable. However, whilst the specific features discovered in *genecorp* would not be present elsewhere it seems plausible to assume that a wide range of lexical epistemic signalling will be present in other scientific disciplines; albeit that it must be acknowledged that teasing out the precise nature of these would not be a small task. However the extent to which the sub-sections of research papers have been shown to contain particular phraseological chunks encoding particular types of epistemic task is striking; an investigation into, for example, clusters found in results sections of a number of scientific disciplines should prove interesting both in revealing the typical phraseological chunks present in each field and

also in providing the opportunity to investigate the collocation data, concordance lines and wider contexts of these in order to discover more about the nature of epistemic signalling.

8.5.6 Statistical significance

In this thesis tri-lexical clusters containing highly frequent lexical items have been explored as a technique to sample epistemic signalling practices in genetics. Whilst the salience of the ten most frequent lexical words in the corpus provides a motivated starting point for further investigation, the statistical significance of the clusters found in the vicinity of those words cannot be measured using *Word Smith Tools 4*. This would certainly constitute a significant weakness in the thesis if the linguistic features that were subsequently identified were relevant to only a few phrases amongst nearly ten million tokens of text. However, the use made here of the corpus as a test bed for assessing the frequency of such items ensures that their use, if not statistically significant, is certainly not a marginal phenomenon. It is ultimately the position of this thesis that the kinds of quantitative analysis that can profitably be carried out on scientific texts are already being done adequately elsewhere; the purpose of the present work is to take a step back from the study of such quantifiable elements as *that*-clauses, items of grammatical modality and other predictable and formally consistent aspects of hedging and to ask whether there might be a great deal more going on that such work misses.

8.6 Recommendations for further study

8.6.1 The discourse of genetics

This study also seems to shed light on the presence of causative language in popular genetics and the nature of the transformations that are taking place in the transitions from the original pieces of research to the popular texts. Contrary to what has been written elsewhere it is argued above that there is a great deal of causative meaning in the findings of geneticists, albeit that it may subsequently be shown that this is more prevalent in *Nature Genetics* than is typical in the discipline at large. What is fascinating however is the nature of the causal relationships apparently construed in media representations of genetics and indeed in the miscommunication that these constitute. Comparisons of the functional relations of the lemma CAUSE across a number of registers containing writing about genetics may well reveal these transformations in detail. If, as is claimed elsewhere, findings in genetics typically misrepresent the original texts by talking about *genes* causing particular disorders and there being a *gene for* a disorder (Carver, Waldahl and Breivik 2008, p.945) it would seem that the source of the misrepresentation is not in the causative language but in the ascription of causation to the *gene* itself rather than the *mutation*. If the popular version does indeed take this form, this could be a very serious failing since the potential confusion between these two versions of genetic causation may have very significant public heath implications. Carver, Waldahl and Breivik (2008) point to the seriousness of confusing the claim that a gene predisposes or slightly raises the chance of developing an illness with the claim that a gene causes the illness, potentially

leading to a fatalist attitude towards personal health and other related issues. In order to research this issue I will create a large corpus of texts reporting findings in genetics and seek to compare the causative language in these with the findings from this thesis.

8.6.2 Epistemic signalling and social epistemology

This thesis is an attempt to develop a greater understanding of the ways in which geneticists encode scientific knowledge. It has inevitably raised many more questions than it has answered. Most fascinating to me is the question of modal meaning in scientific reports. Whilst a naïve view of such findings might expect to find that new findings are typically surrounded by hedging of some type, this seems to be inconsistent with the data examined in thousands of concordance lines from *genecorp*. This is not to say that such hedging is not present: rather it does not seem to be typically present amongst the phrases studied. When a geneticist chooses to say 'X is a candidate gene for' a disorder rather than 'X causes' the disorder she makes it clear that something is unknown in this regard. Geneticists presumably understand by this that there are a range of possibilities and that more work is required to establish the 'correct' version of events. Yet the claim is not realised as a hedged statement in the traditional sense. Geneticists rarely seem to conclude that 'X *may cause*' a disorder. This raises a great number of questions. One hypothesis would be that this way of writing has developed in order to make claims seem more robust and to increase chances of publication in the most prestigious journal in the field. Thus instead of saying what *may* or *may not possibly* be

the case, the writer declares 'BRAC1 *is* a putative gene for breast cancer': apparently making an outright claim of a finding but with modal meaning still being expressed through the adjective *putative*. The use of *candidate* is of even more interest and it has been suggested above that *candidate* may even perform a particular role in the epistemic process, acting not just to express modal meaning but also to classify a particular gene as requiring further study of a very specific kind; work that might very well be carried out by entirely separate researchers within the scientific community. Corpus research into these uses can of course only go so far. It would again be very interesting to compare the findings in *Nature Genetics* with those in other disciplines, this time in order to establish whether there are indeed particular linguistic features of highly esteemed publications. The hypothesis that science writers are attempting to reformulate modal meaning in such a way as to apparently remove hedging from a finding whilst simultaneously leaving a word such as *putative* or *candidate* is a fascinating one and I would be keen to discover how common this is in the discipline. As is suggested in 8.5.1 above, this is another area where language data requires the interpretation not just of the corpus linguist but also of the professional in the field, and in this sense what is really needed is collaborative research combining the observations of those working in the field with the findings of the corpus linguist.

8.7 Concluding remarks

This brings my discussion to an end. In this thesis I have explored epistemic signalling through an investigation into what I have called tri-lexical clusters; a motivated choice of discourse objects based on Hoey's (1991) observation that terminology is often realised by lexical chains. I have tried to show that there are 'preferred ways' of structuring claims in genetics, and my hope is that this thesis has managed in some small way to extend the current understanding of epistemic signalling. In doing so I have also raised a number of future lines of enquiry that I hope will be seen as being worthy of further investigation.

References

Alston, W. (1994). Belief-forming Practices and the Social. In Schmitt, F. ed. (1994) *Socializing Epistemology*. London, Rowman and Littlefield: 29-52.

Aristotle and J.L. Acrill (1974). Aristotle's Categories and De Interpretione. Oxford: Clarendon.

Aristotle and J. H. McMahon (2008). *The metaphysics*. Mineola, N.Y., Dover ; Newton Abbot : David & Charles.

Baker, M. (1993). Corpus Linguistics and translation studies: Implications and applications. In Baker, M., Francis, G. and Tognini-Bonelli, E. (eds.) *Text and Technology: In honour of John Sinclair*. Amsterdam, John Benjamins: 233-252.

Baker, M. (1999). "The role of corpora in investigating the linguistic behaviour of professional translators." *International Journal of Corpus Linguistics* **4**: 281-298.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London, Continuum.

Baker, P. Gabrielatos., C., Khosravinik, M, Krzyżanowski, M., McEnery, T., and Wodak, R. (2008). "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press." *Discourse and Society* **19**(3): 273-306.

Beeching, K. (2006). Synchronic and diachronic variation: The how and why of sociolinguistic corpora. In Wilson, A., Archer, D. and Rayson, P. (eds.) *Corpus Linguistics around the world*. Amsterdam and New York, Rodopi: 49-62.

Bernadini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In Sinclair, J. (ed.) *How to use corpora in language teaching*. Amsterdam, Benjamins**:** 15-36.

Bernhardt, S. (1985). "The writer, the reader and the scientific text." *Journal of Technical Writing and Communications* **15**(2): 163-174.

Bhatia, V. (1993). *Analysing Genre: Language use in professional settings*. London, Longman.

Bhatia, V. (1997). "The power and politics of genre." *World Englishes* **16**(3): 359-371.

Bhatia, V. (2004). *World of written discourse: A genre-based view*. London, Longman.

Biber, D. (1988). *Variation in Speech and Writing*. Cambridge, Cambridge University Press.

Biber, D. (1992). Using computer-based text corpora to analyze the referential strategies of spoken and written texts. In Svartik, J. (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991*. Berlin, Mouton de Gruyter.

Biber, D. (1993). "Co-occurrence patterns among collocations; A tool for corpus-based lexical knowledge acquisition." *Computational Linguistics* **19**: 549-556.

Biber, D. (1995). *Dimensions of register variation*. Cambridge, Cambridge University Press.

Biber, D. (1996a). "Investigating language use through corpus-based analyses of association patterns." *International Journal of Corpus Linguistics* **1**: 171-197.

Biber, D. (2009). "A corpus-driven approach to formulaic language in English." *International Journal of Corpus Linguistics* **14**(3): 275-311.

Biber, D., Connor, U. and Upton, T.A. (2007). *Discourse on the move : using corpus analysis to describe discourse structure*. Amsterdam : Philadelphia, John Benjamins.

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

Bloor, M. and. Bloor., T. (2007). *The practice of Critical Discourse Analysis: an introduction*. London, Hodder Arnold.

Brett, P. (1994). "A genre analysis of the results section of sociology articles." *English for Specific Purposes* **13**(1): 47-59.

Butler, C. S. (1985). *Systemic linguistics:theory and applications*. London, Batsford.

Caldas-Coulthard, C. and Coulthard, R.M. (eds.) (1996). *Texts and Practices*. London: Routledge.

Cameron, D. (2001). *Working with spoken discourse*. London, Sage.

Carter, R. (1997). *Investigating English discourse: language, literacy, literature*. London, Routledge.

Carter, R. (1995) *Keywords in language and literacy*. London: Routledge.

Carter, R. (2004). *Language and creativity: the art of common talk.* London, Routledge.

Carver, R., Waldahl, R. and Breivik, J. (2008). "Frame that Gene: A tool for analysing and classifying the communication of genetics to the public." *EMBO reports* **9**: 943-947.

Coady, C. A. J. (1973). "Testimony and observation." *American Philosophical Quarterly* **10**: 149-155.

Coady, C. A. J. (1975). "Collingwoord and historical testimony." *Philosophy* **50**: 409-424.

Coady, C. A. J. (1981). "Mathematical knowledge and reliable testimony." *Mind* **90**: 542-556.

Coady, C. A. J. (1992). *Testimony: A Philosophical Study*. Oxford, Oxford University Press.

Condit, C. M., Ferguson, A., Kassel. R., Thadhani, C., Gooding, H. C. and Parrot, R. (2001). "An Exploratory Study of the Impact of News Headlines on Genetic Determinism." *Science Communication* **22**(4): 379-395.

Cook, G. (1989). *Discourse*. Oxford, Oxford University Press.

Cook, G. (1992). *The discourse of advertising*. London, Routledge.

Cook, G. (1994). *Discourse and literature*. Oxford University, Oxford University Press.

Cortes, V. (2004). "Lexical bundles in published and student disciplinary writing: Examples from history and biology." *English for Specific Purposes* **23**(4): 397-423.

Coulthard, R.M. (1985). *An introduction to discourse analysis*. Harlow, Longman.

Coulthard, R.M. (1992). *Advances in spoken discourse analysis*. London, Routledge.

Coulthard, R.M. (1993). On beginning the study of forensic texts: Corpus, concordance collocation. In Hoey, M. (ed.) *Data, description, discourse*: *Papers on the English Language in honour of John Sinclair*. London, HarperCollins: 86-114.

Coulthard, R.M. (1994a). *Advances in written text analysis*. London, Routledge.

Coulthard, R.M. (1994b). "On the use of corpora in the analysis of forensic texts." *Forensic Linguistics* **1**: 27-44.

Culpeper, J. (2009). The meta-language of impoliteness: Using Sketch Engine to explore the Oxford English Corpus. In Baker, P. (ed.) *Contemporary Corpus Linguistics*. London and New York, Continuum: 64-86.

Dancy, J. (1987). *Introduction to Contemporary Epistemology*. Oxford, Blackwell.

Descartes, R. and D. M. Clarke (1999). *Discourse on method and related writings*. London, Penguin Books.

Descartes, R. and D. A. Cress (1993). *Meditations on first philosophy : in which the existence of God and the distinction of the soul from the body are demonstrated*. Indianapolis ; Cambridge, Hackett.

Fairclough, N. (1989). *Language and Power*. London, Longman.

Fairclough, N. (1992). "Marketization of public discourse." *Discourse and Society* **4**: 133-162.

Fairclough, N. (1993). *Discourse and social change*. Oxford, Polity.

Fairclough, N. (1995). *Critical discourse analysis: the critical study of language*. Harlow, Longman.

Fairclough, N. (2000). *New Labour, New Language?* London, Routledge.

Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research*. London, Routledge.

Fairclough, N. (2004). Critical discourse analysis in the new capitalism. In Young, L. and Harrisson, C. (eds.) *Systemic Functional Linguistics and Critical Discourse Analysis; studies in social change*. London, Continuum: 103-122.

Fairclough, N. (2006). *Language and Globalization*. London, Routledge.

Feldman (1974). "An alleged defect in Gettier counterexamples." *Australian Journal of Philosophy* **52**: 68-9.

Ferguson, G. (2001). "If you pop over there: a corpus-based study of conditionals in medical discourse." *English for Specific Purposes* **20**(1): 61-82.

Feyerabend, P. K. (1975). *Against Method: Outline of an Anarchistic Theory of Knowledge*. London, New Left Books.

Feyerabend, P. K. (1976). On the critique of scientific reason. In Howson, C. (ed.) *Method and Appraisal in the physical sciences*. Cambridge, Cambridge University Press.

Feyerabend, P. K. (1978). *Science in a free society*. London, New Left Books.

Feyerabend, P. K. (1981a). *Problems of Empiricism. Philosophical Papers Volume II*. Cambridge, Cambridge University Press.

Feyerabend, P. K. (1981b). *Realism, Rationalism and Scientific Method, Philosophical Papers, Volume 1*. Cambridge, Cambridge University Press.

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In Connor, U. and Upton, T.A. (eds.) *Discourse in the professions: perspectives from corpus linguistics*. Amsterdam, John Benjamins**:** 11-33.

Foley, R. (1994). *Egoism in epistemology*. In Schmitt, F. (ed.) Socializing Epistemology. London, Rowman and Littlefield.

Foucault, M. (1980). *Power/knowledge.* Brighton, Harvester Wheatsheaf.

Foucault, M. (1984). The order of discourse. In Shapiro, M. (ed.) *The Language of Politics*. Oxford, Blackwell: 108-138.

Foucault, M., (1989). *The archaeology of knowledge*. London, Routledge.

Francis, G. (2003). A corpus-driven approach to grammar. In Baker, M., Francis, G. and Tognini-Bonelli, E. (eds.) *Text and Technology: In honour of John Sinclair*. Amsterdam, John Benjamins**:** 233-250.

Francis, G., Hunston, S. and Manning, E., (eds.) (1996). *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins.

Francis, G., Hunston, S. & Manning, E., Ed. (1998). *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London, HarperCollins.

Gettier, E. (1963). "Is Justified True Belief Knowledge?" *Analysis* **23**:121-123.

Gledhill, C. J. (2000a). *Collocations in science writing*. Tubingen, Gunter Narr Verlag.

Gledhill, C.J. (2000b). 'The discourse function of collocation in research article introductions.' *English for Specific Purposes*, 19, 115-135.

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge, MA, Harvard University Press.

Goldman, A. (1987). "Foundations of Social Epistemics." *Synthese* **73**: 109-144.

Goldman, A. (1999). *Knowledge in a Social World*. Oxford, Oxford University Press.

Goldman, A. (2001). "Experts: Which Ones Should You Trust?" *Philosophy and Phenomenological Research* **63**: 85-110.

Goldman, A. (2004). "Group knowledge versus group rationality: Two approaches to social epistemology." *Episteme* **1**(1): 11-22.

Greenhalgh (1998). Narrative based medicine in an evidence based trial. In Greenhalgh, T. and. Hurwitz., B., (eds.) *Narrative based medicine*. London, BMJ Books.

Greenhalgh, T. and Hurwitz., B., (eds.). (1998). *Narrative based medicine*. London, BMJ Books.

Groom (2007). *Phraseology and epistemology in humanities writing: a corpus-driven study*. Birmingham, University of Birmingham. Unpublished PhD thesis.

Habermas (1972). *Knowledge and human interests*. London, Heinemann Educational.

Habermas, J. (1984). *The theory of communicative action. Vol. 1: reason and the rationalization of society*. Oxford, Polity.

Habermas, J. (1987a). *The philosophical discourse of modernity*. Cambridge, Polity.

Habermas, J. (1987b). *The theory of communicative action. Volume 2: Lifeworld and sytem*. Oxford, Polity.

Halliday, M. A. K. (1988). On the language of physical science. In Ghadessey, M. (ed.) *Registers of Written English*. London, Pinter.

Halliday, M. A. K. (1993). Quantitative studies and probabilities in grammar. In Hoey, M. (ed.) *Data, description, discourse: Papers on the English Language in honour of John Sinclair*. London, HarperCollins: 1-25.

Harwood, N. (2005). "Nowhere has anyone attempted . . . In this article I aim to do just that' A corpus-based study of self-promotional I and we in academic writing across four discipline." *Journal of Pragmatics* **37**(8): 1207-1231.

Hill, A. A. (1962). Interview with Noam Chomsky. *Third Texas Conference on Problems of Linguistic Analysis in English*. University of Texas, Texas.

Hoey, M. (1983). *On the surface of discourse*. London, Allen and Unwin.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford, Oxford University Press.

Hoey, M. (2001). *Textual interaction: An introduction to written discourse analysis*. London, Routledge.

Hoey,M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge

Howson, C., Ed. (1976). *Method and appraisal in the physical sciences*. Cambridge, Cambridge University Press.

Hume, D. (1975). *A Treatise of Human nature*. Oxford UK, Clarendon Press.

Hume, D. and S. Buckle (2007). *An enquiry concerning human understanding and other writings*. Cambridge, Cambridge University Press.

Hunston, S. (1989). *Evaluation in experimental research articles*. Birmingham, University of Birmingham. Unpublished PhD thesis.

Hunston, S. (1993). Evaluation and ideology in scientific discourse. In Ghadessey, M. (ed.) *Register Analysis: Theory and Practice*. M. London, Pinter**:** 57-73.

Hunston, S. (1994). Evaluation and organisation in a sample of written academic discourse. In Coulthard, R.M. (ed.) *Advances in written text analysis*. London, Routledge**:** 191-218.

Hunston, S. (2008) 'Starting with the small words: Patterns, lexis and semantic sequences'. *International Journal of Corpus Linguistics* 13/3: 271-295

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge, Cambridge University.

Hunston, S. and G. Francis (2000). *Pattern grammar : a corpus-driven approach to the lexical grammar of English*. Amsterdam, John Benjamins.

Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam, John Benjamins.

Hyland, K. (2004). *Disciplinary discourses : social interactions in academic writing*. Ann Arbor ; London, University of Michigan Press.

Hyland, K. (2008). "As can be seen: Lexical bundles and disciplinary variation." *English for Specific Purposes* **27**(1): 4-21.

Johns, T. (1991). Should you be persuaded: Two examples of data driven learning. In Johns, T. and King, P. (eds.) *Classroom concordancing*. *ELR Journal* Volume 4: 1-13. Birmingham, Birmingham University.

Kirkham, R. (1984). "Does the Gettier Problem Rest on a Mistake?" *Mind* **93**: 501-13.

Krishnamurthy, R. (1987). The process of compilation. In Sinclair, J.M. (ed.) *Looking up*. London, Collins: 62-85.

Kuhn,T. (1959) The Essential Tension: Tradition and Innovation in Scientific Research. In: Taylor, C. eds. (1959) *The Third (1959) University of Utah Research Conference on the Identification of Scientific Talent*. 1st ed. Salt Lake City: University of Utah Press, p.162–74. and New York, Continuum.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, London: University of Chicago Press.

Kuhn, T. S. (1970). Logic of discovery or psychology of research. In Lakatos, I.A. (ed.) *Criticism and the growth of knowledge*. Cambridge, Cambridge University Press.

Kuhn, T. S. (1977). *The essential tension*. Chicago, Chicago University Press: 1-23.

Kuhn, T. S. (1983). "Rationality and Theory Choice." *Journal of Philosophy* **80**: 563-70.

Kuhn, T. S. (1996). *The structure of scientific revolutions*. Chicago, University of Chicago Press.

Lakatos, I. (1968). *The problem of inductive knowledge*. Amsterdam, North Holland.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos, I. and Musgrave, A. (eds.) *Criticism and the growth of knowledge*. Cambridge, Cambridge University Press: 91-196.

Lakatos, I. (1971). Replies to critics. In Buck, R.C. and Cohen, R.S. (eds.) *Boston studies in the philosophy of science, Volume 8*. Dordrecht, Reidel: 174-182.

Lakatos, I. (1976). *Proofs and refutations*. Cambridge, Cambridge University Press.

Lakatos, I. (1977). *Philosophical papers, Volume 1*. Cambridge, Cambridge University Press.

Lakatos, I. and Musgrave, A., Ed. (1970). *Criticism and the growth of knowledge*. Cambridge, Cambridge University Press.

Latour, B. and S. Woolgar (1979). *Laboratory life : the social construction of scientific facts*. Beverly Hills ; London, Sage Publications.

Lee, S. and Ziegeler, D. (2006). Analysing a semantic corpus study across English dialects: Searching for paradigmatic parallels. In Wilson, A., Archer, D. and Rayson, P. (eds.) *Corpus Linguistics around the world*. Amsterdam and New York, Rodopi: 121-140.

Leech, G. (1991). The state of the art in corpus linguistics. In Aijmer, K. *English Corpus Linguistics: Linguistic Studies in Honour of Jan Svartvik*. London, Longman**:** 8-29.

Lew, R. (2009). The web as corpus versus traditional corpora: Their relative utlility for linguists and language learners. In Baker, P. (ed) *Contemporary Corpus Lingusitics*. London, Continuum**:** 289-300.

Locke, J. and S. Pringle-Sattinson (1978). *Essay concerning human understanding*. [S.l.], Harvester Press.

Longino, H. E. (1983). "Beyond 'Bad Science': Skeptical reflections on the value-freedom of scientific inquiry." *Science, Technology and Human Values* **8**: 7-17.

Longino, H. E. (1987). "Can there be a feminist science?" *Hypatia* **2**: 51-64.

Longino, H. E. (1990a). *Scientific Inquiry*. Princeton, Princeton University Press.

Longino, H.E. (1990b). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ, Princeton University Press.

Longino, H. E. (1994). "In search of feminist epistemology." *The monist* **77**: 427-285.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Baker,M., Francis, G. and Tognini-Bonelli, E. (eds.) *Text and Technology: In honour of John Sinclair*. Amsterdam, John Benjamins:157-176.

Mahlberg, M. (2003). "The textlinguistic dimension of corpus linguistics: The support function of English general nouns and its theoretical implications." *International Journal of Corpus Linguistics* **8**(1): 97-108.

Mason, O. and Hunston, S. (2004) The automatic recognition of verb patterns: a feasibility study.' *International Journal of Corpus Linguistics* 9: 253-270.

Mauntner, G. (2005). "Time to get wired: Using web-based corpora in critical discourse analysis." *Discourse and Society* **16**(6): 809-828.

Mautner, G. (2007). "Mining large corpora for social information: The case of elderly." *Language in Society* **36**(51-72).

Mautner, G. (2009). *Corpora and Critical Discourse Analysis*. In Baker, P. (ed.) Contemporary Corpus Linguistics. London and New York, Continuum: 32-46.

McCarthy, M. and. Carter, R. (1994). *Language as discourse: perspectives for language teaching*. Harlow, Longman.

McEnery, T. and A. Wilson (2001). *Corpus linguistics : an introduction*. Edinburgh, Edinburgh University Press.

Millar, N. (2009). "Modal verbs in TIME: frequency changes 1923-2006." *International Journal of Corpus Linguistics* **14**(2): 191-220.

Mindt (2000). *An empirical grammar of the English verb system*. Berlin, Cornelsen.

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

Moser (1986). *Empirical Justification*. Lancaster, Reidel.

Myers, G. (1989). "The pragmatics of politeness in scientific articles." *Applied Linguistics* **10**: 1-35.

Myers, G. (1990). *Writing biology : texts in the social construction of scientific knowledge*. Madison, Wisconsin, University of Wisconsin Press.

Myers, G. (1991). "Lexical cohesion and specialised knowledge in science and popular science texts." *Discourse Processes* **14**: 1-26.

Myers, G. (1992). "Textbooks and the sociology of scientific knowledge." *English for Specific Purposes* **11**(3-17).

Myers, G. (1994). Narratives of science and nature in popularising molecular genetics. In Coulthard, R.M. (ed.) *Advances in written text analysis*. London, Routledge.

Noguchi, J., Orr, T., and Tonio, Y. (2006). Using a dedicated corpus to identify features of professional English usage: What do 'we' do in science journal articles? In Wilson, A., Archer, D. and Rayson, P. (eds.) *Corpus Linguistics around the world*. Amsterdam and New York, Rodopi**:** 155-166.

O'Donnell, M. Brook, Scott, M., Mahlberg, M. and Hoey, M., 2012. Exploring text-initial words, clusters and concgrams in a Newspaper Corpus. Corpus Linguistics and Linguistic Theory. 8, (In Press.)

O'Halloran, K. (2007). "The subconscious in James Joyce's 'Eveline': a corpus stylistic analysis which chews on the 'Fish hook''." *Language and Literature* **16**(3): 227-244.

Oakey, D. (2008). *The form and function of fixed collocational patterns in research articles in different academic disciplines*. Leeds, Leeds University. Unpublished PhD thesis.

Okruhlik, K. (1994). "Gender and the biological sciences." *Canadian Journal of Philosophy Supplementary Volume 20*: 21-42.

Orpin, D. (2005). "Corpus Linguistics and Critical Discourse Analysis: examining the discourse of sleaze." *International Journal of Corpus Linguistics* **10**(1): 37-61.

Pennycook, A. (1994). "Incommensurable discourses?" *Applied Linguistics* **15**(2): 115-138.

Plato and R. Waterfield (1987). *Theaetetus*. Harmondsworth, Penguin.

Pollock, J. (1986). *Contemporary Theories of Knowledge.* London, Hutchinson.

Popper (1969). *Conjectures and refutations*. London, Routledge and Kegan Paul.

Popper, K. R. (1972). *The logic of scientific discovery*. London, Hutchinson.

Popper, K. R. (1974). Normal Science and its dangers. In Lakatos, I. and Musgrave, A. (eds.) *Criticism and the growth of knowledge*. Cambridge, Cambridge University Press: 51-8.

Popper, K. R. (1979). *Objective knowledge*. Oxford, Oxford University Press.

Popper, K. R. (1983). *Realism and the aim of science*. London, Hutchinson.

Sinclair, J. M., Ed. (1987b). *Collins Cobuild English Language Dictionary*. London and Glasgow, Collins.

Rogers, M. (2007). *Lexical chains in technical translation: A case study in indeterminacy*. In. B.E. Antia (ed.) Indeterminacy in Terminology and LSP. Amsterdam: John Benjamins:15-35.

Scott, M., 2004, *WordSmith Tools version 4*, Oxford: Oxford University Press.

Shapiro, M. (1984). *The language of politics*. Oxford, Blackwell.

Shope, R. (1983). *The Analysis of Knowing*. Princeton, Princeton University Press.

Sinclair, J. (ed.) (1987a). *Looking up: An account of the COBUILD project*. London, HarperCollins.

Sinclair, J. M., Ed. (1990). *Collins Cobuild Student's Dictionary*. London and Glasgow, Collins.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford, Oxford University Press.

Sinclair, J. M. (1992). The Automatic Analysis of Corpora. in Svartik, J. (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, Stockholm, Mouton de Gruyter: 379-397.

Sinclair, J. M., Ed. (2004). *How to use corpora in language teaching*. Amsterdam, Benjamins.

Sinclair, J. M. (2005) 'Corpus and Text: basic principles.' In M. Wynne (ed.) *Developing Linguistic Corpora: a guide to good practice*. Oxford: Oxbow books, 1-16.

Sinclair, J. M. and R. Carter (2004). *Trust the text : language, corpus and discourse*. London, Routledge.

Sinclair, J. M. and Coulthard, R.M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. London, Oxford University Press.

Sturgeon, S. (1998). *Epistemology. Philosophy 1: A guide through the subject*. A. C. Grayling. Oxford, Oxford University Press.

Swales, J. M. (1981). *Aspects of article introductions*. Birmingham, Language Studies Unit, University of Aston in Birmingham.

Swales, J. M. (1990). *Genre analysis : English in academic and research settings*. Cambridge, Cambridge University Press.

Teubert, W. (1996). "Comparable or parallel corpora." *International Journal of Corpus Linguistics* **9**: 238-264.

Teubert, W. (2000). 'A province of a federal superstate, ruled by an unelected bureacracy: Keywords of the Eurosceptic discourse in Britain. In C. G. A. Musolff, P. Points and R.Wittlinger (eds.) *Attitudes towards Europe: Language in the unification process*. Aldershot, Ashgate**:** 45-86.

Tipton, I. C. (1977). *Locke on human understanding : selected essays*. Oxford, Oxford University Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam, J. Benjamins.

van Dijk, T. A. (1991). *Racism and the press*. London, Routledge.

van Dijk, T. A. (1992). "Discourse and the denial of racism." *Discourse and Society* **3**: 87-118.

van Dijk, T. A. (1993a). *Elite discourse and racism*. Newbury Park, CA, Sage.

van Dijk, T. A. (1993b). "Principles of critical discourse analysis." *Discourse and Society* **4**: 249-283.

van Dijk, T. A. (2009). *Society and discourse : how social contexts influence text and talk*. Cambridge, Cambridge University Press.

van Leeuwen, T. (2004). *Social semiotics*. London, Routledge.

van Leeuwen, T. (2005). *Introduction to social semiotics*. London, Routledge.

West, G. (1980). "That-nominal constructions in traditional rhetorical divisions of scientific research papers." *TESOL Quarterly* **14**: 483-9.

Widdowson, H. G. (1979). *Explorations in applied linguistics*. Oxford, Oxford University Press.

Widdowson, H. G. (2004). *Text, Context, Pretext: Critical issues in discourse analysis*. Oxford, Blackwell.

Wodak, R. (1989). *Language, Power and Ideology: studies in political discourse*. Amsterdam, John Benjamins.

Wodak, R. (1996). *The genesis of racist discourse in Austria since 1989*. In: Caldas-Coulthard, C. and Coulthard, R.M. (eds.) (1996) *Texts and Practices*. London: Routledge:107-28.

Wodak, R. and Meyer, M., Ed. (2001). *Methods of critical discourse analysis*. London, Sage.