GENERAL INTRODUCTION

Patrick Hanks

1. What is lexicology?

Lexicology is the department of linguistics concerned with the lexicon – the study of words and their meanings. It is distinguished from lexicography, which is the practical business of dictionary making. Lexicography is the subject of a separate collection of papers in the *Critical Concepts* series, edited by R. R. K. Hartmann (2003).

Lexicology is very much more than the formalizing of lexicographic practice. It deals with issues that go to the heart of the nature and use of language. Words are the most obvious manifestation of language as a human artefact, and one might have expected therefore that the study of words would have played a central part in linguistics throughout its history. It is therefore somewhat surprising to find that lexicology and word meaning were until recently neglected in many, though not all, schools of linguistics – but such was the case. The central concerns for much of the twentieth century of the so-called 'mainstream' in linguistics (a term which generally denotes American linguistics) were phonology, morphology, and above all syntax. Much of 'semantics' has been and still is concerned with teasing out the logical form and relationships of propositions, which is very different from teasing out the ways in which people use words to make meanings.

The great American linguist Leonard Bloomfield regarded the lexicon of a language as no more than "really an appendix of the grammar, a list of basic irregularities" (1933: p. 274). The main concern of the most influential American linguist of recent times, Noam Chomsky, has been the identification of principles of a 'universal grammar' that is presumed to govern all languages, and the principles that govern the generation of well-formed syntactic structures in particular languages – not the nature of words and their meanings. Chapter 4 of Chomsky (1965) is entitled "Some Residual Problems", and section 2 of that chapter – less than thirty pages – is entitled "The Structure of the Lexicon" (pp. 164–192). It is concerned with rules for inserting lexical items as terminal nodes into syntactic structures ('phrase-markers'). In it, Chomsky briefly introduces the notion of 'selectional restrictions', such that '*he read the book with great enthusiasm*' is a well-formed sentence of English and '*he resembled his father with great enthusiasm*'

is not (because in the Chomskyan model the verb ***resemble*** is presumed to have a selection restriction in its lexical entry that prevents it from governing an adverbial of manner).

As far as I know, no inventory of lexical entries of the kind hinted at by Chomsky has actually been compiled or attempted for any language. Section 4.2 of Chomsky (1965) also discusses economy of lexical classification, arguing that the noun ***destruction*** should be subsumed as part of the verb ***destroy***, rather than treated an independent but related lexical entry. The very considerable progress in Chomskyan theory concerning universal grammar in the ensuing 30 years did not include development of a significant role for the lexicon. Thirty years later, Chomsky (1995) was to write:

> The lexicon is a set of lexical elements, each an articulated system of features. It must specify, for each such element, the phonetic, semantic, and syntactic properties that are idiosyncratic to it, but nothing more: if features of a lexical entry assign it to some category K (say consonant-initial, verb, or action verb), then the entry should contain no specification of properties of K as such, or generalizations will be missed. The lexical entry of the verb ***hit*** must specify just enough of its properties to determine its sound, meaning, and syntactic roles through the operation of general principles, parameterized for the language in question.
>
> Chomsky (1995: pp. 130–131)

Although 'meaning' is briefly mentioned here, there is no discussion of what the meaning of ***hit*** – or any other word – might actually be. This is, as Jackendoff (2000) observed, a syntactocentric approach to language. Actually, it is very close to Bloomfield's "list of basic irregularities", though Chomsky and Bloomfield are worlds apart in other respects.

Is there, then, nothing more to lexicology? On the contrary! A theme running through many of the papers in the present collection is the central role of the lexicon in both linguistic and psychological theory. To understand what it is to be human, it is necessary to understand the nature of language, and to understand this, it is necessary to understand what words are and how they are used. The radically different perspective of a lexicological approach to linguistic theory is summarized by Igor Mel'čuk (who in other respects is a very traditional theorist) in the following words:

> Most current linguistic theories view a linguistic description of a language as a grammar; a lexicon is taken to be an indispensable but somehow less interesting annex to this grammar, where all the idiosyncrasies and irregularities that cannot be successfully covered by the grammar are stored. By contrast, Meaning-Text Theory considers the lexicon as the central, pivotal component of a linguistic description; the grammar is no more than a set of generalizations over the lexicon, secondary to it.
>
> Mel'čuk (2006: p. 228)

Other writers, too, now assign a central role to the lexicon. Anna Wierzbicka, for example, has argued that syntax is essentially semantically motivated. One could go further and say that syntax is no more than the glue that is used to paste words together. It is in the words themselves, their uses, their cognitive and cultural associations, and their combinations, that meaning – the true meat of linguistic action – lies.

It is important to stress in this introductory section that a 'lexical item' may be a single word, but it may also be a phraseological combination. The problems of reductionism (reducing every sentence analytically to the single words and morphemes of which it is composed) have long been recognized, by writers as different as Mel'čuk, Sinclair, Halliday, and Bolinger, to name but a few. Reductionist analyses lose major parts of the meaning and function of texts and utterances. Against this, Sinclair (2004), for one, argues that meanings are created largely from contexts and may override any or all of the meanings normally assigned to the lexical items in isolation. He draws attention to the tension between the 'openness principle', in which words are used freely in a wide variety of contexts, and the 'idiom principle', in which their normal use is governed by collocations and other aspects of linguistic context. Studying how words are used in combination with one another is therefore of the utmost importance (see section 4.2 of this Introduction). Another contrast to the reductionist approach is the holistic approach to the nature of language expounded by Wray (2002). Wray argues that language users make use of formulaic pre-set phrases in most of their day-to-day utterances, only resorting to reductionist analysis of the lexical items when they absolutely need to understand or say something new.

2. Foundations of lexicology

For an understanding of how words work – their meanings and implications – it is necessary to turn away from American so-called 'mainstream' linguistic theory of the 20th century to other disciplines – the philosophy of language, cognitive science, anthropology – and to linguistics in other cultures – European structuralism (as developed by followers of Ferdinand de Saussure), Russian meaning-text theory (as developed by Mel'čuk in both Russian and French), Maurice Gross's lexicon grammar in French, and the British systemic-functional linguistics of followers of J. R. Firth, in particular John Sinclair. Equally influential have been the ideas of philosophers such as Wittgenstein and Quine and anthropologists such as Brown and Rosch, challenging the conception that a word meaning can be formulated as a set of necessary and sufficient conditions expressing essential properties.

This leads us into a veritable ferment of ideas. Writers in these traditions have contributed a wide variety of views – many of them incompatible with one another – and some strong disagreements. What they have in common is that they posit a central role for the lexicon in language and/or cognition. As far as American generative linguistics is concerned, the main contributions to lexicology have come from post-Chomskyans such as Jackendoff, Lakoff, and Pustejovsky. Jackendoff has argued (2000) against a syntactocentric approach to linguistics and in favour of pluralism – a system of of overlapping and interconnected structures – in which the lexicon has an independent role, with interface to other levels. Lakoff turned for inspiration, not to linguistics, but to anthropology – in particular, to the prototype theory of Eleanor Rosch (see, in particular, Lakoff 1987). And Pustejovsky, for his contribution to our understanding of language – Generative Lexicon Theory – went back for inspiration, not to Chomsky or Bloomfield, but to Aristotle.

The founder of modern structural linguistics, Ferdinand de Saussure (1857–1913), made four basic distinctions which are of the utmost importance for the study of language in general and lexicology in particular. These are:

1. **langue** vs. **parole**
2. **substance** vs. **form**
3. **paradigmatic** relations vs. **syntagmatic** relations
4. **synchronic** vs. **diachronic** study of language

The **langue/parole** dichotomy distinguishes language as a system from language as a set of events. From the point of view of lexicology, it is essential to distinguish the central, focal meaning of a word from the various uses to which it may be put. Under the influence of syntactocentric linguistic theory, parole has been somewhat neglected, for grammar does not vary in quite the same way as the semantics of content words. Slight variations in grammar on different occasions are of comparatively little interest. However, the rich variety of uses to which a word may be applied on different occasions can be a subject of great interest, as the papers on vagueness and variability in this collection show. Vagueness and variability of word meaning make it possible for people to use words in new, previously unknown situations and to use existing terms to denote previously unrecognized phenomena, as well as to talk about existing phenomena in new and interesting ways. Parole – language in use – is clearly, therefore, of great importance for lexicology.

The **substance/form** dichotomy at its most basic refers to the distinction between the meaning of a term and its phonological form when uttered (or its orthographic form when written or the hand movements in a sign language). This brings with it another important Saussurean distinction, namely that between **signified** and **signifier**. Within each langue, there is a system of signifieds (meanings, contents, referents), which are expressed by signifiers

(words and phrases). Ideas about the signified or the level of content were developed in different ways by subsequent European structuralist linguists, including German field theorists such as Trier, Porzig, and Gipper and other theorists such as Hjelmslev and Coseriu, all of whom are represented in volume II of the present collection.

The **paradigmatic/syntagmatic** dichotomy is equally important for lexicology. A paradigmatic relation is the one that exists among words that 'compete' for the same slot in sentences, as for example the adjectives **good** and **bad** or the colour terms **green** and **red**. Syntagmatic relations are those that exist among words used in a given utterance, as for example between a modal verb and a main verb. Words of similar semantic class, for example ***apples, pears, cherries***, are held to be in a paradigmatic relationship, but this does not exclude the possibility of them being used syntagmatically (typically, linked by a conjunction) in a given sentence, as in *He offered us apples, pears, and cherries*. More will be said about this dichotomy below.

The **synchronic** study of a language is the study of it as it exists at any one time – typically, the present day – whereas the **diachronic** perspective involves studying its development over time. In the 19th century almost all lexicological scholarship was diachronic. It was Saussure who gave impetus to the synchronic study of language.

Words can change their meaning idiosyncratically or systematically. Examples of systematic change are given by Coseriu in chapter 25, Volume II, in this collection. Coseriu argues that the Latin distinction between bright colours and dull colours (***niger*** 'bright black' vs. ***ater*** 'dull back', for example) was systematic, but came to be lost in Romance languages, not only for black, but also for white, red, yellow, and other colours. An example of an idiosyncratic change is the well-known case of English ***gay***, which in the 1960s developed the sense 'homosexual', with favourable or neutral connotations, rapidly driving out the older meaning 'happy and cheerful'. An even more idiosyncratic case is that of the English word ***size***, which in the 15th century went from the meaning 'court of law' (a short form of ***assizes***) to 'dimension or magnitude'[1]. The two senses co-existed for a couple of centuries, but eventually the older one died out, even before changes in the system of English local justice in 1972 made the more formal term ***assizes*** itself obsolete.

The **diachronic perspective** is important for lexicology and lexicography alike. Most of the great national dictionaries of European languages (such as the *Oxford English Dictionary*, the *Trésor de la langue française*, and the *Deutsches Wörterbuch* founded by the brothers Grimm) are diachonic studies of the history and development of words. They start each entry with the oldest known meaning of the word and, *sub specie aeternitatis*, trace the development of meanings from there. Thus, OED (second edition, 1989) starts its entry for ***camera*** by explaining various Latin, Italian, and Spanish uses of the word as "an arched or vaulted roof or chamber", including "a

department of the papal curia", and only at sense 3b does it indicate that a camera is an apparatus used in photography. Historical principles of this kind are an admirable way of showing the diachronic development of word meaning, but they have had the unfortunate side effect of encouraging the belief that the oldest meaning of a word is its 'true meaning' and that old meanings are somehow better than new ones. For a full understanding of the meaning of a word, it is important to know its history, but this does imply that a value judgement favouring older meanings is appropriate. An important study of the diachronic perspective of word meaning is Sweetser (1990). The importance of the diachronic perspective is underlined by the fact that even dictionaries that purport to be compiled on synchronic principles, 'placing the modern meaning first', generally include in each main entry a section on the etymology and/or semantic history of the word.

3. The metalanguage of lexicology

The epigraph to Quine's *Word and Object* (1960) is a quotation from Otto Neurath: "Wie Schiffer sind wir, die ihr Schiff auf offener See umbauen müssen, ohne es jemals in einem Dock zerlegen und aus besten Bestandteilen neu errichten zu können" – "We are like sailors who must rebuild their ship on the high seas, without ever being able to take it apart in a dry dock and reconstruct it from scratch using the best components."

This could equally well serve as an epigraph for a collection of articles on the lexicon. Lexicologists are obliged to use words as the means for conducting their research into words and reporting their conclusions. There is no metalanguage other than words themselves (in one language or another) – or at a pinch derivatives of words such as logical symbols – for expressing thoughts about words. One consequence of this is that confusion tends to creep into discussions of word meaning between the object language (a natural human language or a subset of one), the metalanguage (the system used to describe and discuss it), and the objects in the world (physical or abstract) that the words denote. It is obvious (when pointed out) that the meaning of a word is not an object in the world (you can cut down a tree, but you can't cut down the meaning of ***tree*** – or, if you do, you are doing something very different). It is less obvious that concepts are distinct from the words used to represent them, but the very fact that word meaning is so variable is evidence that this is so. The nature of this variability is discussed in several papers in this collection, in particular those in Volume I, Part 3.

This three-way relationship was addressed by Ogden and Richards (1923). They drew a 'semantic triangle' (Figure 1) to illustrate their argument that the relationship between words and objects is indirect, and that the words of a language must be separated both from the external world of facts and from the thoughts and concepts that they represent.

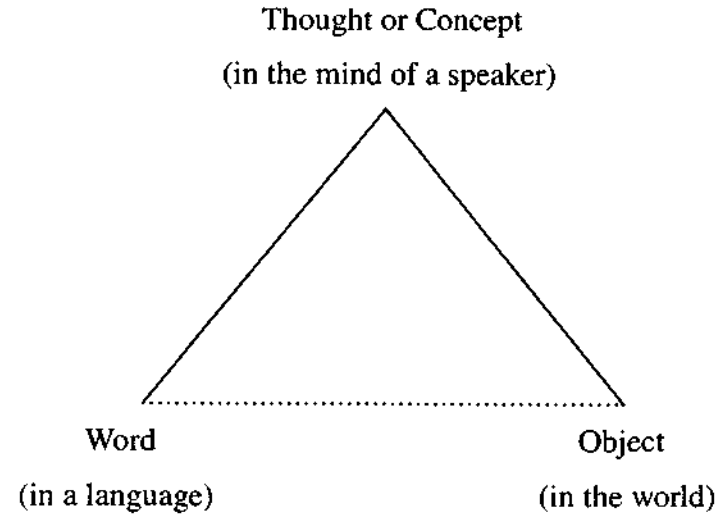


Figure 1 Ogden and Richards' Semantic Triangle.

People use words to talk about objects in the world, but, according to Ogden and Richards, words are not to be confused with objects in the world, nor are they to be confused with the thought or concept that a speaker has in mind when he or she chooses a word to represent a thought. And words stand for objects only indirectly, through the thoughts of people who use them. This three-way separation has some surprising consequences, chief among which is the vagueness or instability of lexical meaning. The meaning of a word may vary slightly in the conceptual schemata of individual speakers, and the way in which the world is conceived may also vary from speaker to speaker, from locality to locality, and from time to time. Human societies exert enormous normative pressures to control and counteract this tendency to variation and to ensure that, even if different speakers do mean slightly different things by a word or phrase, the differences are not great enough to be noticed. If they are noticed, typically they are stigmatized: people who use words unconventionally are generally ignored or laughed at, not congratulated[2].

A second epigraph for this collection, therefore, might be from Wierzbicka (1985: p. 8): "An adequate definition of a vague concept must aim not at precision but at vagueness: it must aim at precisely that level of vagueness which characterizes the phenomenon itself".

Not only are words used as the everyday vehicles of human social intercourse (a fact which makes them a legitimate object of descriptive science); they are also the very tools that are used in pursuit of precision in scientific concepts. This has the paradoxical result that for scientific purposes an ordinary word like ***minute*** or ***second*** may be given a stipulative precision

that it does not have in ordinary language. If your friend asks you to wait a second, he does not mean that you should wait during 9,192,631,770 cycles of radiation corresponding to the transition between two hyperfine levels of the ground state of caesium 133, yet that is the meaning of the word ***second*** assigned to it by the international committee charged with determining the meaning of international scientific units (SI units).

Scientists and philosophers of science have for centuries railed against the imprecision of natural language. Since the 17th century, some have even resorted to constructing artificial languages for the purpose of expressing concepts more clearly (not only inventing symbolic logics but also stipulating very precise definitions for particular words, both newly coined scientific terms like ***ohm***, ***joule***, and ***mammal*** and more familiar terms like ***second***, ***metre***, and ***bird***). Philosophers of science from Wilkins and Leibniz to Russell have wanted to dismantle language and start again from scratch, in order to create a vehicle for conveying concepts with precision. Descriptive lexicologists, on the other hand, want to describe, as precisely as possible, the vagueness and variability that constitutes the reality of words in use. The tension between a scientific desire for precision and the essentially vague, dynamic properties of a natural-language lexicon is another theme that runs through this collection.

3.1 Semantic primitives or primes

Another major issue concerning the language to be used for defining words and concepts is the idea that words with complex meanings should be explained in terms of simpler words. For example, ***run***, ***creep***, ***crawl***, ***ride***, ***drive***, ***climb***, and several other words all contain the basic notion 'move', but differ as to manner of motion, direction, or other semantic feature(s). *Move* is a simple term, suitable for use as the genus word in a definition, whereas a word such as ***motility*** or ***perambulation*** would not be suitable for the same purpose. Many writers in this collection – from Leibniz and Goddard to Wierzbicka – have argued that an explanation of the meaning of a word must be couched in terms simpler than the word itself, and they draw the logical conclusion that there must therefore exist a set of terms that are so simple and basic that they cannot be defined at all. These are called **semantic primitives** or **atomic primes of meaning**. Unlike most theorists who have made such proposals, Wierzbicka has devoted much time and effort to explicitly specifying a set of semantic primes (Figure 2). According to her theory of a **natural semantic metalanguage**, the meaning of all words in all languages can be reduced to combinations of these primes, each in just one sense, which cannot be further reduced. This is a recurrent but controversial claim in lexicology. What is controversial about it is the idea that these basic primitives or primes are psychologically or existentially real, as opposed to being a possibly useful device in the apparatus of lexicological analysis or

Substantives:	I, YOU, SOMEONE, PEOPLE, SOMETHING/ THING, BODY
Determiners:	THIS, THE SAME, OTHER
Quantifiers:	ONE, TWO, SOME, ALL, MANY/MUCH
Evaluators:	GOOD, BAD
Descriptors:	BIG, SMALL
Intensifier:	VERY
Mental predicates:	THINK, KNOW, WANT, FEEL, SEE, HEAR
Speech:	SAY, WORDS, TRUE
Actions, events:	DO, HAPPEN, MOVE, TOUCH
Existence and possession:	THERE IS / EXIST, HAVE
Life and death:	LIVE, DIE
Time:	WHEN/TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT
Space:	WHERE/PLACE, HERE, ABOVE, BELOW; FAR, NEAR; SIDE, INSIDE; TOUCHING
"Logical" concepts:	NOT, MAYBE, CAN, BECAUSE, IF
Augmentors:	VERY, MORE
Taxonomy, partonomy:	KIND OF, PART OF
Similarity:	LIKE

Figure 2 Proposed set of semantic primes (Wierzbicka and Goddard 2002).

computational language processing. For a fuller account of the arguments in favour of semantic primes, see Goddard (1998).

There are several variants of and analogies to this claim. For example, in his theory of **preference semantics** Wilks argues on practical grounds that

for natural language processing a limited set of basic terms (which he calls "aristocrats among words") is necessary for the reduction of the complex meanings actually found in words and texts to something simpler that can be processed by a computer program. In the early 20th century, the philosopher Bertrand Russell argued that concepts should be expressed in terms of 'logical atoms' of meaning. At a rather broader practical level, mention should be made of restricted defining vocabularies for foreign learners of a language, of which the example par excellence is the set of approximately 2000 words developed for the *Longman Dictionary of Contemporary English* (1978; 4th edition 2005)[3].

It is questionable whether there is adequate empirical evidence to justify regarding semantic primitives as linguistically or psychologically real entities. An alternative view, compatible with the more pragmatic line of Wilks, but not with the philosophical argumentation of Leibniz, Russell, and Wierzbicka, is that a natural language consists of a vast network of beliefs, which are of course expressed through and associated with words, some of which have greater connectivity than others (***move*** is connected both semantically and syntagmatically to more words than, say, ***crawl*** or ***motility***). This is not a reason for assigning primitive status to some of them. In other words, the fact that some people would like language to be conceptually neat and well organized around a bunch of primitives does not necessarily entail that it is.

3.2 Componential analysis

Another theme running through many of the chapters in this collection, in particular in Volume II, is that the meaning of a word is not a simple abstract entity, still less a concrete referent to some fixed set of objects in the world, but consists of a number of interrelated components referring to a more or less vaguely defined set of entities, if indeed it refers to entities in the world at all.

At its most basic, the concept of componential analysis enabled lexicologists to express the central, normal meaning of a word – say, ***boy*** – in terms of polarities with plus and minus features, and to contrast the meaning in this way with the meaning of other related terms, as in Figure 3.

boy = +HUMAN, +MALE, −ADULT
man = +HUMAN, +MALE, +ADULT
woman = +HUMAN, −MALE, +ADULT
girl = +HUMAN, −MALE, −ADULT

Figure 3 Example of proposed contrasting binary semantic components.

Tables such as Figure 3 were popular in the 1960s and 70s, but even then they were recognized as an oversimplification. For example, the notion that a girl is −MALE rather than +FEMALE depends on MALE being the 'unmarked' term and FEMALE being the 'marked' term. This was controversial even in the seventies, and is now widely regarded as unacceptable, not to mention politically incorrect. Other objections include the fact that the words ***boys*** and ***girls*** are often used to refer to adult human beings. The obvious counter to this particular objection is that the **core meaning** of ***boy*** and ***girl*** contains the component −ADULT; any uses that contradict this essential feature are metaphorical or otherwise unusual exploitations of the core meaning.

Another objection to componential analysis of the kind expressed in Figure 3 is that not all semantic features consist of binary contrasts. If HAVING LEGS were to be considered as a candidate semantic component (a not unreasonable proposal in itself), problems arise. Mammals, birds, and insects have legs, snakes and worms do not. But it is not merely the presence of legs but the number of legs that is an essential differentia for distinguishing classes of animate beings. Here we have a set of contrasting features that is not binary: it would be +4 LEGS for most land mammals, +2 LEGS for humans and birds, +0 LEGS (or −LEGS) for snakes, worms, and fish, +6 LEGS for insects, +8 LEGS for spiders, +INNUMERABLE LEGS for centipedes and millipedes, and so on. This example suggests that the binary model of positive/negative features for lexical semantic differentiation may be unduly simplistic.

A more serious objection is that it is generally impossible to decide whether a semantic component is essential or accidental. Is +WHITE an essential semantic component of ***swan***? People used to think so until black swans were discovered in Australia. Is having a high-pitched voice a semantic feature of ***boy***? This seems to be carrying things too far, but it is clearly a criterial feature of some kind in the following citation from the British National Corpus:

> Footsteps and the sound of **a boy's voice** were approaching the hedge. Maybe it was the Post Office boy.

(How could they have known it was a boy's voice if voice type is not a criterial feature that distinguishes boys from men?)

This is not the place to go into a detailed discussion of componential analysis. For more detail, the reader should turn to Volume II. Suffice it to say that the notion that the meaning of a word is composed of a structure of components is as old as Aristotle and is still very much with us, most recently having been adapted to probabilistic and contextually governed models of word meaning.

4. Ten questions for lexicology

There are at least ten major questions that may be asked of lexicology:

1. How do words relate to the world around us?
2. How do words relate to one another (within the system of a language)?
3. How is the lexicon of a language structured (insofar as it is structured at all)?
4. What is the relationship between words and concepts?
5. How can word meaning be formalized and so be made machine-tractable?
6. How do young children acquire the words and meanings of their native language?
7. How do the words of one language relate to words with 'the same' meaning in other languages?
8. How do the meanings of words change over time?
9. What is the difference between the literal meaning of a word and figurative meanings such as metaphors?
10. What is the nature of lexical creativity?

In the remainder of this General Introduction, I shall discuss each of these topics briefly.

4.1 Words and the world

To ordinary language users, it may seem a matter of common sense that the meaning of a word represents an entity in the world, or rather a class of entities in the world. However, a little reflection will convince us that this is at best a half truth. It is certainly true that many nouns denote classes of entities in the world: the noun ***tree***, for example, represents a member of a class of physical objects and the word ***oak*** represents a member another class of physical objects. The second is a subclass of the first. But the word ***tree*** is not limited to denoting a class of physical objects (there are, for example, *parsing trees* in grammar and *genealogical trees* in family history, and one can talk about *the trees in a picture*, where they are representations or images of physical objects, rather than physical objects in themselves), while the criteria for defining even those trees that are physical objects are far from absolute. For example, one might start by define a tree as being a plant that has a trunk and branches that consist of wood, but then there are counterexamples like *banana trees*, where the trunk or stem is not woody. If we are tempted to postulate that bearing leaves is a necessary condition for being a tree, not only must we extend the definition of ***leaf*** to encompass the needles of pine trees and other conifers, but we must also deal with the leaflessness of deciduous trees in winter and note the existence of (admittedly rare) species of so-called leafless trees, specifically the ***belah*** and ***buloke***

trees of southern Australia. As soon as a necessary condition for the meaning of a word is proposed, it seems, it has to be modified to accommodate counterexamples.

To take another example, verbs typically denote events in the world – the verb ***climb*** is an example – but not all uses of the verb ***climb*** denote physical events – the value of stocks and shares may be said to ***climb*** to new levels, but this is an abstract events rather than a physical one – while in a sentence such as "The road climbs to the top of the Downs", the verb does not denote an event at all: it denotes a state. A similar situation pertains with regard to adjectives: qualities such as ***good*** and ***big*** denote, not entities or events in the world, but properties of entities and events, and very often the nature of the properties varies greatly according to the kind of entity or event denoted by the noun that the adjective modifies.

Words that represent entities (both physical and abstract), events, and states of affairs in the world – including the mental and emotional world of human beings and the various properties associated with each of these – are called 'content words'. Insofar as content words represent classes of entities in the world, it seems that they do so only loosely. Each content word has a central or focal meaning – sometimes more than one – and each focal meaning is associated with a cluster of typical contexts in which it is used (both other words and situations). In addition, each content word is available for a variety of other, less central uses. Someone learning to use a language to talk about the world (regardless of whether it is their first language or a second language) has to learn not one but two types of competence: the competence to use words normally, which is rule-governed, and the competence to exploit those norms of usage, which is also rule-governed. Lexical metaphors, ellipses, coercions, figures of speech, and other exploitations are all in a rule-governed relationship with focal meaning. Moreover, such exploitations may themselves be secondary conventions (as in the case of ***family trees*** and ***parse trees***) or they may be first-time, one-off coinages (as in the case of someone referring to an ornamental structure hanging from a living-room ceiling as a 'tree', unaware that the conventional term for it in English is ***a mobile***).

At the present time, lexical exploitation rules have not been acknowledged for what they are – second-order regularities governing the dynamic use of words as well as the creation of secondary conventional meanings of a word. They have not been systematically studied and are insufficiently well understood. They are, however, of central importance for understanding the role of the lexicon in language. It is not the case that "anything goes": there is a difference between a dynamic, meaningful exploitation of a norm, such as Dylan Thomas's use of ***ago*** in the phrase 'a grief ago' – and a mistake. Exploitations are deliberate irregularities; mistakes are accidental irregularities.

If this were all, it might be argued, broadly, that the role of words is to denote entities and events in the world, physical and abstract, and the

properties of such entities and events. This is certainly a large part of the role of words, but it is not the whole story. There are other important classes of words which have no denotation in the world at all. Examples include modal and auxiliary verbs, determiners, and conjunctions (what entity or event is denoted by ***may*** or ***shall***; ***the*** or ***his***; ***and***, ***but***, or ***or***?), discourse organizers (***however***, ***anyway***, ***besides***), and sentence adverbs (***broadly***, ***unfortunately***). These are called 'function words'. The meaning of function words does not relate to the world, but to the language, and does not vary or get exploited in the same way as the meaning of content words.

Clearly, there is a lot more to words than the denotation of objects, events, and states in the world. Nevertheless, a very large proportion of word use is devoted to talking about objects, events, and states in the world. Words – individual lexical items – function as interface points between the world about us (as it is perceived by humans as language users) and the system or network of beliefs that all human being have in their heads. At the same time, the way words interact with each other is a defining characteristic of language. These two interfaces (words and the world, words and other words) are the central concerns of lexicology. The first of them is one of the central concerns of the philosophy of natural language; the second is a central concern of lexical semantics. A third interface is that between different users of the same language.

4.2 How words relate to one another

As mentioned above, words relate to one another both paradigmatically, by being members of the same semantic set, and syntagmatically, by being in the same sentence or paragraph.

4.2.1 Syntagmatic relationships

Syntagmatic relationships are called **collocations**, and the words in a collocation are called **collocates** of each other. Theoretically, any co-occurrence of two words in reasonably close proximity may be regarded as a collocation, but in practice the term is generally restricted to terms that co-occur frequently – i.e. more often than chance would lead one to expect – and collocation is now often defined as being co-occurrence within a text window of five words either side of the node word.

Collocational relationships are in some cases governed by syntax and in others not. If the relationship is a syntactically governed one, the term used for it by J. R. Firth and his followers is **colligation**. Some writers regard collocations and colligations as mutually exclusive sets, but that distinction is hard to sustain. For all practical purposes colligations can be regarded as a subclass of collocations. The terminology is complicated by some American computational linguists who refer to colligations as 'true collocations'

and to collocations as 'co-occurrences'. In addition to the major works of Sinclair and Gross mentioned above, a central text for the study of psycholinguistic aspects of collocation in the light of corpus evidence is Hoey's *Lexical Priming* (2005), which argues that from earliest infancy humans build up a store in their brains not just of words as isolated building blocks but of words as nodes complete with a network of collocational and colligational actualities and potentials.

The question whether to say '*forbid someone to speak*' or '*forbid someone from speaking*' is a syntagmatic question about colligation. So are questions about the relationship between the lexical items ***slice***, ***bread***, and ***butter*** in the expression '*a slice of bread and butter*' as contrasted with, say, '*a slice of bread with some butter on it*'. Other collocations may not be in a syntactic relationship with each other at all. For example, the word ***doctor*** is a frequent collocate of ***hospital***, ***disease***, ***illness***, ***antibiotics***, ***pharmacist***, ***treat***, and ***cure***, but there is no fixed syntactic relationship in which these collocations are normally found. The pairs of collocates may be more or less anywhere in a sentence in relation to each other, or even in different but adjacent sentences.

Since the advent of large electronic text corpora in the late 1980s, a variety of computational tools have been introduced to measure the statistical significance of collocations and to discover which words actually do collocate with which other words. The statistical tools used with greatest success used to measure interesting collocations in corpora are **mutual information** (MI score) and **t-score**. The general principle of both is the same – to compare the actual, observable frequency of co-occurrence of pairs of words in a vast collection of texts with the frequency that one would expect through chance co-occurrence of the two collocates, given their respective frequency in the corpus independently of each other. Each statistical tool has its usefulness for different purposes. MI score tends to favour lower-frequency words, so it is particularly valuable to the lexical analyst as a reminder of the collocations in which content words participate. For example, in the British National Corpus the verbs ***prescribe*** and ***advise*** get a high MI score in relation to the noun ***doctor***. Even though these verbs are not very high-frequency words in themselves, they are strongly associated with the noun ***doctor***. The analyst's intuition confirms the importance of the association measured by the computer. (The downside of MI is that it may also favour rare but unimportant collocates, so a certain amount of selection by the analyst is called for.) T-score, on the other hand, favours high-frequency words in the way in which it selects collocates. It is therefore useful for picking out function words and determining idiomaticity of colligations, e.g. identifying the particles and prepositions that are most normally governed by a particular verb or noun. This makes it an attractive tool for foreign learners and course-book writers.

A tool for measuring colligations in corpora is the **Word Sketch Engine** of Adam Kilgarriff, Pavel Rychlý, David Tugwell, and others. This is a

language-independent tool, i.e. it can be applied to any corpus in any language, provided that the right pre-processing steps (principally word-class tagging and lemmatizing) have been carried out. It uses a modified version of MI score (called **salience score**) to measure the comparative significance or salience of lemmatized collocates in a corpus, in a pre-defined set of syntagmatic relations (e.g. subject – verb, verb – object, preposition – prepositional object, adjective – noun, verb – adverb, and so on). The number of syntagmatic relations measured varies from language to language.

4.2.2 Paradigmatic relationships

Syntagmatic relationships describe ways in which words go together to make meanings. Paradigmatic relationships, on the other hand, are about choices and contrasts. Even synonymy may be regarded as a contrastive relationship: a speaker must **choose** whether to say "The shop will close in 15 minutes" or "The shop will shut in 15 minutes." This section offers the reader a brief summary of the main types of paradigmatic lexical relationship. A much fuller account of paradigmatic lexical relationships can be found in Cruse (1986) and (2000).

Synonymy

Synonyms are words that have approximately the same meaning, for example ***tall***, ***high***, ***towering***, ***soaring***, ***lofty***. It is sometimes argued that there are no true synonyms. Even words that are quite close in meaning differ in some respect or other, for example register (e.g. ***lofty*** and ***soaring*** are more literary than ***high***), context (mountains are ***high***, people are ***tall***). Even the closest synonyms (and they do not get much closer than English ***close*** and ***shut***) may have subtle differences of meaning or emphasis. For example, collocational evidence suggests that ***close*** has an essentially extrinsic focus (you ***close*** a road, a library, or a shop typically in order to prevent people going in), whereas ***shut*** has an essentially intrinsic focus (if you ***shut*** a shop or your mouth, the focus tends to be on the fact that things can no longer pass from the inside to the outside). Of course, it hardly needs to be said that these 'focuses' are no more than very faint preferential tendencies. For all practical purposes, the essential meaning of the two terms is identical.

Hyponymy and Hyperonymy

Hyponymy is the relation between a word and the semantic hierarchy into which it fits. Thus, ***hammer*** is a **hyponym** of ***tool***, which is a hyponym of ***artefact***, which in turn is a hyponym of ***physical object***. There are various kinds of hammer, too, some of which (e.g. ***mallet*** and ***sledgehammer***) are denoted by separate terms. ***Hammer*** is a **hyperonym** of ***mallet***, ***sledgehammer***,

etc. The term **hyperonym** is preferred to **hypernym** since 'hypernym' and 'hyponym' are pronounced identically in British English. Alternative terms are **superordinate** (= hyperonym) and **subordinate** (= hyponym). See section 4.3. below for further discussion of semantic hierarchies.

A further important semantic relation is that between co-hyponyms, e.g. ***hammer*** and ***saw*** are co-hyponyms of ***tool***; ***hawk***, ***sparrow***, and ***penguin*** are co-hyponyms of ***bird***.

Qualia

Pustejovsky (1995) revived the Aristotelian notion of qualia and argued that qualia are generally sufficient for the expression of polysemy, and that it is therefore unnecessary, not to say impossible, to enumerate all the different facets of a word's meaning as separate lexical entries. Qualia are properties associated with the meaning of a concept. Pustejovsky distinguishes four qualia:

CONSTITUTIVE: the relation between an object and its constituent parts
FORMAL: that which distinguishes an object within a larger domain
TELIC: the purpose and function of the object
AGENTIVE: factors involved in the origin or "bringing about" of something

Lexical databases such as WordNet have tended to focus on the constitutive (hyponyms) and the formal (hyperonym) properties of words, while traditional dictionaries have tended to focus on the agentive. But it will be readily seen that for certain classes of entities the telic is also an essential property. Thus, it is not sufficient to express the meaning of the noun ***table*** in terms only of its formal (a piece of furniture), its constitutive (flat top, legs, etc.), and its agentive (made of wood, plastic, metal, steel, or some other material). An essential thing to say about a table is that it is for putting things on (its telic), just as an essential thing to say about a ***chair*** is that it is for sitting on.

Antonymy and Opposites

Informally, people often talk about one word as being the 'opposite' of another, but in fact antonymy is not a single lexical relationship all. Cruse (2000) distinguishes the following classes of antonyms:

Polar antonyms (e.g. ***long/short***, ***fast/slow***)
Equipollent antonyms (e.g. ***hot/cold***)
Overlapping antonyms (e.g. ***good/bad***)
Directional reversives (e.g. ***up/down***, ***forwards/backwards***, ***advance/retreat***)
Converses (e.g. ***above/below***, ***doctor/patient***)

The main difference between **polar antonyms** and **equipollent antonyms** lies in whether or not both terms are 'unmarked' (see next section). The main difference between **reversives** and **converses** is that reversives imply movement, whereas converses can denote pairs of static entities or states of affairs.

Marked and unmarked

An important distinction in lexicology, mentioned in the preceding paragraph, is that between marked and unmarked terms. An unmarked term is the general word for something, whereas a marked term denotes a specific subset or contrasting set of items. Thus, ***duck*** is an unmarked term, in that it can denote any bird of the right species, without reference to whether it is male or female. ***Drake***, on the other hand, is a marked term: it denotes only male birds of these species. In pairs of polar antonyms, one member of the pair is marked; the other is unmarked. For example, ***long*** and ***fast*** are unmarked, while ***short*** and ***slow*** are marked. This is because the question ***How long/fast is it?*** does not imply that whatever it is is in itself necessarily long or fast. By contrast, the question ***How short/slow is it?*** implies that whatever it is is in itself slow or fast. The question ***How slow is your car?*** implies that the speaker believes that you have a slow car. Equipollent antonyms are both marked: the question ***How hot is it?*** implies that whatever it is is in itself hot, and the question ***How cold is it?*** implies that whatever it is is in itself cold. Distinctions between marked and unmarked items play an important and sometimes neglected part in lexical analysis.

Meronymy

Meronomy is the technical term for part-whole relations, e.g. ***hand*** and ***finger*** or ***tree*** and ***branch***. The term meronymy embraces a large number of complex relations, which cannot be gone into here. A clear exposition will be found in Cruse (1986).

Metonymy

Metonymy is the practice of referring to something by using an associated word, e.g. ***glass*** to denote a beverage. This practice is common and widespread, probably in all languages. Until it is pointed out, we notice nothing unusual is someone saying, "He drank three glasses". But of course it was not the glasses themselves that he drank, it was the liquid in them. Humans don't have wings, so if I say, "We flew to San Francisco last summer", my listeners automatically know that it was a plane that did the flying; in point of fact, we just sat in it. Metonymy is a common source of nickname and surname formation in European languages; ***Mac the Knife*** is not literally a knife; he is a man known for carrying and using a knife. In medieval Europe,

people were known by nicknames such as ***Tom Hand*** and ***Peter Leg*** because of some noticeable peculiarity of the hand or leg. In due course some of these metonymic nicknames became established as hereditary surnames.

4.3 The semantic structure of the lexicon

From what has been said in the preceding section, it will be clear that the meanings of the items in the lexicon of a language are related to one another in a variety of different ways. This has led people naturally to infer that the lexicon of any language is structured, not merely a random collection of words denoting random entities and events in a random way. This is undoubtedly true – up to a point. However, it is not the case that all words in a language fit systematically into a neat and orderly structure such as a hierarchical ontology of the kind envisaged by Wilkins, Roget, and George Miller (WordNet).

A brief comment is appropriate here on the term 'ontology'. The word derives ultimately from the present participle of the Greek verb meaning 'to be' and in philosophical parlance it denotes Aristotle's conception of the subject of scientific inquiry – *to on* 'that which is', i.e. everything that exists. In modern computational linguistics, however, it is a buzz word used to refer to the entire vocabulary of a language (or at least, all the content words) and very often carries with it a number of unthinking and undefended assumptions, for example that all content words are in a hierarchical relationship – a so-called IS-A hierarchy, such that a ***canary*** is a ***finch*** is a ***bird*** is a ***living being*** is a ***physical object*** is an ***entity***[4]. Some of the problems for defining lexical structure in terms of a hierarchical ontology are identified below.

A word may have more than one meaning, which means that it may go into more than one place in the ontology. Satisfactory simple criteria for deciding which meaning is active in any utterance in which the word is used have not yet been developed, and it may be that such criteria cannot be developed for principled reasons. Different senses of polysemous words are not always mutually exclusive: there is often some semantic overlap. For example, the notion of *consuming a biscuit* (eating it) overlaps with *consuming groceries* (buying them in order to eat them), which in turn overlaps with *consuming durable goods* (buying them in order to keep them and use them), which overlaps with *consuming resources* such as fuel (using them up so that they are no longer available). Eating a biscuit is a very different activity from using oil to fire the central heating, and yet this example shows that, as far as the word ***consume*** is concerned, there are no hard-and-fast dividing lines in the intermediate steps between these semantic extremes of its meaning potential.

There are also lexical gaps (in any natural language) in any ontological model, which have to be filled by phrasal circumlocutions, or by invented words. The word ***mammal*** was invented in the 18th century to fill just such

a gap in the hierarchy of living beings. The term ***living being*** itself is a phrasal circumlocution for a lexical gap in the IS-A hierarchy of English words. Another example of a lexical gap is that in Russian and German there is no single word for 'to go' – speakers of these languages have to commit themselves to the *manner* of motion (walking, running, creeping, driving, riding, etc.) before they can talk about going somewhere. In fact, natural languages are full of lexical gaps, which rarely trouble everyday users of those languages, but become problematic when one tries to build a hierarchy or map the terms of one language onto those of another. Hierarchical ontologists sometimes respond to this problem by inventing artificial terms or forcing natural-language terms into a Procrustean bed and stipulating the meaning that they will have in the hierarchy. (Procrustes was a character in Greek mythology who ensured that his guests exactly fitted the bed he gave them, either by chopping off their feet if they were too tall or by stretching them on the rack if they were too short.)

In the jargon of Semantic Web research, 'ontology' has acquired yet another sense, namely an inventory of all the tagged items – names, addresses, dates, documents, etc. – that are available for computational processing.

Synonyms sometimes 'compete' for the same slot, both in an ontological hierarchy and in usage. For example, the verb ***suppose*** competes with ***imagine***. A speaker has to chose whether to say 'I suppose that . . .' or 'I imagine that . . .'. WordNet solves this particular problem neatly by placing both or all relevant terms in a synonym set within an IS-A hierarchy and by allowing words to be members of more than one synonym set. But this solution brings other problems in its wake. One such problem is that the occurrence of a word in different synonym sets is sometimes equated with its having two different meanings, but this cannot be right. The verb ***think***, for example, is placed by WordNet in one synonym set alongside {***believe***, ***consider***, ***conceive***}, in another with {***opine***, ***suppose***, ***imagine***, ***reckon***, ***guess***}, and in a third with {***cogitate***, ***cerebrate***}. ***Think*** occurs in several more synonym sets, too. It would be hard to defend the position that these are all separate meanings of ***think***. It seem preferable to regard them as closely related facets of a single meaning.

A distinction must also be made between individuals and words. As a general rule, words denote classes of individuals. Both classes and individuals may be members of more than one semantic class. And in any realistic ontology there must be **multiple semantic inheritance** or a **tangled hierarchy**. For example, a ***cat*** is not only a ***mammal*** (i.e. it inherits any properties that are attributed to the term ***mammal***) and in most cases a ***pet*** (i.e. it inherits any properties that are attributed to the term ***pet***), but also a ***carnivore*** and a ***tree-climber*** (amongst other things). ***Cats*** and ***dogs*** are pets as well as mammals, whereas ***parrots*** and ***budgerigars*** may also be pets even though they are not mammals. So the term ***pet*** cannot be neatly slotted into an IS-A hierarchy of animate entities. And on the other hand ***lions*** and

hyenas are indisputably ***mammals*** – but not (normally) ***pets***. So being a pet is a property of certain individuals and classes and not of others within the hierarchy of animate entities.

A similar problem arises with regard to individuals: an individual called Joe may be a doctor by day, a saxophonist by night, and an oarsman at weekends. ***Doctor***, ***saxophonist***, and ***oarsman*** are co-hyponyms of ***human***, but this does not mean that Joe is three different individuals. One possibility is to deal with this problem either by saying that being a pet or a saxophonist is a property of certain classes of animals or humans, like having fur or barking, or by postulating a separate ontology of pets.

Attempts to identify the place of a word in a semantic hierarchy by observing its syntagmatic behaviour run into equally serious problems, for example because of the fact that different entities do different things on different occasions. Humans share some behaviours (***eating***, ***sleeping***, ***moving***) with animals and other behaviours (***planning***, ***negotiating***, ***announcing***) with human groups and institutions. There is a widespread belief that a well-constructed hierarchical ontology will help to disambiguate words. Up to a point this is true: ***firing a gun*** is not the same as ***firing an employee***, so the idea is (or was) that if we could only identify all the lexical items that count as gun-like things and all the lexical items that count as employee-like things, we could use these classes of nouns as a lever to disambiguate these two meanings of the verb fire. But even with such a simple and distinctive example, problems arise: for example, ***Big Bertha*** may denote a person, but was also actually the name of a very large gun used in the First World War. For most words, the distinctive criteria between meanings are not clear-cut.

All this points to the conclusion that hierarchical ontologies may not be very useful for attributing meanings to words. Part of the reason for this, if it is true, is that such ontologies are a summation of the lexical reflexes of the 2,500-year-old European endeavour to construct an organized scientific conceptualization of the universe. This is not the same as investigating how people use words to make meanings. Why ever did anyone think it would be?

A more promising approach may be, rather than attempting to construct one gigantic, monolithic IS-A hierarchy, to group words into classes according to their collocational behaviour. Some local hierarchies may emerge in an empirically well-founded ontology constructed on such a basis, but overall the picture that emerges is likely to be much more tangled and messy. See the chapters in Volume VI for some approaches to this much debated issue.

4.4 Words and concepts: historical development

Philosophers from earliest time have been aware of the vague and relative nature of word meaning. Aristotle, for one, observed that some terms are applied with different but related senses in different contexts. For example, ***healthy*** can be applied to different nouns – a person, a food, a diet, an activity

or action (e.g. *healthy exercise, a healthy workout*), a quality (somebody may be said to have *a healthy colour*), and so on. No single definition of ***healthy*** will fit all these uses, but in Aristotle's view, they are all related to each other, and (more importantly) to a central, primary use. The meaning of a term in its central, primary use has been called the 'focal meaning'. Thus, in the Aristotelian view, the focal meaning of ***healthy*** relates to a human being; other uses are related to or derived from this. Aristotle does not seem to have regarded it as necessary or, indeed, possible, to state necessary conditions that would account for all correct uses of a word like ***healthy***.

Throughout Classical Antiquity and the Middle Ages, little was added to what Aristotle had said about words and meanings. Indeed, Aristotle's remarks on contextual differences were largely ignored, whereas his remarks on essential vs. accidental properties of concepts and on genus and differentiae were studied and developed assiduously and fitted into the development of medieval logic. Perhaps the only development worthy of note for lexicology during this long period was the development of the Latin word ***definitio*** 'definition'. The Roman lawyer and politician Cicero (106–43 B.C.) is largely responsible for popularizing the notion of ***definition*** (as opposed, say, to the ***explanation*** or ***description*** of a concept). The Latin verb ***definire*** 'to define' literally means 'to set boundaries'. It should be noted in passing, however, that Cicero uses the verb ***definio*** in the sense of deliberately stipulating the boundaries of word meaning, not in the sense of determining them through empirical observation of language in use.

The focus in later Antiquity and in the Middle Ages, as far as the study of language and concepts is concerned, was on developing a logic that would distinguish true propositions from false ones and on determining relationships among conjoined propositions. Words and their meanings excited comparatively little interest, largely because a word in isolation is neither true nor false. (***Snow*** is neither true nor false. 'There is snow on the ground' is either true or false, depending on the circumstances in which it uttered; 'Snow is white' is (allegedly) always true.) Thus, in a venerable tradition, the semantics of concepts came to be associated with propositional logic and truth conditions. To this day, this is a very active area of academic activity. For discussion of truth-conditional and logical semantics, which are parallel subjects to lexicology, see *Critical Concepts: Semantics*, in particular Part 1 of Volume I.

A principle in logical semantics and associated truth-conditional theories of meaning is the Principle of Compositionality, attributed to Gottlob Frege. This states, broadly, that the meaning of a sentence is determined by the rule-governed way in which the constituents combine. The Principle of Compositionality does not say anything about the meaning of the constituents. A traditional view is that the meaning of constituent parts of sentences is the subject matter of lexicology, while the rules governing the way in which the constituents combine is the subject matter of logical and

truth-conditional semantics. A problem for this view raised by some of the chapters in this collection (by Sinclair, Hoey, Gross, and Mel'čuk, for example) is that the lexicon now threatens to spill over into the grammar: each word carries its own principles of combination. But of course this view of compositional semantics, while saying much about idiomaticity, says nothing about the truth of the resultant propositions. The relationships among words, concepts, compositionality, and truth are due for revision in the light of recent work on both sides of the lexicon/composition divide.

It was not until the European Enlightenment of the 17th and 18th centuries that people began to think seriously again about the relationship between words, concepts, and the world. At that time, thinkers such as Wilkins and Leibniz (among others) made it a major concern to try to remedy what they perceived as the deficiencies of imprecision and vagueness in natural languages, by inventing a perfect language in which concepts would be precisely defined. A hierarchy of concepts would be represented, one for one, by a hierarchy of precisely defined words. Imprecision would be eliminated. This move was influential in the subsequent development of logic and was taken as axiomatic until challenged by Wittgenstein and others in the mid 20th century, but it is based on a confusion – what Gilbert Ryle called a category mistake. Logic and language are two different things. The rules that govern the functioning of natural language – how people use words to make meanings – are not the same as the logic needed for argumen-tation about scientific discovery. As Hjelmslev pointed out in the 1950s, a separation between natural language (vague, flexible) and logic (precise, inflexible) is necessary for the understanding of natural language.

At a different level, the work of Wilkins and Leibniz has also been influential in the creation of practical language tools such as thesauruses and dictionaries. As far as lexicography is concerned, a major influence on definition writing was Leibniz's principle of identity – that two expressions are identical in meaning if one can be substituted for the other *salva veritate* (without affecting the truth). Leibniz followed Wilkins in planning a perfect language that would eliminate vagueness and uncertainty. In addition to creating a 'perfect language' (with a script of its own), Wilkins had also set about organizing a hierarchy of concepts – such that a hawk is a bird and a bird is a living being and a living being is a physical object and so on. This ontological hierarchy (see Chapter 3, Volume I, and Borges' gentle parody in Volume I, Chapter 14) was the direct precursor both of Roget's Thesaurus and of WordNet. Although, as we have noticed, there are difficulties in applying a hierarchical ontology to all words, the basic approach is sufficiently plausible for both Roget and WordNet to be immensely popular, not as historical curiosities or theoretical possibilities, but as popular practical tools for everyday language users. The possibility of a hierarchical structure of concepts is a theme that starts in Volume I of the present collection and continues in various guises through to Volume VI.

It was left to Ogden and Richards and the 20th-century European structuralists – followers of Ferdinand de Saussure, in particular Hjelmslev and Coseriu – to make the at first sight surprising suggestion that the structure of concepts in a language and the relationships of the words of a language are independent variables. Synchronically, this means that a lexical element may fit simultaneously into several different conceptual structures (a tree can be both a physical entity in the domain of plants and an abstract entity in the domain of grammar), while diachronically it means that the conceptual structures of a language may change independently both of the world and of the words used. Coseriu, in Volume II, Chapter 25, adduces evidence from Latin and the Romance languages in support of the claim that the conceptual structure of a language may vary independently from the words that express the concepts. Alternatively, the scope of a word may become larger or smaller or different, while it remains essentially the same word. Coseriu's example is Latin *passer*, which in Latin meant 'sparrow' but in modern Spanish (*pajero*) has become generalized to mean 'any small bird'. At the same time, the scope of the Latin word *avis* has been cut down in Spanish (***ave***) to denote only large birds. Given the regular rules of phonological change, the words are the same, but the conceptual structure of the two languages (Latin and Spanish) is different.

4.5 Formalizing word meaning

The formalization of word meaning has been an important issue ever since the 1960s, when computers began to be used for processing text as well as for number crunching. It is not possible to compute anything that does not have a formal structure. Early optimism, represented, for example, by the chapters in Volume VI by Michael Lesk and Harold Robison, gave way for a time to disappointment, which itself is now being replaced by renewed optimism as statistical methods begin to bite. The general approach of those early papers has not by any means been invalidated, but the details of actual language use turn out to be much more complex than people originally expected.

A seminal early article by Lesk, subtitled 'How to tell a pine cone from an ice-cream cone' (1986; Volume VI, Chapter 83 in this collection), showed how, if the word ***cone*** occurs in a text that also contains *pine, evergreen, tree*, and/or *needle*, a computer can correctly select the meaning 'fruit of a tree' (rather than 'wafer container for an ice cream' or 'geometric figure') by consulting a machine-readable dictionary. By the same process, the correct meaning of ***pine*** ('evergreen tree', rather than 'waste away through sorrow or illness') is likewise selected. Lesk's system did not always get the right meaning. Despite this, Lesk's comment, "The problem of deciding which sense of a word was intended by the writer is an important problem in information retrieval systems" remains fundamentally true today. Impressive results have been achieved in information retrieval by variations on string

matching (without reference to any supposed meaning of the string), but no natural-language understanding system (or artificial intelligence system) can be properly so called if it has not solved the problem of the relation between word use and word meaning, which at the present time remains still unsolved. At least part of the problem is that specific contextual clues as to the correct meaning of a lexical item do not occur as regularly in ordinary texts as they do in the dictionaries that Lesk was using. Another problem is deciding in advance what is a relevant clue. Lesk's idea was that relevant clues might match words found in the relevant dictionary definition, but there is all too often little overlap between everyday communicative texts and the text of a dictionary.

The second chapter by Lesk included here (Volume VI, Chapter 84, 'They said true things but called them by Wrong Names') was inspired by a quotation:

> They . . . [shall] not perceive the train,
> Till in the engine they are caught and slain.

This seems to modern readers to make perfectly good sense (no doubt something to do with a fatal railway accident) – until they are told that it was written by Ben Jonson in the 17th century (Lesk, personal communication). This forces a recognition that the 'train' in question cannot be a railway train, and the engine, whatever it is (probably a military siege engine), is not a locomotive.

Among the most ambitious attempts to formalize word meaning is the work of Igor Mel'čuk, represented here in section 1 of Volume VI. Among other things, Mel'čuk proposes that a set of lexical functions governing the way in which words relate to one another and what they mean. The meaning of each word can be expressed as a set of formal functions governing its relationship and interaction with other words. He also proposes that word meanings can be expressed as sets of necessary and sufficient conditions.

Some other lexical resources used by the natural language processing community are, unfortunately, built on the introspections of theoretical linguists, proposing formal structures with no empirical validity – i.e. structures that might theoretically exist but (as far as can be discovered from the empirical evidence of actual usage) probably do not. These resources and theories, although fashionable in some circles, are not represented in this collection.

An equally ambitious programme is FrameNet, a lexical resource based on the frame semantics theory of Charles Fillmore. The aim is to discover and document the range of semantic and syntactic combinatory possibilities of each word in each of its senses. Fillmore postulates a number of basic conceptual structures that underlie the meanings of individual words, for example the conceptual structures of commercial transactions or of movement or perception or any of a very large number of other activities. These

conceptual structures are called **frames**. Each of the semantic components or **frame elements** used in them (e.g. ***buy***, ***sell***, ***buyer***, ***seller***, ***goods***, ***price***) has meaning by virtue of its relationship to the other frame elements in the same frame and may be realized by any of a potentially very large number of actual lexical items. It is impossible in a short summary such as this to do justice to the richness of the theory of frame semantics; the reader is referred to the FrameNet website, http://framenet.icsi.berkeley.edu/ and the publications listed there. The papers selected for this collection in Volume IV, section 3, are intended to give some insight into the motivation for building frames rather than separate lexical entries.

4.6 Lexical acquisition by very young children

A topic of absorbing interest is, how do children acquire words in the first place? The first section of Volume V of this collection contains four classic articles on child language acquisition. Roger Brown's 1958 paper established the general tendency of children to learn more concrete terms before abstract terms (rather than vice versa, as some theorists then held). Eve V. Clark's 1973 paper proposed a semantic feature hypothesis, according to which children gradually enrich the meaning of the terms that they acquire by adding more and more semantic information. Clark also drew attention to 20th-century research in many languages establishing that children characteristically overextend the meaning of the words that they acquire, so that, for example, a word meaning 'animal' may be overextended to denote all things that move. A person (child or foreign language learner) who has only a very small vocabulary will cheerfully use the words in that vocabulary to fill the semantic and pragmatic space that he or she needs, regardless of the usage of words by adult native speakers of the language. This is only surprising if one believes that there is a necessary connection between a word and its meaning. From the speaker's point of view, any strategy that achieves communicative success is acceptable. Correctness of denotation is a secondary matter. Clark (1997) develops the theme of the child gradually refining the meaning of his/her terms in the light of guidance from the innumerable pragmatic signals (such as the direction in which someone is looking) that he or she receives from the adults and other children with whom he/she is interacting, as well as other features of the environment. Acquiring the meaning of words is a whole-world activity, not a narrowly linguistic one. For a fuller account of Clark's observations and theory of language acquisition, see Clark (1993).

4.7 Lexicology and translation

To the question, "How do the words of one language map onto those of another?" there is a very simple answer: "They don't". Of course, there are

some simple correspondences: French ***arbre*** corresponds in its normal, most literal uses to English ***tree***, German ***Baum***, and Romanian ***copac***. But even at this basic, concrete level of the lexicon, there are already incongruencies: in Romanian, for example, a fruit tree is ***pom***, while a genealogical tree is ***arbore*** (see Coseriu, Volume II, Chapter 25).

Even 'the same' word can have different scope and implications. For example, the Czech word ***aristokracie*** has a slightly wide scope than the cognate English word ***aristocracy***. It includes the princes of the Church (Roman Catholic cardinals), whereas the English word does not. German ***Nachbarschaft*** corresponds broadly to English ***neighbourhood***. Both words denote a physical location (*Er wohnt in dieser Nachbarschaft, he lives in this neighbourhood*). However, the German term also has an abstract sense (*in einer guten Nachbarschaft wohnen* 'to have good neighbourly relationships'), which the English term does not.

More complex phraseologies and concepts are even harder to map one-for-one onto the terms of another language. There is, for example, no English equivalent for the German verb ***leisten*** that is equally satisfactory in all contexts, and sometimes considerable circumlocutions are required to achieve a sentence with equivalent meaning. *Bitte leisten Sie den Anweisungen von Polizei Folge* and *Please follow the police instructions* mean the same thing, but the phraseology is different.

Translation and lexical equivalents across languages is a complex subject, reserved for a separate compilation in the *Critical Concepts* series. It is not included here. The essential point, as far as the study of the lexicon is concerned, is that the folk belief in one-for-one equivalences, which may serve well enough in basic pragmatic situations such as shopping and tourism, very soon breaks down when serious conversation is attempted. The lexicon of every language is *sui generis*, a thing in itself, with its own rules and constraints.

4.8 Etymology and word history

Lexicology is sometimes defined as including the development of word forms as well as the development of word meanings. In this collection, however, the development of word forms (historical morphology) is not a central theme. Insofar as etymology and word history are represented, the focus here is on changes in word meaning and semantic structures. In particular, scholars such as Jost Trier and Eugene Coseriu brought a structuralist perspective to the history of words and their meanings.

The general idea of Semantic Field Theory is that different languages divide up the world in different ways, and the terms used are determined by different criteria. At its simplest, this means that, for example, there is not necessarily just one word denoting things you sit on. In German, this particular semantic field is divided up among the lexical items ***Sessel*** and ***Stuhl*** (among others), as Gipper discusses in Volume II, Chapter 21.

To make matters even more complicated, the division of a semantic field among its lexical items is not stable, but has a tendency to change over time. Trier, in a famous paper (Chapter 20 in this collection) discusses how concepts in the field of personal and cultural development – terms denoting knowledge, art, and skill (***wisheit***, ***kunst***, ***list***) – developed in Middle High German, carving up this particular cognitive field in different ways at different points in time, and having different cognitive and cultural associations. A historical perspective such as this is found neither in the work of generative lexical theorists (Katz and Fodor, Jackendoff, Pustejovsky) nor in the work of Firthian systemic linguists (Halliday, Sinclair, Hoey, Winter).

4.9 Literal and figurative word meanings

Metaphor and Figurative Language is a separate collection in the *Critical Concepts* series, so little will be said about this subject here. For more detailed discussion of the issues, the reader is referred to those volumes.

The concept of metaphor is only meaningful if it is contrasted with the concept of literal meaning. Various criteria for the literal meaning of a word or lexical item have been proposed, including principally:

- **Historical priority**. The oldest meaning of a word is sometimes said to be the literal one. Problems with this include knowing where to draw the line. Can an obsolete meaning be literal? Some people would say so, but it is hard to defend the notion that the literal meaning of a familiar word like ***sock*** is 'a light shoe or slipper'. That, however, was the medieval meaning of the word, derived from Latin *soccus* 'a slipper worn by a comic actor'.

- **Frequency**. The verb ***backfire*** is much more often used of plans and actions going wrong than of internal-combustion engines making a loud explosions or guns blowing up in the user's face. But it does not seem wrong to say that the abstract meaning (e.g. *his plan backfired*) is a metaphor based on the literal meaning of some mechanical device (a car engine or a gun) misbehaving.

- **Concreteness**. The example of ***backfire*** suggest that, if a word has two meanings, one of which denotes a concrete entity or event, then the concrete one is likely to be regarded as the literal meaning.

- **Phraseological constraints**. A linguistic metaphor is generally much more constrained than a literal meaning. For example, in their metaphorical senses words such as ***torrent***, ***river***, and ***oasis***, are typically followed by ***of*** (*a torrent of abuse, a river of blood, an oasis of calm*) or used in one of a small number of other phraseological constructions, while the literal meanings are much less constrained.

- **Absence of resonance**. Finally, it should be noted that, cognitively, linguistic metaphors have the potential to 'resonate' in the mind of the hearer or reader with a literal sense, but not vice versa. An *oasis of calm in the centre of a city* resonates with the folk notion of peaceful oases with palm trees, water holes, and possibly a few Bedouin and their camels, whereas the notion of an oasis in the Sahara Desert does not resonate with the notion of a place in a city centre.

From the point of view of lexicology, the most important thing to note about literal and metaphorical meaning is that all content words have the potential to be used in new and unusual ways – including metaphors, similes, and other figures of speech. This is not true of function words. The use of function words such as auxiliary verbs, modal verbs, conjunctions, and determiners is governed by logical and grammatical rules that are rarely if ever broken.

Thus, the content word ***lever*** literally denotes a rigid bar, resting on a pivot, used to lift a heavy object or to operate a mechanism. But this same word is regularly found in non-literal uses such as 1.

> *1 Using overseas aid as a* ***lever*** *to promote democracy in third-world countries*

Uses such as 1 are classified as figurative or metaphorical uses of literal concepts, even though they may be highly conventional. A figurative meaning is typically abstract (one cannot hold overseas aid in one's hands), whereas a literal meaning is typically concrete (a literal, physical lever is something that one *can* hold in one's hands). This fragment of text is classified as metaphorical because, in Max Black's terminology, the secondary subject (a lever as a means of coercing people to do something) **resonates** with the primary subject (a lever as a means of lifting or operating something). Criteria for distinguishing literal from figurative senses include not only concrete/abstract but also old/new (old-established senses are more literal than new senses derived from them), frequent/rare (metaphorical uses are often rarer than literal ones), and resonance (metaphorical senses resonate with.

From the point of view of empirical analysis of words and meanings, the distinction between literal and metaphorical word meaning is much less important than the distinction between conventional and creative uses of words. Many so-called metaphors (such as the use of *lever* in the example above) represent perfectly conventional secondary senses. Some scholars

(e.g. Sinclair) even deny the existence of metaphor, since on these grounds metaphors are indistinguishable from other secondary senses. It's all just a matter of how the words are used, and to what extent this is conventional.

Mira Ariel's article (Chapter 77, Volume V) on the demise of a unique concept of literal meaning provides a thoughtful and extended critique of some of the issues involved.

4.10 Lexical creativity

People acquiring a language need to learn not one but two kinds of competence in rule-governed linguistic behaviour: the competence to use the language normally, with the words in their most literal and conventional senses, and the competence to exploit those normal uses in order to say new and interesting things. Any study of dynamic metaphor, dynamic use of other figures of speech, ellipsis, and semantic coercion (such that *enjoying a book* coerces a different underlying semantic frame from *enjoying a meal*) is forced to the conclusion that there must be exploitation rules in languages as well as rules governing normal grammatically well-formed uses of words. Further evidence for the existence of of exploitation rules comes from the utterances of sophisticated foreign speakers who know a second language well, but who sometimes apply the exploitation rules of their native language to the second language, with non-idiomatic-sounding results – for example the English-speaking German who tells you that he lives "in a good neighbourhood" when he means that he enjoys good relations with his neighbours (see 4.7 above). At the present time, comparatively little work has been done on identifying the rules governing exploitation of norms, not least because the norms themselves are not yet well understood: the findings of corpus linguistics (see, for example Sinclair 2004) suggest the uncomfortable conclusion that some adjustment to standard linguistic theories about words in use is required if they are to fit the facts, and this indeed is what has been happening since the late 1990s. The focus must therefore still be on getting the right account of first-order regularities.

5. Structure of the collection

Volume I gives a selection of philosophical writings about the lexicon from Aristotle onwards. Aristotle established the ground rules, so to speak, for lexicology and the definition of concepts, which, with some clarification by Porphyry and other scholars of the Classical period, were accepted almost without question until the 20th century. Indeed, Aristotle's work is still profoundly influential in present-day thinking about the lexicon, as witness the work of, say, Wierzbicka and Pustejovsky (in other respects, very different writers). Unfortunately, Aristotle did not actually write any documents explicitly devoted to the lexicon, so what is offered here is the next

best thing: a selection of passages showing what Aristotle actually said about word meaning, selected, arranged, and edited by Katerina Stathi. Ironically (in view of his enormous influence on subsequent writers on the lexicon), word meaning was one of the few subjects in which Aristotle was not primarily interested. Rather, he regarded words – and language as a whole – as merely a vehicle for the logical organization of concepts for the study of the world, the universe, and the principles of human behaviour. His remarks on the lexicon are scattered widely throughout his writings. One major aspect of Aristotle's comments on the lexicon is not represented here, namely his writings on metaphor and figurative language. These will be found in a companion set to this collection: *Critical Concepts in Metaphor and Figurative Language*.

Volume I continues with some radical 20th century challenges to the Aristotelian tradition, in particular arguments by Wittgenstein and Putnam against necessary conditions for word meaning in natural language. Wittgenstein, too, is represented in this collection by a specially commissioned selection – a selection made, with a commentary, by Yorick Wilks. The two papers here on stereotypes and word meaning by Putnam (a philosopher of mathematics and language) may be compared with the work of Eleanor Rosch on prototypes in cognitive science and anthropology (see Chapter 70, Volume V). The volume concludes with a selection of papers on the systematic variability and vagueness that characterizes word meaning, which has been a focus of research in lexicology since the mid 20th century.

Volume II gives a selection of articles on lexicology from two major traditions in linguistics: European structuralism and American generative linguistics. Some important papers by European structuralists, in particular the German semantic field theorists of the 1920s and 1930s and the structuralist Romance philologist Coseriu, are here translated into English for the first time. Other papers, including those by Hjelmslev and Pottier, are reproduced in the original French. Kinship terms are often cited by structuralist linguists to demonstrate the role of componential semantic analysis, and Volume II, Part 6 includes two of the classic papers in this area, by Goodenough and Lounsbury.

The second part of Volume II turns to a very different kind of structuralism, namely work on the lexicon in the American generative tradition, starting with Katz and Fodor's formerly oft-cited 1964 article (an attempt extend the structures of generative theory into lexical semantics, using semantic "markers" and "distinguishers" of a recognizably Aristotelian kind). This work is set into perspective by Bolinger's insightful discussion of the theory in general and Bierwisch's account of the role of semantic primitives in German adjectives. The last section of Volume II contains Pustejovsky's introductory account of generative lexicon theory and Jackendoff's re-evaluation of the standard division between lexicon and grammar.

Volume III contains important essays elaborating themes that have already been touched on in Volumes I and II, and in particular addresses the issues of of semantic primitives and the question of polysemy in natural language. The concluding section of Volume III represents influential anthropological perspectives.

Another shift of focus comes with Volume IV. Up to now, most of the chapters have focused on the internal structure of word meaning, but now we look at the ways in which words are put together in sentences and how this affects meaning. This is currently a tremendously fruitful area of research, because new resources have become available in recent years (large machine-readable collections of text – corpora – and the Internet itself). Reading the papers of Halliday and Sinclair written in the 1960s, before the full potential of large electronic corpora was known, one cannot but be impressed by the prescience with which they foresaw the direction that empirical study of the lexicon would take.

The second section of Volume IV represents lexicon grammar, the theory developed by Maurice Gross, the general conception of which is that each word in the lexicon is associated with a rich set of grammatical preferences and constraints, rather than merely being a terminal node in a parse tree. The relationship between Gross's work and that of Sinclair and his associates makes an interesting comparison. Next comes a section devoted to the equally important theoretical work of Charles Fillmore on frame semantics. The volume closes with Wilks's account of his theory of preference semantics together with three other important papers by influential theorists on lexical preferences and context.

Volume V represents discussions of the mental lexicon. The first section is about how children acquire language, in particular how they learn what words mean. The second section contains some seminal papers on prototype theory and situation semantics, including Eleanor Rosch's famous work on the cognitive representation of semantic categories. Section 3 of this volume begins with Lawrence Barsalou's account of how people invent categories ad-hoc, even if there are no existing words to represent a required concept, and the volume concludes with some other important discussions of aspects of the mental lexicon.

Volume VI contains discussions of formal and computational approaches to the lexicon. The first section of this volume is devoted to Mel'čuk's lexicocentric formalisms, insisting on the combinations as well as the meanings of words, with examples from his explanatory and combinatorial dictionary. We then turn to the computational analysis of word associations in text, starting with the use of machine-readable dictionaries as a source of evidence and going on to the use of corpora and the statistical analysis of collocations in them. Section 3 gives an overview of some of the principal lexical resources for language processing. The collection concludes with four very different perspectives on computational representations of the lexicon.

Notes

1 This semantic change took place as a result of the cheating habits of medieval bakers, who risked being taken to court or '***sized***' if they gave short measure: a ***size loaf***, therefore, was one that was of a magnitude that would be acceptable to an assizes. Subsequently, the word was applied in its new sense to the dimensions not only of loaves but also of other products and entities. The idiom ***a baker's dozen*** (meaning thirteen, or one more than requested) had a similar origin, representing the medieval baker's insurance policy against being **sized**.

2 An exception to this rule of normative linguistic behaviour is poetry, where innovations are deliberate not accidental and (sometimes) may be admired rather than stigmatized.

3 The restricted defining vocabulary developed for LDOCE was, unfortunately, not matched by a restriction on the sense in which each word may be used. Since the most frequent words in a language are often also the most polysemous, this leads to a combinatorial explosion of theoretically possible meanings when these words are put together in a definition.

4 WordNet's IS-A hierarchy of hyperonyms of ***canary*** is in fact a good deal more complicated than this, including many more subtle steps.

Recommended further reading

One final duty remains for the writer of a general introduction, namely to recommend further reading. Many hundreds of books and articles on the various aspects of lexicology are published each year. Here, mention will be made only of works that give introductory guidance through the complexities of the field.

Probably the best available general introduction to the central issues of lexicology is Part 2 of Sir John Lyons' *Lexical Semantics* (Lyons, 1995). The same author's two-volume survey, *Semantics* (Lyons, 1977) is still an essential source of information, despite being somewhat dated, having been written 30 years ago. A readable and well-judged introduction to basic concepts of lexicology is Jackson and Zé Amvela (2000). Also worth reading is Lipka (2002). A thoroughgoing traditional account of lexical semantic relations is Cruse (1986). Finally, mention must be made of the massive two-volume collection of specially commissioned articles in German and English, Lexikologie/Lexicology, edited by Cruse and others (2002, 2005) in the de Gruyter HSK series, which contains many short, authoritative articles both on mainstream topics in lexicology and on numerous culturally specific matters.

References

Bloomfield, L. 1933. *Language*. New York: Holt, Rinehart, and Winston.
Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
Chomsky, N. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
Clark, Eve E. 1993. *The Lexicon in Acquisition*. Cambridge University Press.
Cruse, D. A. 1986. *Lexical Semantics*. Cambridge University Press.
Cruse, D. A. 2000. *Meaning in Language: an Introduction to Semantics and Pragmatics*. Oxford University Press.
Cruse, D. A., F. Hundsnurscher, M. Job, and P. R. Lutzeier (eds). 2002, 2005. *Lexikologie/Lexicology*. 2 volumes. Berlin and New York: de Gruyter.

Goddard, C. 1998. 'Bad arguments against semantic primitives' in *Theoretical Linguistics* 24: 2–3, pp. 129–156.
Jackendoff, R. 2002. *Foundations of Language*. Oxford University Press.
Jackson, H., and Zé Amvela, E. 2000. *Words, Meaning, and Vocabulary: An Introduction to Modern English Lexicology*. London: Continuum.
Hoey, M. 2005. *Lexical Priming. A New Theory of Words and Language*. London and New York: Routledge.
Lakoff, G. 1987. *Women, Fire, and Dangerous Things: what Categories reveal about the Mind*. University of Chicago Press.
Lipka, L. 2002. *English Lexicology: Lexical Structure, Word Semantics, and Word Formation*. Tübingen: G. Narr.
Lyons, J. 1977. *Semantics*. 2 volumes. Cambridge University Press.
Lyons, J. 1995. *Linguistic Semantics: An Introduction*. Cambridge University Press.
Mel'čuk, I. 2006. 'The Explanatory Combinatorial Dictionary' in Giandomenico Sica (ed.) *Open Problems in Linguistics and Lexicography*. Monza, Italy: Polimetrica.
Ogden, C. K., and I. A. Richards. 1923. *The Meaning of Meaning*. London: Routledge and Kegan Paul.
Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.
Quine, W. V. O. 1965. *Word and Object*. Cambridge, MA: MIT Press.
Rosch, E., and B. Lloyd. 1978, *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.
Sinclair, J. M. 2004. *Trust The Text: Language, Corpus and Discourse*. London: Routledge (Taylor and Francis).
Sweetser, Eve E. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press.
Wierzbicka, Anna. 1985. *Lexicography and Conceptual Analysis*. Ann Arbor: Karoma.
Wierzbicka, A., and C. Goddard (eds). 2002. *Meaning and Universal Grammar: Theory and Empirical Findings*. 2 volumes. Amsterdam and Philadelphia: John Benjamins.

Acknowledgements

My first debt of gratitude goes, of course, to the authors, living and dead, who have contributed chapters to this collection and who have contributed to the intellectual drama of lexicology across many centuries. I would particularly like to thank the many friends and colleagues who have helped me make the selection of papers in these volumes, both with comments on the selection as a whole as it gradually took shape, and by answering my innumerable questions and requests for opinions about specific issues. There were some lively debates. In the course of compilation, the collection grew from three to six volumes, but despite this many painful exclusions had to be made. To have included all the papers suggested would have required many additional volumes. The collection attempts to give a broad, interdisciplinary view of the whole field of lexicology. Nevertheless, the final selection is a personal one, and any eccentricities are of course my responsibility. In particular, I would like to thank (in alphabetical order of surname) Mira Ariel, Sue Atkins, Nick Braisby, Nicoletta Calzolari, Eve V. Clark, Ken Church, Christiane Fellbaum, Thierry Fontenelle, Dirk Geeraerts, Alexander Geyken, Gregory Grefenstette, Michael Hoey, Ray Jackendoff, Adam Kilgarriff, Ramesh Krishnamurthy, Michael Lesk, Birte Lönneker, Anke Lüdeling, Igor Mel'čuk, Karel Pala, Stephen Pulman, James Pustejovsky, Donald Rutherford, Michael Rundell, Katerina Stathi, Piek Vossen, Anna Wierzbicka, Yorick Wilks, and Timothy Williamson. To all of them, and to many others too numerous to name individually, my grateful thanks.

It has been a particular delight to work with Simon Alexander of Routledge (Taylor and Francis) as the publisher of this collection, and I am glad to have this opportunity to acknowledge his unfailing patience, courtesy, good humour, and professionalism.

Above all, my gratitude is due to Anne Urbschat, who began work on the project as an administrative assistant in 2003 and who has put in far more time and effort than was envisioned at the outset. She has not only handled all the rights and permissions requests, dealing efficiently with the different needs of different publishers' rights departments and author's special requests; she has also organized the work flow for the project and made valuable suggestions about the content.

48

LEXIS AS A LINGUISTIC LEVEL

M. A. K. Halliday

Source: C. E. Bazell, J. C. Catford, M. A. K. Halliday and R. H. Robins (eds), *In Memory of J. R. Firth*, London: Longman, 1966, pp. 148–162.

At a time when few linguists, other than lexicographers themselves, devoted much attention to the study of lexis, and outlines of linguistics often contained little reference to dictionaries or to other methods in lexicology, J. R. Firth repeatedly stressed the importace of lexical studies in descriptive linguistics.[1] He did not accept the equation of 'lexical' with 'semantic',[2] and he showed that it was both possible and useful to make formal statements about lexical items and their relations. For this purpose Firth regarded the statement of collocation as the most fruitful approach, and he sometimes referred, within the framework of his general views on the levels of linguistic analysis, to the 'collocational level'.[3]

The aim of this paper is to consider briefly the nature of lexical patterns in language, and to suggest that it may be helpful to devise methods appropriate to the description of these patterns in the light of a lexical theory that will be complementary to, but not part of, grammatical theory. In other words, the suggestion is that lexis may be usefully thought of (a) as within linguistic form, and thus standing in the same relation to (lexical) semantics as does grammar to (grammatical) semantics, and (b) as not within grammar, lexical patterns thus being treated as different in kind, and not merely in delicacy, from grammatical patterns. This view is perhaps implicit in Firth's recognition of a 'collocational level'.[4]

One of the major preoccupations of grammatical theory in present-day linguistics is the extension of grammatical description to a degree of delicacy greater than has hitherto been attained, and it is rightly claimed as a virtue of contemporary models that they permit more delicate statements to be made without excessive increase in complexity. A grammar is expected to explain, for example, the likeness and unlikeness between *this brush won't polish and this floor won't polish*, the three-way ambiguity of *John made Mary a good friend*, and the non-acceptability of *beautiful hair was had by*

Mary beside the acceptability of *the last word was had by Mary*. Such explanations require the recognition of distinctions which, as is well known, begin to cut across each other at a relatively early stage in delicacy, and the model has to accommodate cross-classification of this kind. The form of statement adopted (and the terminology) will of course depend on the model; but it is generally agreed that all such patterns need to be accounted for.[5]

As part of the process of accounting for these distinctions the grammar attempts, both progressively and simultaneously, to reduce the very large classes of formal items, at the rank at which they can be most usefully abstracted (for the most part generally as words, but this is merely a definition truth from which we learn what 'word' means), into very small sub-classes. No grammar has, it is believed, achieved the degree of delicacy required for the reduction of all such items to one-member classes, although provided the model can effectively handle cross-classification it is by no means absurd to set this as the eventual aim: that is, a unique description for each item by its assignment to a 'microclass', which represents its value as the product of the intersection of a large number of classificatory dimensions.

If we take into account the amount of information which, although it is still far from having been provided for any language, contemporary grammatical models can reasonably claim to aim at providing, there would seem to be two possible evaluations of it. One is that, when the most delicate distinctions and restrictions in grammar have been explained, all formal linguistic patterns will have been accounted for; what is left can only be accounted for in semantic terms. The second is that there will still remain patterns which can be accounted for in formal linguistic terms but whose nature is such that they are best regarded as non-grammatical, in that they cut across the type of relation that is characteristic of grammatical patterning.

The particular model of grammar that is selected may suggest, but does not fully determine, which of these two views is adopted operationally. For example, a model which distinguishes sharply between the grammar of a language and the use of the grammar, regarding corpus-based statistical statements as proper only to the latter, and therefore as outside the range of validity of a descriptive statement, is less easily compatible with the second view than is a model which does not make this distinction and which allows statistical statements a place in linguistic description; nevertheless it is not wholly incompatible with it.[6] Lexical statements, or 'rules', need not be statistical, or even corpus-based, provided that their range of validity is defined in some other way, as by the introduction of a category of 'lexicalness' to parallel that of grammaticalness.

One may validly ask whether there are general grounds, independent of any given model, for supplementing the grammar by formal statements of lexical relations, at least (given that the aim of linguistics is to account for as much of language as possible) until these are shown to be unnecessary. It may be a long time before it can be decided whether they are necessary or

not, in the sense of finding out whether all that is explained lexically could also have been incorporated in the grammar; there still remains the question whether or not it could have been explained more simply in the grammar. The question is not whether formal lexical statements can be made; they are already made in dictionaries, although at a low level of generality, in the form of citations. The question of interest to linguists is how the patterns represented by such citations are to be stated with a sufficient degree of abstraction, and whether this can best be achieved within or outside the framework of the grammar.

Let us consider an example. The sentence *he put forward a strong argument for it* is acceptable in English; *strong* is a member of that set of items which can be juxtaposed with *argument*, a set which also includes *powerful*. *Strong* does not always stand in this same relation to *powerful*: *he drives a strong car* is, at least relatively, unacceptable, as is *this tea's too powerful*. To put it another way, *a strong car* and *powerful tea* will either be rejected as ungrammatical (or unlexical) or shown to be in some sort of marked contrast with *a powerful car* and *strong tea*; in either case the paradigmatic relation of *strong* to *powerful* is not a constant but depends on the syntagmatic relation into which each enters, here with *argument, car* or *tea*.

Grammatically, unless these are regarded as different structures, which seems unlikely, they will be accounted for in a way which, whatever the particular form of statement the model employs, will amount to saying that, first, *strong* and *powerful* are members of a class that enters into a certain structural relation with a class of which *argument* is a member; second, *powerful* (but not *strong*) is a member of a class entering into this relation with a class of which *car* is a member; and third, *strong* (but not *powerful*) is a member of a class entering into this relation with a class of which *tea* is a member. It would be hoped that such classes would reappear elsewhere in the grammar defined on other criteria. *Argument, car* and *tea* will, for example, already have been distinguished on other grounds on the lines of 'abstract', 'concrete inanimate' and 'mass'; but these groupings are not applicable here, since we can have *a strong table* and *powerful whisky*, while *a strong device* is at least questionable.

The same *patterns* do reappear: *he argued strongly, I don't deny the strength of his argument, his argument was strengthened by other factors. Strongly* and *strength* are paralleled by *powerfully* and *power, strengthened* by *made more powerful*. The same restrictions have to be stated, to account for *the power* (but not *the strength*) *of his car* and *the strength* (but not *the power*) *of her tea*. But these involve different structures; elsewhere in the grammar *strong, strongly, strength* and *strengthened* have been recognized as different items and assigned to different classes, so that *the strong of his argument* has been excluded on equal terms with *the strong of his car. Strong* and *powerful*, on the other hand, have been assigned to the same class, so that we should expect to find *a powerful car* paralleled by *a strong car*. The classes set up to

account for the patterns under discussion either will cut across the primary dimension of grammatical classification or will need to be restated for each primary class.

But the added complexity involved in either of these solutions does not seem to be matched by a gain in descriptive power, since for the patterns in question the differences of (primary) class and of structure are irrelevant. *Strong, strongly, strength* and *strengthened* can all be regarded for this present purpose as the same item; and *a strong argument, he argued strongly, the strength of his argument* and *his argument was strengthened* all as instances of one and the same syntagmatic relation. What is abstracted is an item *strong*, having the scatter *strong, strongly, strength, strengthened*, which collocates with items *argue* (*argument*) and *tea*; and an item *power* (*powerful, powerfully*) which collocates with *argue* and *car*. It can be predicted that, if *a high-powered car* is acceptable, this will be matched by *a high-powered argument* but not by *high-powered tea*. It might also be predicted, though with less assurance, that *a weak argument* and *weak tea* are acceptable, but that *a weak car* is not.

As far as the collocational relation of *strong* and *argue* is concerned, it is not merely the particular grammatical relation into which these items enter that is irrelevant; it may also be irrelevant whether they enter into any grammatical relation with each other or not. They may be in different sentences, for example: *I wasn't altogether convinced by his argument. He had some strong points but they could all be met*. Clearly there are limits of relevance to be set to a collocational span of this kind; but the question here is whether such limits can usefully be defined grammatically, and it is not easy to see how they can.

The items *strong* and *power* will enter into the same set as defined by their occurrence in collocation with *argue*; but they will also enter into different sets as defined by other collocations. There is of course no procedural priority as between the identification of the items and the identification of the paradigmatic and syntagmatic relations into which they enter: 'item', 'set' and 'collocation' are mutually defining. But they are definable without reference to grammatical restrictions; or, if that is begging the question, without reference to restrictions stated elsewhere in the grammar. This is not to say that there is no interrelation between structural and collocational patterns, as indeed there certainly is; but if, as is suggested, their interdependence can be regarded as mutual rather than as one-way, it will be more clearly displayed by a form of statement which first shows grammatical and lexical restrictions separately and then brings them together.[7] If therefore one speaks of a lexical level, there is no question of asserting the 'independence' of such a level, whatever this might mean; what is implied is the internal consistency of the statements and their referability to a stated model.

Possible methods of lexical analysis, and the form likely to be taken by statements at this level, are the subject of another paper, by J. McH. Sinclair,

appearing in the present volume.[8] Here I wish to consider merely some of the properties of this type of pattern in language, and some of the problems of accounting for it. Clearly lexical patterns are referable in the first place to the two basic axes, the syntagmatic and the paradigmatic. One way of handling grammatical relations on these two axes is by reference to the theoretical categories of 'structure' and 'system', with the 'class' definable as that which enters into the relations so defined. In lexis these concepts need to be modified, and distinct categories are needed for which therefore different terms are desirable.

First, in place of the highly abstract relation of structure, in which the value of an element depends on complex factors in no sense reducible to simple sequence, lexis seems to require the recognition merely of linear co-occurrence together with some measure of significant proximity, either a scale or at least a cut-off point. It is this syntagmatic relation which is referred to as 'collocation'. The implication that degree of proximity is here the only variable does not of course imply how this is to be measured; moreover it clearly relates only to statements internal to the lexical level: in lexicogrammatical statements collocational restrictions intersect with structural ones. Similarly in place of the 'system' which, with its known and stated set of terms in choice relation, lends itself to a deterministic model, lexis requires the open-ended 'set' assignment to which is best regarded as probabilistic. Thus while a model which is only deterministic can explain so much of the grammar of a language that its added power makes it entirely appropriate for certain of the purposes of a descriptive grammar, it is doubtful whether such a model would give any real insight into lexis. Collocational and lexical set are mutually defining as are structure and system: the set is the grouping of members with like privilege of occurrence in collocation.

Second, in grammar a 'bridge' category is required between element of structure and term in system on the one hand and formal item on the other; this is the class. (This specific formulation refers to the 'scale-and-category' version of a system-structure model; but it is probably true that all models make use of a category analogous to what I am here calling the 'class'.) In lexis no such intermediate category is required: the item is directly referable to the categories of collocation and set. This is simply another way of saying that in lexis we are concerned with a very simple set of relations into which enter a large number of items, which must therefore be differentiated qua items, whereas in grammar we are concerned with very complex and variable relations in which the primary differentiation is among the relations themselves: it is only secondarily that we differentiate among the items, and we begin by 'abstracting out' this difference. In other words there is a definable sense in which 'more abstraction' is involved in grammar than is possible in lexis.

Third, the lexical item is not necessarily coextensive on either axis with the item, or rather with any of the items, identified and accounted for in the

grammar. For example, on the paradigmatic axis, in *she made up her face* one can identify a lexical item *make up*$_1$ whose scatter and collocational range are also illustrated in *your complexion needs a different makeup*. This contrasts with the lexical item *make up*$_2$ in *she made up her team* and *your committee needs a different makeup*. That the distinction is necessary is shown by the ambiguity of *she made up the cast, she was responsible for the makeup of the cast*. Grammatically, the primary distinction is that between *made up* and *makeup*; this distinction of course involves a great many factors, but it also relates to many other items which are distinguishable, by class membership, in the same way. If the grammar is at the same time to handle the distinction between *make up*$_1$ and *make up*$_2$ it must recognize a new and independent dimension of class membership on the basis of relations to which the previous dimension is irrelevant. Any one example can of course be handled by *ad hoc* grammatical devices: here for instance the potentialities of *make up*$_2$ in intransitive structures are more restricted. But such clearly grammatical distinctions, even when present, are so restricted in their range of validity that the generalizing power of a grammatical model is of little value as compared with the cost, in increased complexity, of the cross-classifications involved.

It may be worth citing a further example of a similar kind. We can distinguish grammatically, but not lexically, between *they want the pilot to take off* (= 'so that they can take off') and (= 'they desire him to do so'): these are not necessarily distinguished by intonation, although the unmarked tone selections are different. On the other hand it is easier to distinguish lexically than grammatically between *he took two days off* (= 'he did not work') and (= 'he reduced the time available'). In *the takeoff of the president* (= 'his becoming airborne') and (= 'the imitation of the president') the distinction can usefully be made both in grammar and in lexis.

On the syntagmatic axis, it may be useful to recognize a lexical item which has no defined status in the grammar and is not identified as morpheme, word or group. For example, in *he let me in the other day for a lot of extra work*, one could handle *let in for* as a single discontinuous item in the grammar; but this complexity is avoided if one is prepared to recognize a lexical item *let in for* without demanding that it should carry any grammatical status. Similarly the ambiguity in *he came out with a beautiful model* may be explained, instead of by giving two different grammatical descriptions, by identifying two distinct lexical items, *come* and *come out with* (and of course two different lexical items *model*).

It is not suggested of course that such non-coextensiveness between the items of grammar and those of lexis is the norm, but merely that for certain purposes it is useful to have a descriptive model of language that allows for it. At the same time the above considerations suggest that the lexical component requires not, as it were, a second 'runthrough' of the model

designed for the grammar but rather a specifically lexical model with distinct, though analogous, categories and forms of statement.

Nor is it suggested that the set of patterns recognized as language form is neatly divided into two types, the grammatical and the lexical. A model for the description of language form may recognize only one kind of pattern and attempt to subsume all formal relations within it: some grammatical models, as has been noted, envisage that it is the grammar's task to distinguish *strong* from *powerful* as well as to distinguish *a* from *the* and 'past' from 'present'; while a lexicographical model in which *a* and *the*, as well as *strong* and *powerful*, are entered in the dictionary and described by means of citations could be regarded as in a similar way attempting to subsume grammar under lexis. Even where the model recognizes two distinct kinds of pattern, these still represent different properties of the total phenomenon of language, not properties of different parts of the phenomenon; all formal items enter into patterns of both kinds. They are grammatical items when described grammatically, as entering (via classes) into closed systems and ordered structures, and lexical items when described lexically, as entering into open sets and linear collocations. So in *a strong cup of tea* the grammar recognizes (leaving aside its higher rank status, for example as a single formal item expounding the unit 'group') five items of rank 'word' assignable to classes, which in turn expound elements in structures and terms in systems; and the lexis recognizes potentially five lexical items assignable to sets.

But, to take a further step, the formal items themselves vary in respect of which of the two kinds of pattern, the grammatical or the lexical, is more significant for the explanation of restrictions on their occurrence qua items. The items *a* and *of* are structurally restricted, and are uniquely specified by the grammar in a very few steps in delicacy; collocationally on the other hand they are largely unrestricted. For the item *strong*, however, the grammar can specify uniquely a class (sub-class of the 'adjective') of which it is a member, but not the item itself within this class; it has no structural restrictions to distinguish it from other members of the class (and if the members of its 'scatter' *strong, strength*, etc., turn out to operate collocationally as a single item then this conflated item is not even specifiable qua class member); collocationally, however, it is restricted, and it is this which allows its specification as a unique item. There might then appear to be a scale on which items could be ranged from 'most grammatical' to 'most lexical', the position of an item on the scale correlating with its overall frequency ranking. But these are three distinct variables, and there is no reason to assume a correlation of 'most grammatical' with either 'least lexical' or 'most frequent'. The 'most grammatical' item is one which is optimally specifiable grammatically: this can be thought of as 'reducible to a one-member class by the minimum number of steps in delicacy'. Such an item may or may not be 'least lexical' in the sense that there is no collocational environment in

which its probability of occurrence deviates significantly from its unconditioned probability.

In a lexical analysis it is the lexical restriction which is under focus: the extent to which an item is specified by its collocational environment. This therefore takes into account the frequency of the item in a stated environment relative to its total frequency of occurrence. While *a* and *of* are unlikely to occur in any collocationally generalizable environment with a probability significantly different from their overall unconditioned probabilities, there will be environments such that *strong* occurs with a probability greater than chance. This can be regarded, in turn, as the ability of *strong* to 'predict' its own environment. As extreme cases, *fro* and *spick* may never occur except in environments including respectively *to* and *span* (the fact that *to and fro* accounts for only a tiny proportion of the occurrences of *to*, while *spick and span* may account for all occurrences of this item *span*, is immaterial to the specification of *fro* and *spick*); here it is likely that, for this very reason, *to and fro* and *spick and span* are to be regarded as single lexical items.

It is the similarity of their collocational restriction which enables us to consider grouping lexical items into lexical sets. The criterion for the definition of the lexical set is thus the syntactic (downward) criterion of potentiality of occurrence. Just as the grammatical system (of classes, including one-item classes) is defined by reference to structure, so the lexical set (of items) can be defined by reference to collocation. Since *all* items *can* be described lexically, the relation of collocation could be regarded as being, like that of structure, chain-exhausting, and a lexical analysis programme might well begin by treating it in this way; but this is not a necessary condition of collocation, and if closed-system items turn out, as may be predicted, to be collocationally neutral these items could at some stage be eliminated by a 'deletion-list' provided either, by cross-reference to the grammar or, better, as a result of the lexical analysis itself. Once such 'fully grammatical' items are deleted, collocation is no longer a chain-exhausting relation.

Moreover while grammatical structures are hierarchically ordered, so that one can recognize a scale of 'rank' each of whose members is a chain-exhausting unit (text items being then fully accounted for in sentence structure and again in clause structure and so on), it does not seem useful to postulate such an ordered hierarchy for lexis. Lexical items may indeed enter into a sort of rank relation: it is likely, for example, that on collocational criteria we would want to regard *stone, grindstone* and *nose to the grindstone* each as a separate lexical item, and though triads of this kind may be rare it looks as though we need the categories of 'simple' and 'compound', and perhaps also 'phrasal', lexical item, in addition to 'collocational span', as units for a lexical description. Since the only 'structural' relation in lexis is one of simple co-occurrence, these represent a single serial relation: the item *stone* enters (say) into the collocation *grindstone*, which then does not itself collocate exactly like the sum of its parts but enters as an item into (say) the

collocation *nose to the grindstone*, which likewise does not collocate like the sum of its parts but enters as an item into (say) the collocation *he's too lazy to keep his nose to the grindstone*. The first stage of such compounding yields a morphological (upward) grouping of items, the 'lexical series' which, like its analogue in grammar, may or may not coincide with the syntactic grouping recognized as a 'set': *oaktree ashtree planetree beechtree* presumably do operate in the same set, while *inkstand bandstand hallstand grandstand* almost certainly do not. The 'series' is formed of compound items having one constituent item in common; this item, here *tree* and *stand*, is the 'morphologically unmarked' member of the series and, likewise, if the series forms a set it may or may not be the 'syntactically unmarked' member of the set. Equivalence or nonequivalence between series and set is an interesting feature of lexical typology: one would predict that in Chinese, for example, practically all such series do form sets (with an unmarked member), whereas in Malay and English they very often do not.

The lexical item itself is of course the 'type' in a type-token (item-occurrence) relation, and this relation is again best regarded as specific to lexis. The type-token relation can be made dependent on class membership: just as in grammar two occurrences assigned to different primary classes, such as *ride* (verb) and *ride* (noun), can be regarded as different (grammatical) items, so in lexis two-occurrences assigned to different primary sets can be regarded as different lexical items. This can then be used to define homonymity: if the two occurrences of *model* in the example above are shown to differ according to criteria which would assign them to different sets then they represent two homonymous items. It is not to be assumed, of course, that grammatically distinguished items such as *ride* (verb) and *ride* (noun) may not also operate as distinct lexical items, as indeed they may; merely that *if* they turn out to belong to the same set they will on that criterion be said to constitute a single lexical item, as also will *strong, strength, strongly* and *strengthen*, and perhaps also (if they can be suitably delimited) non-cognate 'scatters' such as *town* and *urban*. This would provide a basis for deciding how many lexical items are represented by 'expressions' such as *form, stand* and *term*.

If we say that the criterion for the assignment of items to sets is collocational, this means to say that items showing a certain degree of likeness in their collocational patterning are assigned to the same set. This 'likeness' may be thought of in the following terms. If we consider n occurrences of a given (potential) item, calling this item the 'node', and examine its 'collocates' up to m places on either side, giving a 'span' of $2m$, the $2mn$ occurrences of collocates will show a certain frequency distribution. For example, if for 2,000 occurrences of *sun* we list the three preceding and three following lexical items, the 12,000 occurrences of its collocates might show a distribution beginning with *bright, hot, shine, light, lie, come out* and ending with a large number of items each occurring only once. The same

number of occurrences of *moon* might show *bright, full, new, light, night, shine* as the most frequent collocates.

On the basis of their high probability of occurrence (relative to their overall frequency) in collocation with the single item *sun*, the items *bright, hot, shine, light, lie, come out* constitute a weak provisional set; this resembles the weak provisional class recognizable in the grammar on the basis of a single 'item-bound' substitution frame—although in lexis it is *relatively* less weak because of the lower ceiling of generality: lexis is more item-bound than grammar. If we intersect these with the high frequency collocates of *moon* we get a set, whose members include *bright, shine* and *light*, with slightly greater generality. That is to say, *bright, shine* and *light* are being grouped together because they display a similar potentiality of occurrence, this being now defined as potentiality of occurrence in the environment of *sun* and in that of *moon*. The process can be repeated with each item in turn taken as the node; that is, as the environment for the occurrence of other items. The set will finally be delimited, on the basis of an appropriate measure of likeness, in such a way that its members are those items showing likeness in their total patterning in respect of all those environments in which they occur with significant frequency.

This is of course very much oversimplified; it is an outline of a suggested approach, not of a method of analysis. As Sinclair has shown, however, methods of analysis can be developed along these lines. Many other factors are involved, such as the length of the span, the significance of distance from the node and of relative position in sequence, the possibility of multiple nodes and the like. One point should be mentioned here: this is the importance of undertaking lexicogrammatical as well as lexical analysis. It is not known how far collocational patterns are dependent on the structural relations into which the items enter. For example, if *a cosy discussion* is unlikely, by comparison with *a cosy chat* and a *friendly discussion*, is it the simple co-occurrence of the two items that is unlikely, or their occurrence in this particular structure? All that has been said above has implied an approach in which grammatical relations are not taken into account, and reasons have been given for the suggestion that certain aspects of linguistic patterning will only emerge from a study of this kind. But it is essential also to examine collocational patterns in their grammatical environment, and to compare the descriptions given by the two methods, lexical and lexicogrammatical. This then avoids prejudging the answer to the question whether or not, and if so to what extent, the notion of 'lexicalness', as distinct from 'lexicogrammaticalness', is a meaningful one.[9]

An investigation on the lines suggested requires the study of very large samples of text. The occurrence of an item in a collocational environment can only be discussed in terms of probability; and, although cut-off points will need to be determined for the purpose of presenting the results, the interest lies in the degree of 'lexicalness' of different collocations (of items

and of sets), all of which are clearly regarded as 'lexical'. Moreover the native speaker's knowledge of his language will not take the form of his accepting or rejecting a given collocation: he will react to something as more acceptable or less acceptable on a scale of acceptability. Likely collocations could be elicited by an inquiry in which the subject was asked to list the twenty lexical items which he would most expect to find in collocation with a given node;[10] but the number of such studies that would be required to cover even the most frequent lexical items in the language is very large indeed. Textually, some twenty million running words, or 1,500–2,000 hours of conversation, would perhaps provide enough occurrences to yield interesting results. The difficulty is that, since lexical patterns are of low generality, they appear only as properties of very large samples; and small-scale studies, though useful for testing methods, give little indication of the nature of the final results.

It is hard to see, however, how the results could fail to be of interest and significance for linguistic studies. Their contribution to our knowledge of language in general, and of one language in particular, may perhaps be discussed in relation to the use of the term 'semantics'. If lexis is equated with semantics, the implication is that lexical patterns can only be described either externally (that is, as relations between language and non-language, whether approached denotatively or contextually) or lexicogrammatically (that is, in dependence on grammatical patterns). This restriction leaves two gaps in our understanding of language: the internal relations of lexis, and the external relations of grammar—that is, lexis (lexical form), and grammatical semantics. But linguistics is concerned with relations of both types, both internal (formal, within language) and external (contextual or 'semantic' between language and non-language); and all linguistic items and categories, whether operating in closed contrasts, like *the* and *a*, or 'past' and 'present', or in open ones, like *strong* and *powerful*, enter into both. Moreover, as Firth stressed, both these types of relation are 'meaningful': it is part of the meaning of 'past' that it contrasts with 'present', and it is part of the meaning of *strong* that it collocates with *tea*. The fact that the labels for grammatical categories are chosen on semantic grounds should not be taken to imply that they represent an adequate substitute for grammatical semantics; but equally the existence of traditional methods in lexical semantics does not mean that lexical items display no internal, formal patterns of their own.

A thesaurus of English based on formal criteria, giving collocationally defined lexical sets with citations to indicate the defining environments, would be a valuable complement to Roget's brilliant work of intuitive semantic classification in which lexical items are arranged 'according to the *ideas* which they express'.[11] But even such a thing as a table of the most frequent collocates of specific items, with information about their probabilities, unconditioned and lexically and grammatically conditioned, would be of

considerable value for those applications of linguistics in which the interest lies not only in what the native speaker knows about his language but also in what he does with it. These include studies of register and of literary style, of children's language, the language of aphasics and many others. In literary studies in particular such concepts as the ability of a lexical item to 'predict' its own environment, and the cohesive power of lexical relations, are of great potential interest.[12] Lexical information is also relevant to foreign language teaching; many errors are best explained collocationally, and items can be first introduced in their habitual environments.[13] A further possible field of application is information retrieval: one research group in this field is at present undertaking a collocational analysis of the language of scientific abstracts.[14]

Only a detailed study of the facts, such as that now being undertaken by Sinclair (the principles and methods worked out by him are described in the paper already referred to in the present volume), can show in what ways and to what extent the introduction of formal criteria into the study of lexis, as implied by the recognition of a 'lexical level', are of value to any particular applications of linguistics. But there seem to be adequate reasons for expecting the results to be interesting; and if they are, this is yet another indication of the great insight into the nature of language that is so characteristic of J. R. Firth's contribution to linguistic studies.

Notes

1 See especially 'The Technique of Semantics', *TPS* 1935 (reprinted in J. R. Firth, *Papers in linguistics 1934–1951*, London, Oxford U.P., 1957).

2 'Modes of meaning', *Essays and Studies* (*The English Association*), *1951* (in *Papers in linguistics 1934–1951*; pp. 195–196): 'It must be pointed out that meaning by collocation is not at all the same thing as contextual meaning, which is the functional relation of the sentence to the processes of a context of situation in the context of culture. . . . Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words'. Compare also 'The Use and distribution of certain English sounds', *English Studies* 17.1 (1935) (in *Papers in linguistics 1934–1951*; p. 37), in reference to 'lexical substitution-counters' (lexical items): 'This (sc. the lexical) function should not be misnamed *semantic*'.

3 See 'A Synopsis of linguistic theory', *Studies in linguistic analysis* (Special Volume of the Philological Society), Oxford, Blackwell (1957), p. 12. In the present paper 'lexical level' has been used in preference to 'collocational level' in order to suggest greater generality and parallelism with the grammatical level.

4 It is also stated explicitly by Firth in 'A Synopsis of linguistic theory', p. 12: 'Collocations of a given word are statements of the habitual or customary places of that word in collocational order but not in any other contextual order and emphatically not in any grammatical order'. Note that here 'order' refers to the 'mutual expectancy' of syntagmatically related categories, such as elements of structure in grammar or phonology, and *not* to linear sequence: cf. ibid., pp. 5, 17 and my 'Categories of the theory of grammar', *Word* 17.3 (1961), pp. 254–255.

5 That is, that distinctions are made which involve the recognition of more finely differentiated syntagmatic relations in the grammar, and that these in turn define further sub-classes on various dimensions within previously defined classes.

6 The place of collocational restrictions in a transformational grammar is considered by P. H. Matthews in 'Transformational grammar' (review article), *ArchL* 13.2 (1961).

7 For a discussion of the relation between grammatical and lexical patterns see Angus McIntosh, 'Patterns and ranges', *Lg* 37.3 (1961).

8 See J. McH. Sinclair, 'Beginning the study of lexis', *passim*.

9 The implication is, in effect, that 'wellformedness' is best regarded as 'lexicogrammaticalness', and that a departure from wellformedness may be ungrammatical, unlexical or unlexicogrammatical. That the last two are distinct is suggested by such examples as *sandy hair, sandy gold* and *sandy desk: sandy desk* is unlexical, in that this collocation is unlikely to occur in any grammatical environment, whereas *sandy gold* is merely unlexicogrammatical: there is nothing improbable about *golden sand*. An analogous distinction is observable in clichés: in *shabby treatment* the mutual expectancy is purely lexical, and is paralleled in *they treated him shabbily, a shabby way to treat him* and so on, whereas the collocation *faint praise* is restricted to this structure, in the sense that it will not occur with similar probability under other grammatical conditions.

10 Compare the methods used to assess the *disponibilité* of lexical items in the development of 'Français Fondamental': see G. Gougenheim, R. Michea, P. Rivenc and A. Sauvageot, *L'élaboration du français élémentaire*, Paris, Didier (1956).

11 P. M. Roget, *Thesaurus of English words and phrases* (1936 ed.), London, Longmans (1960), p. xiii.

12 Cf. J. R. Firth, 'Modes of meaning'.

13 The following text examples may be cited in this connection: festive animals, circumspect beasts, attired with culture, funny art, barren meadows, merry admiration, the situation of my stockings was a nightmare, lying astray, fashionable airliner, modern cosy flights, economical experience, delightfully stressed, serious stupid people, shining values, a wobbly burden, light possibility, luxurious man, whose skin was bleeding, driving a bicycle, old and disturbed bits of brick wall, a comprehensive traffic jam, her throat became sad, my head is puzzled, people touched with assurance, thoughts are under a strain, a sheer new super car.

14 This research is being undertaken by Dr. A. R. Meetham and Dr. P. K. T. Vaswani at the National Physical Laboratory, Teddington, Middlesex.

49

BEGINNING THE STUDY OF LEXIS

John Sinclair

Source: C. E. Bazell, J. C. Catford, M. A. K. Halliday and R. H. Robins (eds), *In Memory of J. R. Firth*, London: Longman, 1966, pp. 410–430.

It might sound strange to suggest that great difficulties lie in the way of anyone who wants to study the vocabulary of a language. One could point to the great dictionaries and thesauri, the fat concordances and the humbler word-lists, or one could note that the popular accounts of the history of languages are often little more than histories of their vocabulary. Nevertheless, if one wishes to study the 'formal' aspects of vocabulary organization, all sorts of problems lie ahead, problems which are not likely to yield to anything less imposing than a very large computer. The reasons why these problems arise plunge us into the study of lexis.[1]

I am not here going to defend the setting-up of lexis as an independent part of language **form**. That is done in a paper presenting the present approach to lexis by M. A. K. Halliday elsewhere in this volume.[2] I take it for granted that it is desirable and necessary to look at the internal patterns of language from two different, interpenetrating aspects. When we are studying language form, we are concerned to examine the way in which patterns recur, without taking into account other patterns which lie outside language, patterns of social or natural organization. It would be surprising indeed if a great many of these external patterns were not reflected in language, but we try not to inflate this expectation into preconception.

Let us consider three words: *tome, paperback* and *cruelty*. The first two share a non-linguistic notional similarity—in an 11-plus test, we would all select *cruelty* as the odd word out. The first two occur in the same paragraph in Roget.[3] But if we consider the chances relative to the total frequency of any given two of them occurring within a short stretch of language, there is little to choose between them. No pair would get better odds than the others. All the **collocations**[4] are unlikely.

This is not to say that the patterns of language form give us no way of pairing the first two against the third; we may well find that *tome* and *paperback*, while showing no special tendency to co-occur, may share collocations with other words which are not associated with *cruelty*, words like *edition, bookshop, paper, print*. If we find such patterns, we satisfy our expectations, but express them in terms of the way in which language is internally organized, without having assumed in the first place that their similarity of reference was *a priori* a similarity reflected in linguistic form.

The two interpenetrating ways of looking at language form are **grammar** and **lexis**. In grammar we look at the patterns of language as if they could be described by a large number of separate choices, each choice being from a small list of possibilities. In each case, the possibilities can be itemized in full, and we can talk of choosing one item *rather than* another. The choice between Active and Passive Voice in the verbal group in English offers a typical example of a grammatical **system**. Every verbal group is either one or the other, and there are only two possible choices.

I used the words *as if* above. Grammar organizes language form in its own terms, and generally ignores any patterns that do not resolve themselves into systems. Items involved are listed, at the end of each piece of grammatical description, as items between which grammar has failed to distinguish.[5] If grammar accounted for the whole of language form, then this situation would not be satisfactory, and the grammarian would have to try to continue, using material which would get more and more intractable to his methods, and getting less and less generality for his efforts.

But running parallel to grammar is lexis, which describes the tendencies of items to collocate with each other. A study of these tendencies ought to tell us facts about languages that cannot be got by grammatical analysis, since such tendencies cannot be expressed in terms of small sets of choices. One lexical item is not chosen *rather than* another, lexical items do not contrast with each other in the same sense as grammatical classes contrast. There are virtually no impossible collocations, but some are much more likely than others.[6] At the present time, lexical statements look very much weaker than statements made using the precise and uncompromising machinery of grammar. There is a much less elaborate framework to lexical descriptions and much less certainty in the statements. Thus *he is taught* contrasts with *he teaches* in one grammatical system, with *he was taught* in another and with *he isn't taught* in another; but we cannot so describe the difference between *teaches, learns, reads, studies, . . .* etc.; what we can say about these is that they often occur close to each other and in the company of items like *history, desk, class*.

This aspect of the patterning of language has been known for a long time, and occasional references to it are found in description, particularly when the patterns are strong and tending towards the cliché. It is easy enough to hint at it, give a few examples, and move on. It is at present impossible to

prove even the assertions about lexical patterns that I have made, or to justify calling any of the words quoted **lexical items** (and so suggest that they are identified units of a description) or to say anything at all objective about the lexical structure of a language. Whereas we can collect a few thousand examples of the verbal group and notice what patterns recur, there is no such general lexical category as 'verbal group' turning up several times a sentence, and there is no easy way of collecting a few thousand occurrences of any lexical item. Furthermore, it seems likely that the more common items will be much more difficult to describe than the rare ones.[7]

Consequently the theory of lexis is fairly rudimentary; it may satisfy our intuitions but it has not been shown to be valid and we have yet to see what a comprehensive description of the lexis of a language looks like. The central problem is the circularity in the definition of the basic unit of description, the lexical item.

The theory can readily define a lexical item as follows:

> a formal item (at least one morpheme long) whose pattern of occurrence can be described in terms of a uniquely ordered series of other lexical items occurring in its environment.[8]

The criterion throughout will be that of collocation in a given stretch of text. Unfortunately, the stretch of text is *not* 'given'; we are trying to define the item in terms of its environment and the environment (by implication) in terms of the item. The implied definition of the environment is that it is the extent of text which is relevant in the description of an item.

To make these statements clearer, let us consider the nature of the predictive power of items in a text. In the first place, it seems clear that we shall have to measure the extent of each environment in lexical items, thus dealing another blow to the usefulness of our definition in practice. But if we do not do this, then we must use a unit of measurement (e.g. the orthographic word) from another level of language description, thus violating the autonomy of the level of lexis which we have made a basic part of our theory. Let us look at an example.

it was an auspicious occasion
the occasion on which it was done was not an auspicious one

It is likely that the lexical value of the collocation of *auspicious* and *occasion* is similar in each sentence. The two items are lexically more or less contiguous, and the description of the material intruding in the second sentence will be made at a more delicate stage of analysis.

In the sentence

I posted the letter in the pillar-box

we might expect the items *post, letter, pillar-box* to predict each other much more than *drop, letter, puddle* in

I dropped the letter in a puddle.

It seems reasonable to hope that in

he flipped the letter with careless elegance into the pillar-box

we can count an instance of the collocation *letter, pillar-box*. But if we find examples like:

> *The letter in his hand, that strangely anonymous, insignificant monument to his dreadful toils of composition, he posted, with a careless elegance which belied the gnawing terror he felt at the thought of whose cruel hands might open it. Then he continued past the pillar-box into the engulfing night*

we might still feel that there was a pattern *post, letter, pillar-box* but it would be difficult indeed to devise a workable procedure for recognising it. A thread runs through: *letter - hand - composition - post - hands - open - pillar-box* but we cannot build a superstructure on such a slender thread.

We do not have to assume that patterns which we perceive in examining a text and patterns that our techniques allow us to describe are identical. Neither of these may be the same as the most efficient possible account of the lexical structure[9] of the text. The patterns perceived by a trained linguist examining a text are unreliable and usually extremely tentative. Our techniques are perforce based on the capabilities of machines, though the demands made by the machines do not affect the theory.

At this point we must note that the techniques of analysis are extremely crude. The lexical patterns we perceive in examining texts occur over varying extents of text according to the particular circumstances of each example. The existence of a mutual prediction can depend on any or all of

(a) the strength of the predictions of items over each other
(b) the distance apart of the items
(c) the nature of the items which separate them, whether continuing a 'thread' as above, or not
(d) the grammatical organization.

The feature of physical distance apart of two items is one that requires further comment. It is clear that distance ultimately sets the boundary of collocation, although it will be a different distance in different cases. Ought we to argue from this that the proximity of items within a collocation should

be assessed? If *post* in a text occurs 100 times just next to *letter* and 100 times with another item in between, do we have a less strong pattern than say, if *pillar-box* occurs 150 times next to *letter* and only 50 times next but one? Often we might feel that there are grounds for distinguishing collocations on this basis, rather than merely noting that both *post* and *pillar-box* occur 200 times in collocation with *letter* and relegating the differing proximities to a less important place. But with some frequent collocations (e.g. *buy* and *cheap*) we might find that they are habitually separated, but we will be reluctant to argue from this that the collocation *buy* + *cheap* is less strong than another that occurred without any interventions. The theory, as far as it goes, gives us no grounds for weighting proximities, and the study of examples tends to deny that items become *steadily* less mutually predicting as the distance between them increases. So we reject, for the moment, the suggestion that degree of proximity within the chosen boundaries of collocation should be considered of primary importance, thus also rejecting a neat but dangerous way of fixing the boundaries of each collocation.

It is also likely that there are cases of two or more items collocating more frequently with each other in certain of their possible sequences. *Buy* and *cheap* might well exemplify this. But as the theory stands at present, the primary structural criterion is that of co-occurrence, in any sequence, with or without intervening material; features such as preferred sequences, or habitual interventions, are secondary in structure. It remains for the facts to confirm or refute this order of priority, and a manual study has been started to collect hints as to the extent to which simple co-occurrence is inaccurate in describing the facts.

We may use the term **node** to refer to an item whose collocations we are studying, and we may then define a **span** as the number of lexical items on each side of a node that we consider relevant to that node. Items in the environment set by the span we will call **collocates**. The extent of the span is at present arbitrary, and depends mainly on practical considerations; at a late stage in the study we will be able to fix the span at the optimum value, but we start with little more than a guess.

What we call **items** are of course not really entitled to the term, since only after looking at the collocations over a long text will we be able to identify what the actual items are. Let us examine the collocations in a short piece of text and introduce some more terms.[10]

> / here that's where we can save money / in what way? / we can't drink. We don't drink and we don't smoke and we spend all our money on clothes / We do spend all our money on clothes / I know, and what an excuse for having to spend our money somehow / I wish I could save mine / It's impossible to save money / that's my lament yes / unless you never go out and are very strong willed / Oh it's not so bad for women, you know, it's / is it not? / us poor blokes

> that suffer. / It's poor blokes like me that gets persuaded to take some woman to the what d'you call it ball at two guineas a ticket and a quid for drink and a quid for taxis / They're very lucky that can persuade you / They're lucky they can persuade you. I'm in with the wrong set / Well she didn't persuade me. It was someone else / ha / persuaded me to buy a ticket / — / I'm still not sure whether it's the thing to do to accept a quid from the girl towards the ticket / — / I see no reason why not / no not for not if the tickets are two quid / two guineas and she says she'll pay half you think that's reasonable do you? / I'd get the extra shilling off her / Oh you wouldn't / oh medics honestly they're worth every inch of their reputations / if a bloke were taking you to a ball and you offered to pay / I suppose I better offer eh? / no and you offered to pay two half of a ticket you don't think there's anything wrong with the boy to accept it? / half of your own ticket? / no no you offered to pay / but you're not a student (well I have got) you're earning are you / you should be earning / my income is not no larger really than anybody who's on a grant / well I think it would be all right to accept the money / yes I think so 'cos it cost plenty of money in drinks, I mean, I can just say 'well I can get all of the drink' /

We will take the span as 3 lexical items before and after each node, and we will only select fairly obvious lexical items. The first occurrence of *money* embraces the piece (items underlined)

> *save money / in what way / we can't drink. We don't drink*

and the second embraces

> *drink and we don't smoke and we spend all our money on clothes / I know, and what an excuse*

and so on. Very little valuable information can be got from such a small text because we cannot assume that each item is behaving at its most typical. However, to show the processes, let us look at the collocations of *money*, *pay* and *ticket* displayed on p. 22.

Below each item is listed its total environment in the text, so that in this text, with a span of 3 items on either side, *spend* occurs four times as a collocate of *money*, *think* three times, etc. This is in miniature the source data for what we call the **cluster** of a lexical item. The Total Environment Table for an item differs from a cluster in two main respects:

(a) It does not relate the number of occurrences of a collocate to the total frequency of that collocate. Obviously the more frequent an item is, the

Total Environment Tables

Item:	*money*		*pay*		*ticket*	
occurrences	7		4		6	
collocates			offer	5		
(*frequency ordered*)	drink	4				
	save	4			quid	4
	spend	4				
(*2 or more occurrences*)	clothes	3	half	3	accept	3
	think	3				
	cost	2	better	2	boy	2
	excuse	2	think	2	guinea	2
	impossible	2	ticket	2	half	2
	know	2			offer	2
	money	2			pay	2
	plenty	2			persuade	2
					reason	2
					see	2
					ticket	2
(*1 occurrence only*)	accept		ball		ball	
	all right		earn		buy	
	get		guinea		drink	
	go out		have got		girl	
	lament		quid		say	
	say		reasonable		student	
	smoke		say		sure	
	strong willed		student		the thing to do	
	way		suppose		think	
	wish		take		woman	
					wrong	

greater is the chance of it appearing regardless of lexical connections. In the environment for *money*, for example, both *think* and *save* occur, but while this fact accounts for all the occurrences of *save* in the passage, there are two occurrences of *think* which are not in collocation with *money*.

(b) In this example, the lexical items have been isolated. An objective study would consider each morpheme as potentially a lexical item, and so total environment tables would be prepared for what are often called 'grammatical items'. If we are correct in assuming that such items are lexically independent (that is, ultimately they can show no significant cluster since other items appear in their environments in accordance with predictions based on the total frequency of the items in a text), the total environment tables will not show this independence, because they do not relate the total number of occurrences of node and collocate to each other.

To summarize, we measure both the way in which an item predicts the occurrences of others (and get results such as the table on page 22), and also the way in which it is predicted by others. Then we choose the second of these measurements for our statement of the lexical meaning of the item. This statement is called a **cluster** and is derived from the total environment tables.[11]

Firth said (op. cit., page 196) 'One of the meanings of *night* is its collocability with *dark*' and we can go on from there to say that the **formal meaning** of an item A is that it has a strong tendency to occur nearby items B, C, D, less strong with items E, F, slight with G, H, I, and none at all with any other item. And that is exactly what is tabulated in the cluster.

Already we can see how crude the technique of collocation is. There is an enormous amount of **casual collocation**, that is, the span setting has netted a lot of items which are most unlikely to have any predictive power over the node. They may be accidental, reflecting the place, perhaps, where someone breaks into a committee meeting with the coffee; or they may include the magnificent images of some of our greatest poetry. But lexis has a broad mesh in the early stages, looking always for typicality, and rejecting the atypical. The unusual collocations will come into their own only when they have been proved to be unusual, and their degree of unusualness has been measured.

The vital distinction between casual and significant collocation is, as we have seen, made according to the frequency of repetition of the collocates in several occurrences of an item. The more occurrences of an item that we study, the clearer the picture will be; each particular casual collocate will be unlikely to occur again, and though there will be more and more of them the significance of any one of them will steadily decrease. At the same time collocates typical of the item in question will impress their pattern more and more strongly until the pattern is broadly speaking complete and the evidence of further occurrences does not materially alter the pattern.

If p = the total number of occurrences of items in a text, and f = the total number of occurrences of a particular item, then the probability of that item occurring at each item-place in the text is $\frac{f}{p}$. If s = the span setting multiplied by two (since the span measures the number of item-places on *each* side of a node), the probability of our item collocating with each successive node is $\frac{sf}{p}$. Then if we consider a particular node which occurs n times in the text, the probability of our item collocating with this node is $\frac{nsf}{p}$. This figure can then be compared with the actual facts of the text, the actual number of times this collocate occurs with this node. Statistical tests can assess the significance of any discrepancy between the predicted and the actual figures, giving a positive correlation (where the collocate

attracts the node to itself), a negative correlation (where the collocate repels the node), or an absence of correlation. The last state of affairs tells us that the items in question are neutral with regard to each other in the text, possibly because they just *are* neutral with regard to each other, or possibly because the text is not long enough to reflect their performance in the language as a whole.

No matter how long a text was chosen, there would be some items that did not occur often enough to be described on the evidence of this text alone, and they could only be studied with reference to items that had been established in the text. In a very long text, these would probably only amount to a few barely-assimilated foreign words, stray technical terms, and, of course, misprints.

We have reached the stage of postulating the arrangement of the clusters of all 'text-described' items, and devising methods of removing the non-significant 'tails' of these clusters. Since our items are only provisionally identified, we must not permanently discard the tails. Subsequent adjustments may well call for part of a tail to be incorporated in the cluster. What sort of adjustments can we make?

To answer this question we must return to the stage where we guessed at the provisional items, and refine our guesswork, relying on the near certainty that the more the occurrence of two items in collocation defies the chance predictions, the nearer we are to a correct identification of them. Our procedures depend on the way in which we guessed in the first place.

(a) We could guess, throughout the text, where each item stopped and the next one started, and mark forms differently where we felt sure that more than one item had the same form (like $quack_1$ collocating with *duck*, and $quack_2$ collocating with *medicine*). This method ought to yield a good approximation to lexical items, but it is impracticable, since it depends on a plentiful supply of trained editors; and unreliable, since it would hardly be possible to keep track of the millions of subjective decisions that would be introduced at an early stage.

(b) We could rely on the orthographic word (in a transcribed text) as the best approximation. This is, we think, the best unit from the point of view of quick results, but it would require that many words which appeared to vary only on a grammatical axis would be conflated. *Drive, drives, drove, driving* and *driven*, for example. It is a fair guess that it is proper to conflate these forms, but it may not be a correct guess, and we would have to retain the possibility of reconstructing the original form quickly. If it happened, for example, that *-ing* was an essential component of a lexical item, as in

I'm dying to see the show

we would lose a very useful criterion for identifying the item be *dying to*. All we really know about the two parts of *driving* is that the first is lexically

important and probably has very little influence in the grammar while the second is operative in the grammar, and probably has little influence in the lexis.[12]

(c) Perhaps the most reliable method, then, would be to split *driving* into its two parts (which we crudely call **morphemes**), being confident that *-ing* is not very important, except when considered in conjunction with other morphemes. At least we could be confident that no lexical item was smaller than the chosen unit, and the remaining problems would be:

(i) detecting multi-morpheme items;
(ii) detecting more than one item with the same **form**.[13]

The method requires a minimum of guesswork, and readily adapts to mechanical editing. If we assume that this method has been adopted, we can now look into the way in which we can make our provisional analysis more accurate. At this stage it becomes obvious that the work will get beyond the capacity of the most dedicated human drudge. Because, if we get evidence to split the form *quack* into $quack_1$ and $quack_2$, then all clusters in which *quack* is a significant collocate will have to be adjusted, the various occurrences of *quack* will have to be identified as one or other of the two possibilities, and the arithmetic will have to be done all over again. Similarly if we get evidence to identify the item *run to seed* we will have to look at each occurrence of *run* and examine its environment, and identify whether or not it is part of the item *run to seed*. There are probably twenty or more items incorporating the form *run*,[14] so the ultimate significance of an occurrence of *run* will look very different from the provisional results. The process of refinement, one must stress, is not one of merely jettisoning collocates because they appear in clusters under false pretences. Since significance is related to the total number of occurrences of an item, the recognition of a fairly rare item like *run* to *seed* among the nondescript *runs* may bring it into prominence in places where before the vast frequency of *run* had obscured the significance.

Before we go further into the processes of refinement, we had better set out certain assumptions about the nature of polymorphemic lexical items. We shall try to avoid preconceptions, and, particularly, to avoid analogies with other levels of language, particularly grammar.

We fix no upper limit to the size of polymorphemic items. *Take the bull by the horns, don't put all your eggs in one basket*, may well turn out to be best regarded as single lexical items. A polymorphemic item may have its components in any sequence. Though no clear examples of varying sequence have come to light yet, the best model we have for the behaviour of component morphemes in lexical items is collocation, and there is certainly no sequence restriction there.[15] The components of a polymorphemic item may be discontinuous. Oaths can interrupt them almost anywhere, and concoctions like

you must cut your coat, I'm afraid, according to your cloth
he's run utterly to seed
put all his nuclear eggs in the West German basket

do not sound wildly esoteric (and one is remembered from a Sunday newspaper).

As soon as we accept these assumptions, we must face the charge of distorting the facts in order to exclude grammatical considerations, and we must take a closer look at the influence of grammar on lexis. But now we have a framework of lexical theory within which to study whatever influence grammar has.

It is quite true that most of the polymorphemic lexical items that spring to mind are grammatically restricted in some way. Alter the grammar and you no longer have the item. The *seed* in *run to seed* cannot inflect, for example, and *cats* in *raining cats and dogs* cannot remain a constituent of the item if it is made singular. The key point is that such grammatical considerations are not typical of grammar. It is characteristic of a word like *cats* in grammar that there *is* a word *cat* with which it contrasts, not that there is not.

One's attitude to grammar may change a little when it becomes clear that lexis can bear some of the troublesome burdens of description. The ambiguity in a sentence like

I looked over the house

may be explained

(a) by assigning different grammatical analyses to the two meanings; considering *look over* to be the verb when it is also a lexical item, and *look* to be the verb when the *over* is not an essential constituent of a lexical item. The primary analytical cuts would then be, respectively,

I / looked over / the house
I / looked / over the house

It is too early to be sure that lexis and grammar will always correlate so neatly: some results of an investigation of about 600 'phrasal verbs' carried out by Halliday and Dixon suggest that an intuitive answer to the question 'Is this phrasal verb a single lexical item?' correlates very poorly with nineteen other questions, all of which are grammatical.

(b) by forcing lexis and grammar artificially apart, and assigning but one grammatical analysis to the example above, leaving the lexical analysis only to show the ambiguity. This separation would help cases like

(i) It's raining cats and dogs.
(ii) He's training cats and dogs.

Grammatically, one could then say, both occurrences of *cats* are plural; the lack of choice in (i) is a consequence of lexical item componence.

It is important to notice that all the components of a polymorphemic item are equally relevant, no matter what other patterns any of the same forms may have elsewhere. In *run to seed* the *to* is just as much an identifying component of the item as *seed*, even though there may be a (different) lexical item *run* but no item *to*. The sequence is also a component of this item, but the various inflections of *run* probably are not. So with the *-s* of *cats and dogs*, the *and* and the sequence.

Grammar has ranks,[16] and it is possible to talk of items as being 'free' at one rank and 'bound' at another, higher rank. The form *the* as a morpheme is free, that is, it can stand alone as a word. But as a word it is bound, since it cannot stand alone as a nominal group. We must not take this concept over into lexis, since we can detect only one rank in lexis, the item, which must therefore be free.[17] Its components may have the same forms as other free items but there is no way of relating them at the level of lexis. There is no way of relating *cats* as part of *cats and dogs* and *cat* on its own as a free item, except trivially. Lexically the two are no more likely to be related than any other two forms, and since this lack of relationship is the criterion for separating them, the situation is not to be lamented.

These points regarding the componence of lexical items can be used to question a fundamental assumption about the natures of grammar and lexis. Because of the pre-eminence of grammar, one is inclined to think of a text as 'having a grammatical structure' throughout, but only sporadically 'having a lexical structure'. We speak casually about 'fully grammatical items' or 'function words' as if there were items which were entirely irrelevant in the study of lexis. In the early examples in this article this assumption was made for the sake of exposition, but it seems to have no foundation in fact. Every morpheme in a text must be described both grammatically and lexically; against the fact that *boy*, *boys*, *boy's* and *boys'* are likely to be lexically distinguished only at a very delicate stage in the description, if at all, we must set the fact that *boy*, *man*, *youth* are in the same position grammatically. Each successive form in a text is a lexical item or part of one, and there are no gaps where only grammar is to be found. The so-called 'function words' are presumably words which never attain the status of separate lexical items, but neither do *amok*, *hale*, *eke* as words, or the famous *cran* of *cranberry*, *ceive* of *receive*, etc., or the splendid truncations of Wodehouse, as morphemes.

To conclude the remarks on polymorphemic items, we must anticipate how we will recognize them. Why do we expect that some occurrences of *cold* + *feet* will be singled out for treatment as occurrences of a separate polymorphemic item whereas *cold* + *hands* will not so tempt us? In essence we pick out a polymorphemic item when its cluster cannot be predicted

from the clusters of its components. The 'macro-cluster' of *cold* + *hands* will indeed be different from the cluster of *cold* and the cluster of *hand* but it should be predictably different. The common collocates in the two clusters we might expect to be strengthened in significance, and we would be surprised to find any item appearing that did not occur prominently in one of the two original clusters. With *cold* + *feet*, on the other hand (assuming that we have managed to isolate the occurrences of the polymorphemic item[18]), we would expect items like *risky*, *call off* to appear, although in the two original clusters their significance was masked by the numerous collocates relevant to the two words as separate items. Instances of *amok*, *cran*, *umbrage*, etc., will be identified by the feature of 100 per cent prediction of another form. There is, therefore no choice, and so without looking at their clusters we can combine *run* + *amok* and *cran* + *berry*, rearrange the clusters and spans (since we will have counted each as two items before) and try again. So the question of whether the *berry* in *cranberry* is the same morpheme as the *berry* in *raspberry* or just plain *berry* can be easily answered, without recourse to notional contemplation. From a lexical standpoint, if the clusters of *cranberry*, *raspberry* and *berry* are identical then it is the same morpheme *berry*, and if not, it isn't. This follows from the point that the *berry* in *cranberry* has as little choice as the *cran*, and we must follow the fortunes of the two together.

The argument is perhaps worth illustrating with a further example, involving three sentences:

(a) He might come soon.
(b) He strove with all his might.
(c) He strove with all his might and main.

The same form *might* appear in all three sentences, and the problem is how this form relates to one or more than one lexical item. The *might* of sentence (a) is clearly different from the others; whether it is or is not the exponent of a lexical item can be left aside for the moment. The reason for its separation can be expressed lexically, in terms of its differing pattern of collocation from the others.

The form *main* cannot significantly collocate with *strive*, etc., unless *might* is also present, so *main* can only be one component of a polymorphemic lexical item, which we can identify as *might and main*. So the question whether the two *mights* of (b) and (c) are lexically the same (they seem notionally the same) depends on whether or not the cluster of *might* is identical to the cluster of *might and main*. We can only hazard a guess as to this at the present time; it seems not unlikely that there may be 'long forms' and 'short forms' of the same item (compare *fed up* and *fed up to the back teeth*).

We will now further discuss the lexicogrammatical features in relation to the second of the two dimensions of refinement on page 26, the detection

of two or more items having the same form. Here grammar can again help a lot, but it must not be used in defiance of lexical considerations.

It is true, for example, that grammatical word-class distinctions may parallel lexical distinctions, though a glance at the repetitiveness of dictionaries which give separate entries according to the word-class of an item shows that the parallelism may be carried too far. But we may point to examples like:

mat, noun, collocating with *door*, *wipe*, *hall*, etc.
mat, verb, collocating with *hair*, *jam*, *thick*, etc.
mat, adj., collocating with *paint*, *finish*, *surface*, etc.

where the coincidence of grammatical and lexical boundaries is considerable. In contrast with this we can advance examples like *quack* noted earlier, where the pattern is different:

LEXIS / GRAMMAR	$quack_1$ (duck)	$quack_2$ (medicine)
quack (noun)	x	x
quack (verb)	x	

Here the grammatical and lexical patterns only partly coincide. If we started with the grammatical evidence, treating lexis as a follow-up from grammar, then it is true that some distinction between *quack* (noun) and *quack* (verb) would emerge in collocation. But it would not be as clear or as accurate a distinction as the lexical item distinction which cuts across the grammar. On the other hand, the grammatical information can be used in lexical description. It will emerge clearly that *quack* + *ing* only occurs in $quack_1$, for example.

On what grounds do we split *quack* into two or more lexical items? Or the several hundred much more complex cases like *hand*, making new items with *bird*, *in*, *over*, etc., and collocating with *marriage*, *give*, *factory*, *whist*, *horse*, *tall*, *legible*, *big*, *finger*, *fortune*, etc.? Grammar is hardly any help at all, and the distinctions gained by a word-class division make little inroads on the complexity of this form. We are forced to examine the 'internal consistency' of the cluster.

There is no theoretical requirement that collocates of an item must intercollocate but in practice it can be seen that some occurrences of *hand* are collocationally distinct from other occurrences. That is, there will be a group of collocates, e.g. *marriage*, *daughter*, *engagement*, which will occur together frequently as collocates of *hand*, and another group including *whist*, *rummy*, *ace*, *flush* also often together within collocations of *hand*,

But the chances of *whist* and *marriage* co-occurring are as poor as those of *archbishop* and *fish*. The only way in which these groups are joined is by co-occurrence with the form *hand* (and possibly other accidental ambiguities, e.g. *card*). The way in which the various groups of items can be separated is to compare each collocation with all the others, drawing together those that are like each other until as clear groupings as possible are obtained. Groupings which separate themselves from each other will suggest different lexical items, while groupings which shade into each other, even though opposite ends do not intercollocate, suggest one item with a wide **range**.

These proposals are extremely tentative, since they are based on the very limited results of studies of small clusters. It is not at all clear what sort of generalizations will emerge which can be applied to a text as a whole, and what use such generalizations will be in determining cases where one's intuition is helpless in deciding whether or not a form should be split. Nor will it be clear, in a monolingual study, what procedures are workable only because of the features of the particular language under study. But let us assume success, and a final, accurate description of the lexical items discernible in a text, in terms of their collocational patterns.

We will then know, or be able to find out:

(a) the total number of items in the cluster;
(b) a measure of the 'internal consistency' of the cluster, which can now be called **range**, that is, the chances of the collocates themselves intercollocating. The lexical item *vote*, for example, will probably have a wider range, over much the same ground, as *poll*;
(c) the variation in degree of significance between items in a cluster. We may assume that some items are only barely significant, and that others approach the cliché state. Clusters will differ in the range of the significance degree of the collocates;
(d) the distribution of items according to the degree of significance. This allows us to distinguish between clusters where the (c) figures are similar, but where the number of collocates at each degree of significance is different. (c) and (d) together measure the dependence of an item on its collocates.

In all our comments so far, we have taken the text more or less item by item. Pairs of items, perhaps, and clusters relative to a single node item. We have said nothing about the lexical structure of a whole text. Even when we have accumulated a vast number of facts, tiny details and generalized cluster-shapes, the total interrelations between all the items in our text are fragmented in such a way that it will be impossible to make clear statements on the basis of the processes already described. We have been forced by the nature of the material to unravel all the strands at once, without perceiving

which are the main lines of the organization. We must move towards a framework of some kind in which each item has as far as possible only one 'address'—its 'best fit' in the lexical structure of the language exemplified in the text. Most of our details will have to be discarded as soon as they have contributed towards underlining the general connections, and we must arrive at a list of what we call the **lexical sets** of a language. A lexical set is a discrete part of an organization of the lexical items of a text where each lexical item appears once only.

There are two assumptions made here. One is that the lexical organization will in fact divide into discrete parts. This is further discussed below. The other is that it will be possible to limit the appearance of an item to once only without undue distortion of its patterns. This is a problem with items like *put* and *thing* which may never show a strong enough tendency to one set rather than another. The idea of a 'setless' item may become a reality.

Although we are many years away from producing definite sets from a sufficiency of actual text, it is worth while discussing tentative procedures within the framework set up for the earlier part of the study. As in so many pieces of research, the very last results may well form the best start to applications of the research. Lexical sets parallel the categories of a thesaurus, but are linguistically arranged, and are the distillation of the massive evidence got, initially, from the study of collocations.

There seem two starting points from which sets can be got; the list of clusters, and the individual attraction of each pair of items. Starting from the clusters we can envisage:

(a) selecting the best list of actual clusters, by correlation of all possible combinations of clusters. Each set would then be the cluster of an item, and could have as a head-item the node of the cluster. There might be several possible lists which satisfied the criterion of maximum differentiation, and some of them might have little value. For example, if it happened that the clusters of *make* and *mandolin* together listed all the items in the text once only, this would be of merely passing interest.
(b) conflating the clusters and producing macro-clusters where two or more clusters share similar patterns (in effect extending the node from one item to more than one, without extending the span). This could be additive, that is the two (or more) clusters would simply be thrown together, inclusive of the items that were unique to each of them; or subtractive, where only the items common to the two clusters would be put together, and the unique items would be rejected, to find a home where they were in fact common to two or more clusters. So we might conflate the clusters of *post* and *letter*; if *write* occurred only with *letter*, and *parcel* only with *post* then the additive procedure would accept them both, while the subtractive procedure would attempt to find a better-fitting place for them.

Alternatively, starting from a huge matrix, where the power of each item over each other item was plotted, the work on Factor Analysis done by Shepard[20] might be applied, adapted to cope with the enormous amount of data. The 'distance' of one item from another might be represented by 1 minus its predictive power (on a scale from 0 to 1) and the position of each item relative to each other item in a multi-dimensional space would be plotted by using the distances as co-ordinates. Shepard's programme seeks to reduce the number of dimensions in the space with the minimum of distortion to the co-ordinates. Each resulting dimension would be a lexical set.

Such a piece of work is further complicated by a feature of lexis that we have only hinted at so far. We are interested in the lexical *meaning* of items as represented by their collocations, and it has already been shown that the number of times two items inter-collocate is not a direct measure of the meaning of either item, which must be based on the total frequency of the two items. The same collocation has a different significance to the items involved. Consider a collocation like:

a good omen.

It is of greater significance to *omen* that it occurs with *good* than it is to *good* that it occurs with *omen. Good* occurs so very often that *omen* should not feature large in its cluster, while for *omen*, a few items like *good, bad, propitious* will very frequently collocate. So, paradoxically, the distance between *good* and *omen* is different according to the viewpoint.

It is this feature which allows some morphemes and words to be frequent collocates of other items but never items themselves, and it is the incorporation of this feature which complicates the preparation of clusters, and forces us to describe an item as an 'address list' in the environment of other items rather than according to the patterns that emerge with itself as node.

Whichever method or combination of methods is finally chosen for set description, it is likely that a very large computer will be strained to the utmost to cope with the data. Pilot studies can be done up to the stage of arriving at the clusters of individual items, but there seems no accurate way of simulating the organization into sets.

Many questions have not been faced at all in this article. Of particular importance is the problem of language varieties or **registers**, where items, collocations and clusters may group themselves together according to features of the situation in which utterances are made. *Horse* and *hand* may not collocate significantly at all except in a register where utterances like

my smallest horse is thirteen hands

feature, and the item *hand* exemplified here will probably not emerge at all unless texts from this register are collected. *Vigorous depression* and *dull*

highlights[19] may stand as unusual collocations unless found in the registers of meteorology and photography respectively. The additional dimensions of register have been excluded from this discussion; we assume that our text is reasonably homogeneous, and particular stray register collocations will be rejected on a frequency basis. The rigorous description of registers is still a long way off, and it is probably a better approach to make a series of lexical descriptions of different registers and then search for their common elements.

The theory of lexis opens up exciting areas for describing language more accurately and more usefully. The practical problems are immense, and no secret has been made of them here, but the results that they promise are, possibly because of their novelty, no less fascinating than those of any other branch of linguistics.[20, 21]

Notes

1 The initial stages of a study of lexis on the lines described in this article have been made possible by a grant from the Ford Foundation to the University of Edinburgh, Most of the linguists there have given help to the study; particular acknowledgement for help in preparation of this article is due to Professor Angus McIntosh of Edinburgh, Professor M. A. K. Halliday and Mr. Robert M. W. Dixon of University College London, and Dr. P. J. Wexler of Manchester.

2 See pp. 3–15; also 'Categories of the Theory of Grammar', *Word* 17.3 (December 1961), pp. 273–277.

3 *Roget's Thesaurus of English Words and Phrases* (London, 1962 edition, Longmans, Green & Co.), paragraph 589.

4 For this and other technical terms, and discussions of lexis on which this article draws heavily, see: J. R. Firth, *Papers in Linguistics* 1934–1951 (London, 1957); M. A. K. Halliday, op. cit.; Angus McIntosh, 'Patterns and Ranges', *Lg* 37.3 (1961); T. F. Mitchell, 'Syntagmatic Relations in Linguistic Analysis', *TPS* 1958.

5 Commonly called **classes**.

6 cp. Angus McIntosh, op. cit., for an extended discussion of this point.

7 cp. pages 31–32.

8 Notes on this definition: (a) the term **formal item** is used rather than **stretch of language**, etc., to emphasise that, just as with grammatical items, a lexical item is not uniquely related to a piece of language substance; (b) I make no apology for misusing the term *morpheme* since this article indirectly throws doubt on its status. All I want it to mean is 'a stretch of language which is formally indivisible', and I wish to beg the question of whether it is a unit of grammar, or of lexis, or both, or whether two terms are required, one for grammar and one for lexis.

9 The use of *structure* here is by analogy from grammar. The analogy may not be a good one, and perhaps a neutral term such as **componence** will be preferred; it is used here to refer to that area of lexical analysis which corresponds to structure in grammar.

10 The solidus indicates 'new speaker'. The text is from a transcription of an impromptu conversation.

11 This distinction, which may not be obvious at first, is further discussed on page 32.

12 This is a general statement. It may happen, with examples like *driving rain*, that the *-ing* is important lexically. There is evidence that it is not always sufficient to consider the grammatical morphemes independently of their grammatical function. The item *quarry* which collocates with *chase*, *corner*, *hunter*, etc. will also collocate frequently with a grammatical class we could call **possessive**. The exponents could be *his*, *their*, *Bill's*, etc., and the lexical description would be more accurate if these varying exponents could be conflated.

13 A **form**, in this article, is a stretch of language which has not yet been assigned a lexical status. It is relevant to language form but its relevance has not been determined in detail.

14 e.g. *run up*, *run out of*, *run down*, *runaway*, *runabout*, *run to earth*, *run in*, *runaround*, *overrun*, *run aground*.

15 If the forms *dirt cheap* and *cheap as dirt* have identical clusters, then they would be regarded as instances of the same lexical item with variable sequence.

16 cp. Halliday, op. cit., p. 261, note 47.

17 It may be helpful to think of lexis as concerned with 'degrees of freedom' of items, a concept foreign to grammar, which makes a sharp distinction between free and bound.

18 There may be in English a tendency to avoid saying 'You've got cold feet' when one is not intending an occurrence of the polymorphemic item. If so, our work will be the easier for it.

19 These are genuine.

20 R. N. Shepard 'The Analysis of Proximities', *Psychometrika* 27.2, 3 (1962). I am indebted to Dr. Donald Michie for this reference. A recent unpublished paper by P. Leroy of Rennes 'On the Factor Analysis of Proximities' suggests that the prohibitive computational complexity of an extended Shepard programme may be avoided.

21 This paper was written at the outset of research in this field, which is now supported by a grant from the Office for Scientific and Technical Information. For a clarification and reformulation of some basic concepts as a result of more recent work see Paul van Buren, 'Preliminary Aspects of Mechanization in Lexis' (1965, mimeographed).

50

A LOOK AT THE ROLE OF CERTAIN WORDS IN INFORMATION STRUCTURE

Eugene Winter

Source: K. P. Jones and V. Horsnell (eds), *Informatics 3: Proceedings of a Conference Held by the Aslib Co-ordinate Indexing Group*, 1978, pp. 85–97.

Summary

In linguistics, *What do words mean*? is a key question underlying all discussions of meaning. Indeed, we have always recognised the need to distinguish between at least two categories of word: the *closed-system words* which have small finite vocabularies whose main function is the organisation of open-system words, and the *open-system words* which have large open-ended vocabularies (the creative end). We have also long recognised that there was a cline of the open- and closed-ness of words, but its implications for the semantic classiciation of words have been largely ignored.

The aim of this paper is to explore an interesting aspect of this cline of meaning and to show that there is a whole class of open-system words (lexical items like nouns, verbs and adjectives) which have much of the meaning of the closed-system in sentence connection. Such nouns, verbs and adjectives will be considered as an important part of the contextual grammar of clause relations.

Using Mrs. W. Dea's description of clause relations as our starting point, we will consider in some detail the role of certain of these lexical items in the organisation of groups of sentences in typical contexts. In particular, it will entail discussing how the lexical item may *anticipate* the relation between sentences which follow it in the context, what bearing this may have on the explicit connection of those sentence (pairs), and what implications the semantics represented by these lexical items may have for the study of information structure. (All examples are taken from written English).

51

THE LEXICAL ITEM

John Sinclair

Source: Edda Weigand (ed.), *Contrastive Lexical Semantics*, Amsterdam: John Benjamins, 1998, pp. 1–24.

1. Introduction

In the early days of computational lexicology, some forty years ago, it was felt important to distinguish between a 'word' and a 'lexical item'. This was a period for emphasising the complexity of language, and proposing abstract categories of language form to escape from the confines of surface phenomena; procedural models, like Immediate Constituent Grammar (e.g., Stageberg 1966:262ff.) were held in suspicion because position and sequence were non-negotiable. On the other hand, Firth (1957) made a distinction between 'sequence', the physical positioning of linguistic events relative to each other, and 'order', realised mainly by sequence, which was an abstract representation of language form.[1]

Hence an orthographic word was recognised as a string of characters lying between spaces, but the equivalent lexical unit had greater freedom, sometimes being more than a word, and possibly even less than a word in extent, with some variation and discontinuity. At the time it was felt that in most cases the physical and the formal categories would coincide exactly, but there was a small amount of superficial evidence to justify the distinction (Sinclair 1996b). Words such as *another*, *maybe*, *wherever* were so obviously concatenations of otherwise separate words that little phrases such as *in order to*, *as if*, *of course* could easily be interpreted as essentially the same structures with the word space(s) retained. Compounds and (in English) phrasal verbs were so important that large sections of the morphology – for compounds – and substantial digressions in the syntax – for phrasal verbs – were regularly devoted to these in descriptive publications. The phrasal verb resisted all efforts to accommodate it in a description where word and lexical item were fused.

Idioms and other phraseological conventions were regarded as important by the powerful language teaching lobby, but as little more than a nuisance

by grammarians and lexicographers. Is there any point in analysing *Don't count your chickens before they're hatched*? On the one hand, it looks like a well-formed sentence of two clauses, but on the other hand there seem to be hardly any alternatives to the succession of word choices, and a grammatical analysis in such circumstances has little value if one believes that meaning arises from choice.[2]

At this time, computers were just beginning to cope with text rather than numbers, and they were ill-equipped for the task, both in hardware provision and suitable programming languages. One of the few tasks they could perform reliably was the identification of orthographic words, and to leave aside the distinction between word and lexical item seemed at the time an innocent simplification in practice, since the distinction had been clearly recorded for future attention.

The simplification was not challenged; in fact it was supported by two other contemporary models of language description. One was the dictionary model, which for practical reasons – ease of consultation – if nothing else, has always been based on the rough equation of a word and a unit of meaning. Phrases and idioms warrant a note at the end of some entries, but there is no attempt at all to articulate a theory that allows for them, and for the relationship between the 'independent' and the 'dependent' uses of a word (Sinclair 1987a).

The other model of the period was the lexical component of a transformational grammar (e.g., Katz & Fodor 1963), which was clearly word-based. The representation of text, for example as the output of the phrase structure component, was as a string of words, each of which became an entry in the lexicon. Although the argumentation in support of this decision was not provided, the unrecognised assumption that the word is the unit of lexical meaning remains largely unchallenged.

As a result of this insensitivity to lexis, the last generation of linguistics has seen grammar going through many stages of sophistication, whatever the theoretical model of preference, while the unit of lexis has remained fixed and concrete, pinned to the surface of language. The apparatus of grammatical description has acquired many components, levels, categories and scales, while lexical description has nothing but feeble surface categories like word and collocation.[3]

The need for a more detailed and abstract model for lexical description became clear when lexical information began to be extracted from multimillion word corpora in the early eighties.[4] Several long-accepted conventions in lexicography were called into question – for example the idea that a word could inherently have one or more meanings. The working assumption was that when these meanings were explicated (or translated, in a bilingual dictionary) and, in the better dictionaries, exemplified, the lexicographer's job was done. This practice proved to be incapable of organising the strong, recurrent patterns that were shown by corpus analysis to

be present in the way words were used in texts; the importance of the surrounding language far outweighed the question of how many meanings and how they were related to each other.

A dictionary is a practical tool, and no place to introduce a new theory of meaning, even if one had been available. So the first dictionary derived from a corpus (Sinclair & Hanks 1987) was conservative in its design, in relation to the disturbing nature of the evidence encountered by the lexicographers. Despite the massive organising effect of cotext, pride of place in an entry was usually awarded to one of the meanings least dependent on the cotext. At one time in the design phase of Cobuild it was suggested that the normal pattern of an entry should be turned upside down, since the phrasal uses, which are habitually summarised crudely at the end of entries, if at all, were so important; but this was judged to be unhelpful for normal users, who might become very confused, no matter how theoretically sound the structure might be.[5]

As research progressed in the following decade, it became clear that the original distinction between a word and a lexical item was the key to a more helpful description of the vocabulary, and a more accurate account of meaning. The 'innocent simplification'[6] obscured two basic facts:

(a) many, if not most, meanings require the presence of more than one word for their normal realisation;
(b) patterns of co-selection among words, which are much stronger than any description has yet allowed for, have a direct connection with meaning.

2. Meaning

The establishment of a lexical item as an abstract category distinct from the word requires us to take meaning into account, and it is important not to overburden the argument with all the problems that the concept of meaning brings along with it. Only one problem will be tackled, a central structural problem that can be discussed without raising the others.[7]

The problem is as follows: in dictionaries, lexicons, thesauruses etc., meanings are linked primarily with words; how can the appropriate meaning be identified in a text, at the point where a word occurs? If a word has fifty different meanings in a dictionary, how is an occurrence of the word related to just one of those meanings? And what happens if none of the fifty meanings precisely covers the case?

It is contended here that the theoretical frameworks that linguists use are inadequate to solve this problem – or even to state the problem in such a way that a solution could be attempted. So there must be adjustments made to our theoretical perspectives, after which the problem can be restated.

The initial statement of this problem of meaning has some similarity with Chomsky's well-known view of a grammar as a finite set of rules that specifies the non-finite set of sentences of a language. Chomsky was concerned primarily with syntactic well-formedness, but the same problem arises with the lexis. The problem in both cases is how to relate a finite resource to an unlimited set of applications; in the case of syntax the set of rules is finite and the set of sentences is not; in the case of lexis the set of meaningful items is finite and the set of meanings in use does not appear to be limited.

Chomsky's solution to the problem of linking the rules and the sentences was to introduce recursive rules, but recursion is not a useful concept in the lexicon. Recursion is one of the simpler types of combination; a single rule-form in the grammar[8] provides the link between finite and unlimited sets. For the combinatorial relations of the lexicon, a more complex relationship needs to be defined. That is the purpose of this paper.

A lexicon, or a dictionary, consists of a list of words, to each of which is attached a number of statements, features etc., which together express what can be said about the meaning of the word. The words are arranged according to the rules of grammar to make text.[9]

The problem is that a text is a unique deployment of meaningful units, and its particular meaning is not adequately accounted for by any organised concatenation of the fixed meanings of each unit. This is because some aspects of textual meaning arise from the particular combinations of choices at one place in the text, and there is no place in the lexicon-grammar model where such meaning can be assigned. Since there is no limit to the possible combinations of words in texts, no amount of documentation or ingenuity will enable our present lexicons to rise to the job. They are doomed.

However, the meaningful combinations can now be described in new ways which make them much more tractable. At the time lexicons of the familiar kind were being designed, much of the regularity of the combinations was obscured by the inadequate means of observation of the data. It was impossible to gather together a sufficient quantity and selection of data in which the underlying patterns could be identified. The application of powerful computers to large text corpora has begun to improve our methods of observation.

3. Reversal

To give just one example of the inadequacies of a lexicon built by established methods, they do not take into account the common phenomenon of semantic reversal.[10] Situations frequently arise in texts where the precise meaning of a word or phrase is determined more by the verbal environment rather than the parameters of a lexical entry. Instead of expecting to understand a segment of text by accumulating the meanings of each successive meaningful unit, here is the reverse; where a number of units taken together

create a meaning, and this meaning takes precedence over the 'dictionary meanings' of whatever words are chosen. If the two meanings (the one created by the environment and the item, and the one created by each item individually) are close and connected, the text is felt to be coherent; if they do not, some interpretation has to be made – perhaps the meanings of the items are neutral with respect to the semantic demands of the environment, perhaps there is a relevant metaphorical interpretation, or an irony, or a very rare meaning of an item, or a special interpretation because of what the text is about at this point.

Whenever the meaning arises predominantly from the textual environment rather than the item choice, it is considered to be an instance of semantic reversal. The flow of meaning is not from the item to the text but from the text to the item.

In practice, the flow is rarely in one direction only. The textual environment will nearly always have some effect on the meaning of a unit, and the accepted features of the meaning will not often be totally ignored.[11] So the problem of the description of meaning can be traced back to the rigidity and frequent irrelevancy of the lexicons that supply the meanings. Not only do they (a) only supply some of the meaning, but also (b) they often supply meaning components that do not fit the particular environment; in addition (c) there is no mechanism that I am aware of for adapting a lexical entry to suit a particular recurrent set of circumstances, far less an individual instance.

The effects of reversals can be seen in dictionaries and lexicons when a word is frequently found in collocation with another, and this has an effect on the meaning. For example, *white wine* is not white, but ranges from almost colourless to yellow, light orange or light green in colour. That is to say, the meaning of *white* when followed by *wine* is a different colour range than when it is not.

Traditional dictionaries tend to obscure this point by using encyclopedic information to explain the meaning, for example:

> (of wine) made from pale grapes or from black grapes separated from their skins (Collins English Dictionary, 1991).

This assumes that the user already knows roughly what colour *white* is when collocated with *wine*.

Such examples are familiar enough. But when the word *holy* is interpreted as an abnormal mental state, as in the example *The ambience borders on the holy* . . . we must assume that this semantic feature is assigned not by any lexicon. There is no lexicon that sanctions such a meaning, and indeed, if there were, it would be a distraction with reference to most of the occurrences of the word *holy*. The meaning is created in the collocation with *borders on*. Whatever follows this phrase indicates the limits of normality by

specifying a mental state that lies just outside normality. When the adjective is *obsessional*, the feature of abnormality is already present in the meaning of the word, and the co-ordinated choice will be felt to be coherent; in the case of *holy*, the required feature has to be added by reversal. And if there were an instance of *borders on the normal*, this would be interpreted as fully ironical, suggesting that the normal is unexpected.

The way in which such a semantic problem is tackled is parallel to (or even identical with) a wide range of potential difficulties of interpretation; for example if in a conversation someone says *Wasn't that awful what happened to Harry*?, and if you, as the receiver of this query do not know a Harry to whom something awful has happened (or you know more than one unfortunate Harry), you have two basic courses of action. You can either challenge the presupposition that there is no identification problem, by pointing out your ignorance and/or your interlocutor's lack of clarity, or you can use a reversal technique – try to pick up from the following conversation what you need to know. Or if you are reading a book that is outside your normal area of expertise, and come across items that the author assumes you know the meanings of – like abbreviations – you can either break off in your reading and consult a specialised glossary, or plough on with whatever understanding you can glean from reversal techniques. After a while either your interpretation will break down altogether or you will survive the particular passage because none of the unresolved meanings are critical to your overall understanding.

4. Theory adjustment

'Reversal' is one of the new descriptive categories that will be required to account adequately for the data. The problem goes deeper, however, and requires some reorientation of theory. As a start, here are three hypotheses that, in various ways, cut across established theoretical norms and assumptions. Without the acceptance of these hypotheses, or their refutation, progress towards a better account of meaning will be slow.

I Language text is not adequately modelled as a sequence of items, each in an environment of other items.

We normally accept an underlying model of language as 'item-environment'. At any point in a text we can interpret the occurrence of an item in terms of what other choices are possible, given the environment. Hence each item is both an item in its own right and a component of the environment of other items.

This model must be examined carefully, because it seems inherently implausible. Each item would have to be interpretable, simultaneously, as having many different meaning-relations with other items, equally

multifaceted. As each item came to be processed for its contribution to the meaning of the text, every other item nearby would change not only its meaning, but the basis of its meaning – whether central or peripheral to the node in focus.

This would lead to a huge multiplicity of meanings, and the need for a complex processor to relate them to each other, discard irrelevant ones, etc. Further, since we have no reason to believe that the interpretation will proceed on a strict linear basis, the model would become extremely complex.

Such complexity reflects the interdependence between words and their environments, and makes it clear that all the patterning cannot be described at once; some elaboration of the model is needed in order to disperse the complexity. This could take the form of a diversification of units into different types, as grammar has nouns, verbs etc. – we already recognise 'grammatical' and 'vocabulary' words, without being able to distinguish them formally, and we use equally informal terms such as 'idiom', 'figurative' and 'metaphorical' in lexicography. It could also take the form of erecting a hierarchy of units, a *rank scale* in Halliday's (1961) terms, where structural patterns of different types and dimensions are arranged in a taxonomy.

Exactly what model will emerge is not possible to predict at present; my aim in this paper is to establish the need for it, and to put aside some well-respected assumptions about language that may be hampering our thinking – such as the imbalance at present in favour of the independence of the word, rather than its interdependence on its 'cotext', or verbal environment. To put these notions aside is not to discard them for ever, or to attack their intellectual integrity, or the competence of those who uphold them; it is merely to suspend their operation temporarily to see if the picture that emerges from a rearrangement is more satisfying than the present one.

This point is particularly important with respect to the second hypothesis that I would like to put forward:

II Ambiguity in a text is created by the method of observation, and not the structure of the text.

If a word is likely to be intricately associated with the words that occur round about it, then the consequences of studying its meaning in isolation are unpredictable. Dictionaries, which have little choice than to organise their statements of meaning around the word, present a picture of chaotic ambiguity. Words have many meanings, and there is no way of working out in advance which one is appropriate in a text.

However, if we extend the viewpoint to two or three words (which is normal when lexicographers recognise a relatively fixed phrase) much of the ambiguity drops away.[12] People use this extended viewpoint so naturally in

reading and listening, and language teachers labour the importance of concentrating on the broad aspects of meaning and not the particulars of a single word.

Despite almost universal accord with the position that the environment of occurrence is important in text structure, every machine lexicon I know persists in starting with the inappropriate unit, the word. When such a lexicon is applied to a text, all the possible meanings of a word are listed, including whatever phraseological meanings have been noticed, as if all were potentially relevant on each occasion of the occurrence of the word. Having created quite fictitious ambiguities, the researcher then multiplies them with similarly complex possibilities for the next word, and the next . . . leading to the most innocuous sentence having many thousands of possible 'meanings'.

Here is a superficial example.[13] In Collins English Dictionary, words are assigned a number of meanings as follows, approximately:

cat	24	sit	18	on	25
mat	17	the	15		

The number of possible combinations of meaning, on these figures and multiplying *the* in twice, for *The cat sat on the mat* is thus 41,310,000.

In this desperate situation, there arises a need for sensitive algorithms to filter out all the meanings that do not apply in the particular instance, all except one. The mess is so serious by now that this cannot be achieved by automatic process alone; humans must be trained and employed to clear it up.

This process must be compared with the normal linguistic activity of an ordinary person. All day long, effortlessly, this person interprets passing sentences, usually correctly, and often against a background of high levels of distraction. He or she is not even aware that any of them are potentially ambiguous.

The discrepancy that we perceive between human and machine behaviour is so gross that the model behind the machine's performance must be questioned. The human, perhaps, works with a better notion of meaningful units, and does not encounter ambiguity.[14]

This position casts doubt on the relevance to language study of all the work over many years in Artificial Intelligence and Natural Language Processing which concentrates on the resolution of ambiguity. The ambiguity that is studied is evidenced in carefully contrived short utterances, often of a kind which would be very unlikely to occur in texts. This possible objection to them is dismissed on the grounds that grammar deals with potential utterances, and whether or not an utterance actually occurred is uninteresting.[15] Now that large corpora can provide, with growing reliability, statements of regularity and norms in language usage. the marginal status of the work on artificial ambiguity can be clearly seen.[16]

This is not the place for a thorough examination of ambiguity and associated phenomena; the aim here is merely to open a case for reorientation of our attitudes to a very firmly established viewpoint. A representative account of the current NLP positions on the topic can conveniently be found in Monaghan (1996b). Most of the types presented there can be classified as one of:

(a) Created by the observer's perspective – this is the commonest;
(b) suggested by generality or vagueness of reference; one of the strengths of human language is its ability to avoid having to discriminate (Channell 1994);
(c) created by reference to a formally marked distinction in another language – a genuine problem for translators but not an ambiguity in any one communication system;
(d) created by a grammar which lays claim to too much potential meaning
    (i) combinations which might be contrastive if they occurred, but are inhibited by other factors;
    (ii) contrasts which are not formally distinguished in the language system and therefore can only be a matter for interpretation outside the system.

In other words, none of these count as ambiguities that have to be resolved in language description. Indeed, it can justifiably be claimed that a model of language is inappropriate if it obliges the description to make distinctions in a particular text segment which are not necessary for the interpretation of the text as a whole.

III The form of a linguistic unit and its meaning are two perspectives on the same event.

At first sight, most scholars in a structuralist tradition would not see anything exceptional in this statement. It is accepted that form and meaning are very closely related, and that variation in one normally leads to variation in the other. Even for those whose viewpoint leads them to perceive ambiguity, the close alignment of form and meaning would not be undermined; one would have to accept a lot of overlap, which is not the same as either a confusion between categories or a fusion of them.

Indeed in retrospect it is clear that the association between form and meaning has had to be somewhat loose, to allow for such notions as transformation (Harris 1957). As has been argued in the case of words, syntactic patterns may seem to vary independently of the meaning, as long as the cotext is kept to a minimum. So if active and passive constructions are presented bereft of cotext, their similarity of meaning ("who did what to

whom") is highlighted, and their differences, which show in the higher organisation of the discourse, are not obvious at all.

The position adopted in this paper is intended to be in sharp contrast to the approximateness of the traditional view. It is asserted that form and meaning cannot be separated because they are the same thing. Considered in relation to other forms, a lexical item is a form; considered in relation to other meanings, it is a meaning.

It follows from this tighter statement that ambiguity must in practice be very close to zero, or the statement would have to be seriously weakened. Also, the form of a lexical item must include all the components that are realised in the example. Meaning cannot inhere satisfactorily in just a selection of the components of an item when there are other components left in the cotext, but requires them all to be assembled together, and a way of stating the structure of an item has to be devised.

It follows from the requirement that all the components of a lexical item must be included in its specification, that these genuine meaning-bearing items will have very little connection with their cotexts; all the choices that depend substantially on other choices will be grouped together in the item, and the text will be represented essentially as a succession of relatively large-scale and independent choices.

No doubt the reality will be a good deal more complicated than this sketch supposes. Discontinuity of lexical items is a strong possibility, and various kinds of embedding cannot be ruled out. The method of work, which is based on studying and processing concordances of words and phrases, draws out and highlights those choices that contain an element of coselection or conditioned selection. Other choices, which may also be structural, are less obvious in the data studied at present because they are sporadic with reference to the pattern of choices.

5. The axes of patterning

In order to restate the problem of meaning, we must draw some general implications from the three hypotheses that have been discussed above. The main one, which pervades the whole argument, is that the tradition of linguistic theory has been massively biased in favour of the *paradigmatic* rather than the *syntagmatic* dimension. Text is essentially perceived as a series of relatively independent choices of one item after another, and the patterns of combination have been seriously undervalued.

It is easy to understand how this has happened; once again it depends on the nature of the observations and the stance of the observer. The difficulty has been how to cope with the large range of variation that is apparent in most uses of language. In presenting structure, traditional linguistics puts most of the variation to one side through the device of separating grammar and semantics at the outset. This then obscures most of the structural

relevance of collocation, and removes any chance of the precise alignment of form and meaning. It also presents the semantic level with the kind of problems that this paper is discussing.

The opportunity to observe recurrent patterns of language in corpora has shown how choices at word rank co-ordinate with other choices round about in an intricate fashion, suggesting a hierarchy of units of different sizes sharing the realisation of meaning. The largest unit will have a similar status to the sentence in grammar (and may coincide with sentence boundaries in many instances) in that it will be relatively independent of its surroundings with respect to its internal organisation (see the distinction between *rank* and *level* in Halliday 1961).

Meaning appears to be created by paradigmatic choice; this is within the orthodoxy of most theories, whether or not it is explicit. This perception also relates meaning to the information of Information Theory. However, the mechanism of paradigmatic choice is so powerful that constant vigilance should be exercised to make sure that it is not misapplied. Sometimes in the actual use of language there is less choice than the paradigm is capable of creating, as in the example of counting chickens given earlier; in these cases to present the paradigm unqualified is to distort the description by claiming more meaning in an expression than is actually usable.

It seems that this happens as a matter of routine in most published descriptions, and that a language is characterised as having hundreds of thousands of meanings that are not in fact available, because they are constrained by the need for other choices in the environment. By giving greater weight to the syntagmatic constraints, units of meaning can be identified that reduce the amount of meaning available to the user to something more like his or her normal experience; the balance between the two dimensions will more accurately represent the relation between form and meaning.

Such a conclusion calls for nothing less than a comprehensive redescription of each language, using largely automatic techniques. Problems remain, particularly one concerning the inability of the paradigmatic and the syntagmatic dimensions to relate to each other. They have no contact with each other, they are invisible from each other, and to observe one, the other has to be ignored. The phenomenon is similar to the observational problems that led to Heisenberg's famous Principle of Uncertainty in atomic physics. An atom can have both position and momentum, but these cannot be observed simultaneously, because the techniques for observing one cut out the possibility of observing the other. Similarly, a word gives information through its being chosen (paradigmatic) and at the same time it is part of the realisation of a larger item (syntagmatic); in order to observe either of these, however, we lose sight of the other. Unless the requirements of the cotext are precisely stated, the word as a paradigmatic choice will be invested with far too much independent meaning; on the other hand when observed purely as a component of a larger syntagmatic pattern, it can have very

little freedom, and therefore can give very little information; it might be no more meaningful than a letter in a word, serving only the purpose of recognition. A means must be found of relating the two dimensions in order to give a balanced picture.

We are now in a position to restate the problem of meaning in tractable terms by means of the following hypothesis:

IV The meaning of a text can be described by a model which reconciles the paradigmatic and syntagmatic dimensions of choice at each choice point.

The model is set out in a preliminary fashion in Sinclair (1996a). Five categories of co-selection are put forward as components of a lexical item; two of them are obligatory and three are optional. The obligatory categories are the *core*, which is invariable, and constitutes the evidence of the occurrence of the item as a whole, and the *semantic prosody*, which is the determiner of the meaning of the whole, as we shall see in the example below. The optional categories realise co-ordinated secondary choices within the item, fine-tuning the meaning and giving semantic cohesion to the text as a whole.

The optional categories serve also as a means of classifying the members of a paradigm, and thus the two axes of patterning, the paradigmatic and the syntagmatic, are related; the relationship is in principle capable of automation, and is quantifiable. The three categories that relate words together on either dimension are *collocation, colligation*, and *semantic preference*. The first two are Firth's (1951, 1957) terms.

Collocation (at present) is the co-occurrence of words with no more than four intervening words[17] (given the arguments of this paper, the word is no more reliable as a measure of the environment than it is as a unit of meaning, so this measure will have to be revised, but it is at present the only measure in general use). On the syntagmatic dimension, collocation is the simplest and most obvious relationship, and it is fairly well described. On the paradigmatic dimension it is defined rather differently, because items can only collocate with each other when present in a text, and two items in a paradigm are by that arrangement classed as mutually exclusive. The relationship is that of *mutual* collocation, i.e., that they both collocate (on separate occasions, usually) with the same item or items. So whereas *manual* and *restoration* are both significant collocates of *work*, they themselves do not co-occur significantly.

Colligation is the co-occurrence of grammatical phenomena, and on the syntagmatic axis our descriptive techniques at present confine us to the co-occurrence of a member of a grammatical class – say a word class – with a word or phrase. Colligation as a paradigmatic concept is displaced, like collocation, to that of a mutual relationship; so a possessive may colligate with a particular noun, and the so-called 'periphrastic' construction, *the . . . of . . .* may occasionally occur as an alternative. So *your* (etc.) *true feelings* is the norm, but *the true feelings of people* is an example of the less common structure in this phrase.

Semantic preference is the restriction of regular co-occurrence to items which share a semantic feature, for example that they are all about, say, sport or suffering. This feature is relevant in the same way to both syntagmatic and paradigmatic phenomena.

The three categories are related to each other in increasing abstraction; collocation is precisely located in the physical text, in that even the inflection of a word may have its own distinctive collocational relationships. To observe colligation one has to assign a word class to each word under examination; where there is a preponderance of one particular word class, this is colligation. Within the abstraction of the word class, of course, there may be one or more collocations.

Semantic preference requires us to notice similarity of meaning regardless of word class; however there may well be found within a semantic class one or more colligations of words which share both the semantic feature and a word class. There may also be collocates, specific recurrent choices of word forms carrying the semantic preference.

6. Example

The word *budge* in English poses a problem for dictionaries; for example:

> to (cause to) move a little (Longman Dictionary of Contemporary English)

Leaving aside the tortuous syntax of this sort of definition, it is easy to demonstrate that there is no feature of the meaning of *budge* that restricts movement. Budging is the overcoming of a resistance to movement, and so it concentrates on the beginning of movement. Even a little movement constitutes a budging, which might explain the definition above, without justifying it, because something once budged (i.e., set in motion) might move a great deal.

The point is that English does not talk much about budging at all, but about *not* budging, where the quantity of movement is irrelevant. The two examples that follow the definition above are indeed both negative, but the entry reads as if the lexicographers had not noticed this primary fact of usage, and the user, of course, has no idea how to evaluate the presence of the negatives since they are not referred to.

It would be difficult to find an instance of this word which is semantically positive. The figure (see next page) gives 31 instances from a corpus, all that there are in almost 20 million words. Of these I propose to ignore line 22, which is clearly written to represent the dialect of a region. Most of the

1. ight be out of his mind and refuse to budge. In that case, the Vice-President
2. ergencies. But Mr Volcker has yet to budge on changing his controls over domest
3. off scrubbers' hands before it would budge. It was rumoured to be make-work to
4. to do so, but she knew she could not budge me from my view. We spent several v
5. he recognizes it, he'll refuse to budge off that stool where he's sitting n
6. side, but still the snake will not budge. He keeps banging it on the head wi
7. away louder than ever. I wouldn't budge either, or come back, till a boy w
8. now. We won't none of us be able to budge tomorrow. "They sat at their tea
9. blow. The virus fanciers refused to budge. Whatever the diagnosis, my recove
10. sat in a corner; I determined not to budge from it until closing-time. I also
11. hen neither death nor? disease could budge her. She wrote a cheque for more th
12. it with my shoulder, but it will not budge. I go to the backdoor. I find that
13. ng the following months and would not budge – "What's done can not be undone
14. ooden door of the museum. It didn't budge. Hastily, I looked round for a bel
15. another snail near him he refused to budge, even in the mating season. I ofte
16. me into the dining room, refusing to budge, so that no one else budged, and s
17. It was a dismissal. Bonasera did not budge. Finally, sighing, a good-hearted
18. .9o caliber pezzonovante. You can't budge him, not even with money. He has
19. fternoons when the thermometer won't budge above minus twenty." "And those
20. be so heavy that two horses could not budge it even in moist earth. Although Wa
21. between the duellists and refuse to budge. Often to everyone's great relief
22. the coroner himself are gawn "t" budge on that. In the firrst place, d "
23. omise up to a point but he refuses to budge on design principles he knows to be
24. The humanity here just refuses to budge." "That's ridiculous," says
25. out of the packet. When it did not budge he shook it more fiercely like' at
26. ed at the doorknobs the doors didn't budge or even rattle. "Oh, my God!"
27. and hesitated. He knew he couldn't budge Ben Canaan. He walked to the alcove
28. at they might, the BritiSh would not budge from their immigration policy. In m
29. pressure any delegation. They won't budge from that position." "What a ti
30. the wings of the eagle and refused to budge . after three thousand years of wait
31. tried the idea on him. He wouldn't budge. He seemed to have already faded aw

indications of colligation with a negative are to be found to the left of the central, or 'node' word; immediately to the left we find eight instances of words ending in *n't* and eight of *not* – together making slightly over half the total. Most of the others show the word *to* in this position, and by examining the word previous to that, there is a strong collocation with forms of the lemma *refuse* – nine in all. Although not a grammatical negative, *refuse* can reasonably be considered as a lexicalisation of the kind of non-positive meaning that characterises *budge*.

There are, then, just five remaining instances that do not follow one of the three prominent ways of expressing negativity. The eighth line has a double negative in an extended verbal group, the tenth has *determined not to*, the eleventh Has a *neither/nor* construction. The two remaining instances show neither grammatical nor lexical negation; the second line expresses the refusal aspect with *has yet to* – implying that Mr Volcker refuses to budge, and the third line draws attention to a presumably long and unpleasant period preceding eventual budging (the extended cotext of this line is *so deep with caustic dirt that skin would come off scrubbers' hands . . .*).

The negative quality of the phrase centred around *budge* is thus expressed in different ways, but with a predominance of collocations *refuse to* (and inflections), *wouldn't, didn't, couldn't*. Colligation is with verbs, with modals (including *able to*) accounting for half the 30 instances.

From this point I will not attempt to describe comprehensively the two instances above that imply rather than express negativity (lines 2 and 3). I will include them when they conform to a choice pattern, and ignore them silently when they diverge. They are sufficiently conformant with the general usage of *budge* to be given low priority, so little will be gained by an exhaustive account of their minor deviations; but also I am concerned to establish a methodology that concentrates in the first instance on recurrent events rather than on unrepeated patterns. When the habitual usages of the majority of users are thoroughly described, we will have a sound base from which to approach the singularities, which may of course include much fine writing.

Budge is an ergative verb, in that whatever is to be moved may figure either as subject or object of the verb; subject in an intransitive clause, and object in a clause where the subject is the person or thing making the attempt to move. A guide to these alternatives can be found by looking at whatever immediately follows *budge*. Intransitive clauses may well end with the verb, so where there is a punctuation mark following *budge* we may expect that the clause is intransitive. Twelve times there is a full stop, twice a comma and once a dash – fifteen instances in all, or half of the total.

The distinction between *won't* and *can't* draws attention to two different reasons why people or things do not budge, refusal or inability. Refusal is ascribed to whoever or whatever is not budging (*won't, wouldn't . . .* etc., and *refuse . . .*), while inability is usually ascribed to someone who is trying to get something or someone to budge (can be expressed by *can't, couldn't*). Of the first type, there are twenty instances expressing refusal or interpreted as implying it, all intransitive; where the subject is non-human and the verb modal (a snake, a quotation and a thermometer) we anthropomorphise (which is a kind of reversal). Of the second type, there are four examples; the non-budging is ascribed to inability, the 'agent' is in subject position, the clause is transitive and the person or thing that is not budging is named in the object.

One instance has indications of both possible reasons; the modal cluster is *won't be able to*, and the clause is intransitive. This is a prediction of a future inability to budge, and *won't* does not indicate refusal.

There remain five instances, of which four are *didn't* or *did not*. This usage is neutral with respect to refusal and inability; the structure is intransitive and so suggests that an agent is not important, but a person energetically trying to move a physical object is apparent in adjoining clauses in three of the cases. In the fourth the subject of *budge* is a person; the cotext makes it clear that he is under pressure to move, so it is closer to refusal than inability.

To summarise this matter, we note that twenty-five of the thirty instances are intransitive, with the item that does not budge as the subject; where this item is animate, it strongly collocates with *refuse*, and the semantic preference of refusal is found in most of the other instances, mainly through colligation with certain modals. In the transitive instances the non-budging item is object, the agent of movement is in the subject, the semantic preference is inability and there is strong colligation with the modals of ability.

A minor optional element of the cotext of *budge* is the expression of the position from which there is to be no budging. There are eight instances, all beginning with a preposition; *from* four times, *on* twice, and *above* and *off* once each. Most are of the 'refusal' type.

At this point we may argue that most of the patterning in the cotext has been accounted for, with the possible exception of the word *even*, which occurs four times to the right of *budge*, and links semantically with *yet*. Something fairly extreme is being referred to. We consider why people use this word, why they do not just use the common verb *move*, with which any use of *budge* can be replaced. Something does not budge when it does not move *despite* attempts to move it. From the perspective of the person who wants something moved, this is frustrating and irritating, and these emotions may find expression, because this is the 'semantic prosody' of the use of *budge*.

The semantic prosody of an item is the reason why it is chosen, over and above the semantic preferences that also characterise it. It is not subject to any conventions of linguistic realisation, and so is subject to enormous variation, making it difficult for a human or a computer to find it reliably. It is a subtle element of attitudinal, often pragmatic meaning and there is often no word in the language that can be used as a descriptive label for it. What is more, its role is often so clear in determining the occurrence of the item that the prosody is, paradoxically, not necessarily realised at all. But if we make a strong hypothesis we may establish a search for it that will have a greater chance of success than if we were less than certain of its crucial role. For example we can claim that in the case of the use of *budge* the user wishes to express or report frustration (or a similar emotion) at the refusal or inability of some obstacle to move, despite pressure being applied. Then there is an explanation for *even* and *yet*, and other scattered phrases from the immediate and slightly wider cotext of the instances. A selection of the evidence for pressure and frustration is given below, with reference to the figure.

(1) the President may be out of his mind
(2) should no longer . . . intervene . . . but . . .
(3) stained so deep
(4) she knew she couldn't
(5) <prediction based on experience>
(6) blows his pipe furiously

(7) especially if ordered to do so . . . indignant and thunderous
(8) <prediction based on present circumstances>
(9) the polio faction . . . the virus fanciers
(10) <decision based on prior events>
(11) neither death nor disease
(12) nudge it with my shoulder, but . . .
(13) stuck in his mind and . . .
(14) leant against the heavy wooden door
(15) even in the mating season
(16) naughty . . . ignoring requests . . . forced to . . . capricious
(17) It was a dismissal.
(18) not even with money
(19) <typicalisation of experience>
(20) two horses could not . . .
(21) thrust himself between the duellists and . . .
(22) [not considered]
(23) he may lose a client
(24) That's ridiculous
(25) he shook it more fiercely
(26) no matter how hard he tugged
(27) he knew he couldn't
(28) do what they might,
(29) <diplomatic pressure>
(30) no amount of arguing
(31) I tried the idea on him

The above quotations and remarks are taken from a wider cotext than that printed in the figure, but still a small one. The evidence is probably enough to convince many human readers that the prosody exists and is expressed, implied or alluded to in most of the instances. But the range of expression is apparently without any limit, and the amount of inference necessary to identify it as evidence is often great; moreover in several instances there is no actual piece of language that can be quoted, but a more general appeal is made to experience.

This amorphous collection is an unlikely starter for being related to a structural category, and yet the claim is made that it is the most important category in the description. Without a very strong reason for looking, a computer would find virtually no reason for gathering this collection, but if we can predict a structural place for it, then at the very least the computer could pick out the stretch of language within which a prosody should lie, and whose absence was as significant as its presence. There is good reason to believe, also, that as the number of instances available rises, so do regularities appear that were not reliably shown in smaller sets. The occurrences of *even* are already obvious, and the phrase *he/she knew* recurs even in this

small sample; there is similarity in the phrases *blow . . . fiercely, shook fiercely, hard . . . tugged*, and money is mentioned in the wider cotext of (11) (*Money would send her home when neither . . .*) as well as (18).

It is impossible to predict how much or what will be left over at the end of an extensive study of this usage, but it certainly looks unlikely to be neat and tidy; hence we need a clear and strong hypothesis about the nature and structure of the lexical item of which *budge* is the core, in order at least to search for indications of the semantic prosody. The core gives us the starting point, in the case of *budge* one that anticipates the prosody fairly clearly; the optional patterns of collocation, colligation and semantic preference bring out relevant aspects of the meaning, and the prosody can then be searched for in the close environment.

It is not surprising that this is a very common structure in language, because it allows the flexibility that was identified earlier in this paper as essential for an adequate lexical item. The prosody is normally the part of an item that fits in with the previous item, and so needs to have virtually no restriction on its formal realisation, whereas the core, often in the middle or at the end of an item, is buffered against the demands of the surrounding text so that it can remain invariable. An item of this shape and structure makes it possible for the lexicon to have finite entries which are adequate to describe the way the meaning is created by the use of the item.

In this lengthy description of the lexical item whose core is NEG *budge* I have not had reason to make a distinction that most lexicographers would regard as primary – the literal and figurative uses of the word. For example:

> (cause to) move very little, make the slightest movement; (fig) (cause to) change a position or attitude (Oxford Advanced Learners Dictionary)

It is easy enough to go through the examples and pick the eleven that show the figurative use, where views, opinions, policies, principles etc. occur instead of doors, stools and thermometers. Where the option to express position is taken, the preposition *on* seems to be restricted to the figurative use, while *from* occurs with both. In a few instances the distinction hardly seems necessary, because both the literal and figurative aspects of the meaning of other words are also relevant; for example the wider cotext of the first instance, about the President being out of his mind, includes the use of *office* and *sitting*; moving the President from his office simultaneously requires his physical relocation and the cancellation of his authority.[18]

7. Conclusion

I have tried to show that there is a seriously weak point in the automation of language description, in the design of a lexicon. Current models do not

overcome the problem of how a finite and rigidly formalised lexicon can account satisfactorily for the apparently endlessly variable meanings that arise from the combination of particular word choices in texts. I have suggested that the word is not the best starting-point for a description of meaning, because meaning arises from words in particular combinations.

The term 'lexical item', used to mean a unit of description made up of words and phrases, has been dormant for some years, but is available for units with an internal structure as outlined above. Elements in the surrounding cotext of a word or phrase are incorporated in a larger structure when the pattern is strong enough.

The lexical item balances syntagmatic and paradigmatic patterns, using the same descriptive categories to describe both dimensions.

The identification of lexical items has to be made by linguists supported by computational resources, and in particular large general corpora. The impact of corpus evidence on linguistic description is now moving beyond the simple supply of a quantity of attested instances of language in use. It is showing that there is a large area of language patterning – more or less half of the total – that has not been properly incorporated into descriptions; this is the syntagmatic dimension, of co-ordinated lexicogrammatical choices. Because of the great range of variation in realisation, the regularities of this dimension have been overlooked, whereas from the perspective of a computer they become both more obvious and easier to describe.

Acknowledgements

The instances of language in use that I quote in my work come from a number of sources, in particular The Bank of English in Birmingham and The British National Corpus. I am grateful for permission to make use of these corpora.

Some of the arguments in this paper were first put forward at the Second European Seminar of TELRI in Kaunas, Lithuania in April 1997, and a very short account was included in the Proceedings (Markinkevičiene & Volz 1997) under the title "The Problem of Meaning". I profited a lot from the discussion there and in Münster, and I would also like to thank Anna Mauranen and Mike Stubbs for their helpful comments on the written version.

Notes

1 The problem is still with us in the attempts to write grammars that are explicit enough to drive machine applications. Chomsky (e.g., 1965) managed to combine freedom from 'surface structure' with a procedural model of sentence generation. This was appealing on presentation, but when attempts were made a few years later to build in some of the complexities it exposed, the problem of the ordering of rules became one of the main preoccupations of the period; present-day grammars in this tradition are notoriously incapable of analysing open text.

2 Now that the behaviour of such phrases can be studied in large corpora, it is clear that there is a substantial amount of variation, but of a lexicosemantic or phraseological nature rather than a grammatical one. Of the 16 instances of *chicken* and *count* in a corpus of 200 million words, only one instance of the presumed canonical form is to be found. Two others have the *before* clause, and three have another timing expression; eleven have a possessive adjective in front of *chickens*, and there is one instance each of *no* and *any*; the remaining three are in the structure *of counting chickens*, which refers to the idiom rather than quoting it.

3 I do not refer here to the theories and descriptions of semantics, which have often been richly complex; however, since they are not systematically related to language in use, they are therefore not relevant to the present stage of this argument.

4 See Sinclair (1987c) for a detailed account of the way in which this problem was encountered and dealt with in lexicography. Chapter 4 is particularly relevant.

5 That early response to corpus data has since been repeated many times as more lexicographers encounter corpus evidence.

6 There was more than one parameter of simplification; for example the organisation of meanings around 'headwords' – lemmas in computational linguistics – carries an assumption that, by and large, the inflected forms of a word do not have distinctive meanings. This view is now regarded as rather suspect (Tognini-Bonelli 1995), and it is to be expected that a new generation of dictionary will arise where the indexing is through the form and not the lemma.

7 In particular, this paper does not address the general matter of what is meant by stretches of language; where it is necessary for presentation of the argument to indicate meanings or distinctions of meaning, the statements are quite informal. Only when the meaningful units of a language have been reliably identified will it be useful to examine this matter thoroughly, and then, since the links between form and meaning will not have been broken, the task of description will be different from current work in semantics.

8 One where the same symbol appears on both sides of the operator, like 'X→XY', meaning 'rewrite X as XY'. Obviously the second X can also be expanded, and introduces another X, and so on indefinitely. See Bach's (1964) treatment of recursion.

9 It is conceded throughout this argument that occasionally a multi-word phrase may be used as an item in the wordlist. However this raises a further problem – when two words appear together in a text, how do we know whether they realise one meaning or two? The process known as *tokenisation* in computational linguistics is a relatively straightforward matter when the orthographic word can be trusted, but expands infinitely as soon as multi-word units are recognised.

10 Reversal at a propositional level is introduced in Sinclair (1987b) but not given a name.

11 Presumably the often invoked and rather vague criterion for a compound or idiom (e.g., Cruse [1986:37]: "an expression whose meaning cannot be inferred from the meanings of its parts") applies to the more extreme cases of the effect of the textual environment.

12 This can be made clear in a bilingual context; see Sinclair *et al.* (1996:173–174) for a practical demonstration.

13 I hope nobody actually does this, because I have deliberately exaggerated the problem to show how close it is to absurdity.

14 Here as elsewhere one cannot make absolute statements. Accidental ambiguities, that may be irresolvable even when the viewpoint is extended to the limits of

practicality, are bound to arise, but very occasionally indeed; linguistic communication would be severely strained if it was more common than the sort of coincidence that happens once or twice in a lifetime. Ambiguity above the phrase – at propositional or pragmatic levels – is no more common than ambiguity at word level, but needs separate treatment which would not be appropriate here.

15 Monaghan (1996a:12) goes as far as to say that the cases where corpus evidence does not work "will probably be the most interesting and most crucial ones, since these will be the rarest in most corpora".

16 Marginal for the theory of how language works, that is. There may be relevance, for example, to research in cognitive psychology.

17 The calculation of the optimal size of the collocation *span* for English was first published in Sinclair, Jones & Daley (1970). At that time corpora were very small, so the figure was recently recalculated with reference to a much larger corpus, and with surprisingly little change. Oliver Mason programmed the recalculation, and calls it *lexical gravity*; it is a function in the CUE system of corpus query language.

18 W. E. Louw (personal communication) argues that 'literal' and 'figurative' are points close to the extremities of a continuum of 'delexicalisation'. Words can gradually lose their full lexical meaning, and become available for use in contexts where some of that full meaning would be inappropriate; this is the so-called figurative extension. Louw points out that a writer, especially a literary writer, must exercise vigilance so that the meaning of each word is interpreted at the intended point on the continuum. Such features as collocation are part of the control mechanism available to the writer. So in this example the use of several other words that could be interpreted literally keeps available the physical meaning of *budge*, while the overall interpretation of the passage will be institutional.

References

Bach, Emmon 1964. *An Introduction to Transformational Grammars*. New York: Holt, Rinehart and Winston.

Channell, Joanna 1994. *Vague Language*. Oxford: Oxford University Press.

Chomsky, Noam 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Firth, J. R. 1957[1951]. "Modes of Meaning". *Papers in Linguistics 1934–1951*, 190–215. London: Oxford University Press.

—— 1968[1957]. "A Synopsis of Linguistic Theory 1930–55". *Selected Papers of J. R. Firth 1952–59* ed. by F. R. Palmer, 168–205. London: Longman.

Halliday, M. A. K. 1961. "Categories of the Theory of Grammar". *Word* 17:3.241–292.

Harris, Zellig S. 1957. "Co-occurrence and Transformation in Linguistic Structure". *Language* 33:3.283–340.

Katz, Jerrold J. & Jerry A. Fodor 1963. "The Structure of a Semantic Theory". *Language* 39:2.170–210.

Markinkevičiene, Ruth & Norbert Volz, eds. 1997. *Language Applications for a Multilingual Europe*. Proceedings of the Second European Seminar, TELRI. Kaunas, Lithuania: Institut für deutsche Sprache, Mannheim.

Monaghan, A. I. C. 1996a. "Do we Need to Resolve Ambiguities?". Monaghan 1996b.1–16.
—— ed. 1996b. *CSNLP 1966* Dublin, Ireland. Natural Language Group, Dublin City University.
Sinclair, John M. 1987a. "Collocation: A progress report". *Language Topics: Essays in honour of Michael Halliday* ed. by Ross Steele & Terry Threadgold, 319–331. Amsterdam & Philadelphia: John Benjamins.
—— 1987b. "Fictional Worlds". *Talking about Text* ed. by R. Malcolm Coulthard, 43–60. Birmingham: ELR Monograph, University of Birmingham. (Rev. version, "Fictional Worlds Revisited". *Le Trasformazioni del Narrare* ed. by Erina Siciliani *et al.*, 459–482. Brindisi: Schena Editore, 1996.)
—— ed. 1987c. *Looking Up: An account of the Cobuild project in lexical computing*. London: HarperCollins.
—— 1996a. "The Search for Units of Meaning". *Textus* 9:1.75–106.
—— 1996b. "Beginning the Study of Lexis". *In Memoriam J. R. Firth* ed. by C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins, 410–430. London: Longman.
——, S. Jones & R. Daley 1970. *English Lexical Studies*. Birmingham: Report to the Office of Scientific and Technical Information.
—— & Patrick W. Hanks, eds. 1987. *Collins Cobuild English Language Dictionary*. London: HarperCollins.
——, Jonathan Payne & Chantal Pérez Hernández, eds. 1996. "Corpus to Corpus: A study of translation equivalence", *International Journal of Lexicography* 9:3.171–276.
Stageberg, Norman C. 1966. *An Introductory English Grammar*. New York: Holt, Rinehart & Winston.
Tognini-Bonelli, Elena 1995. "Italian Corpus Linguistics: Practice and Theory". *Textus* 8:2.391–412.

52

THE TEXTUAL PRIMING OF LEXIS

Michael Hoey

Source: Guy Aston, Silvia Bernardini and Dominic Stewart (eds), *Corpora and Language Learners*, Amsterdam: John Benjamins, 2004, pp. 21–41.

This paper sketches a theory of language that gives lexis and lexical priming a central role. All lexical items are primed for grammatical and collocational use, i.e., every time we encounter a lexical item it becomes loaded with the cumulative effects of those encounters, such that it is part of our knowledge of the word that it regularly co-occurs with particular other words or with specific grammatical functions. Priming also goes beyond the sentence, i.e., a lexical item may be primed (i) to appear in particular textual positions with particular textual functions, a phenomenon heavily influenced by text domain and genre, and (ii) to participate in cohesive chains. Because as individuals our exposure to language is unique, i.e., different from everybody else's, it follows that a word is primed for the individual language user. In other words, priming belongs to the individual and is constantly in flux. The theory of priming clearly has relevance for pedagogical issues, with important implications both for language learning, i.e., the way priming is to be tackled within the walls of the classroom, and for language production, in terms of routine and creativity.

Sinclair (e.g., 1991) has argued that the study of lexis leads to results incompatible with the descriptions provided by conventional grammars. Biber et al. (1999) have argued that lexical bundles characterize text types. Farr and McCarthy (2002) argue that the function of conditionals is specific to particular types of interaction. Morley and Partington (2002) argue that syntax is an epiphenomenon of lexis. All this suggests we need a new theory.

In this paper I want to put forward a theory of language that places lexis at its very centre and gives to vocabulary the pivotal status once awarded to

syntax. What I have to say is only a beginning – a mixture of the self-evident and the unproven. Much of what I am going to say will seem obvious but I want to build from a shared position to positions that may seem novel or wrong, though I shall defend those positions fiercely. The theory I shall briefly outline here has links with Brazil's work on the grammar of speech (Brazil 1995), with Construction grammar (e.g. Goldberg 1995) and with Pattern Grammar (Hunston and Francis 2000). It assumes the correctness of Halliday's interpersonal and ideational metafunctions (Halliday 1967–8) but rejects, and attempts to supersede, his account of the textual metafunction (while retaining the insights that Halliday's model provides).

The classical theory of the word is epitomized by those two central nineteenth and early twentieth century compendia of lexical scholarship – Roget's Thesaurus (1852) and the Oxford English Dictionary (Murray *et al.* 1884–1928). According to such texts, lexis can be described in terms of hyponymy and co-hyponymy, near synonymy and antonymy and has meaning(s) which can be defined using the lexical relations just mentioned. Every word, furthermore, belongs to one or more grammatical categories and has pronunciation, etymology, and history. According to the theory that underpins these positions, words interact with phonology through pronunciation, with syntax through their grammatical categories and with semantics through their senses; they find their place in diachronic linguistics through etymology. In such a theory, the lexical item is reactive to other systems, particularly those of grammar and phonology, and in some versions of the classical theory, the relationship between the word and the other systems has been so weak that grammar has been generated first and the words brought in as the last stage in the process (Chomsky 1957, 1965) or that the semantics have been generated first and the words seen as merely expressing the pre-existent meaning. Systemic-functional linguistics has an altogether more central place for lexis, but even in this model the systems can sometimes make it seem as if lexical choice is the last (because most delicate) choice to be made. Even where theory starts from the assumption that lexis is chosen first, or at least much earlier, the assumption is still that it passes through a grammatical filter which organizes and disciplines it.

I referred above to those great 19th century works of scholarship – the OED and Roget's Thesaurus. It is interesting that these works have outlived almost all the theoretical work (apart from that of Saussure's) from the same period. In the same way I am convinced that, when linguists look back at the 20th century, it will not be the grammatical theories that will be admired as permanent works of the highest scholarship but the corpus-backed advanced learners dictionaries, starting with Collins COBUILD and continuing with Oxford, Longman and Macmillan, and it is these works, and of course in particular the first Collins COBUILD Dictionary, that have shown the traditional view of lexis outlined in previous paragraphs to be suspect. In particular what these dictionaries and accompanying corpus-linguistic

work have established beyond doubt is the centrality and importance of collocation in any description of lexis.

Collocation no longer needs support but it demands explanation. The only explanation that makes sense of its ubiquity, and indeed its existence, is psychological in nature. Every lexical item, I want to argue, is primed for collocational use. By *primed* I mean that as a word is acquired through encounters with it in speech and writing, it is loaded with the cumulative effects of those encounters such that it is part of our knowledge of the word (along with its senses, its pronunciation and its relationship to other words in the same semantic set) that it regularly co-occurs with particular other words. As Sripicharn (2002) put it, "years of listening to people speaking make me know which words sound right together." Each use we make of the word reinforces the priming (unless our use runs counter to the priming it has received), as does each new encounter with the word in the company of the same co-occurring other words. Each encounter that we have with the word that does not reinforce the original priming either weakens that priming slightly or complicates it. A word may, and routinely does, accumulate a range of primings which are weighted in our minds in a variety of ways that take account of relative frequency, mode, genre and domain. Part of our knowledge of a word is that it is used in certain kinds of combination *in certain kinds of text*. So I hypothesize (supported by small quantities of data) that in gardening texts *during the winter* and *during the winter months* are the appropriate collocations, but in newspaper texts or travel writing *in winter* and *in the winter* are more appropriate; the phrase *that winter* is associated with narratives.

It follows from the processes involved in collocational priming that it is not in principle a permanent feature of the word. As new encounters alter the weighting of the primings, so they shift in the course of an individual's lifetime, and as they do so (and because they do so) words shift imperceptibly in their meaning or their use. I suspect that for many older linguists such a shift has occurred in the priming of the word *collocation* itself! Its collocations, post-Halliday and Hasan (1976) and pre-Sinclair (1991), were, I suspect, predominantly with the words *text* and *sentence*, rather than with *corpus* and *word*.

So collocational priming is context specific and subject to change. It is also, importantly, a matter of weighting rather than requirement. So the relatively rare phrase *through winter* is English as well as *in winter*. Priming belongs to the individual. A word is primed *for a particular language user*. A corpus cannot demonstrate the existence or otherwise of a priming for any individual. It can only show that a particular combination is likely to be primed for anyone exposed to data of the kind represented in the corpus in question.

If we accept these positions, as I think we must if we are to account for the existence and prevalence of collocation, we open the way to a more general recognition of the notion of priming. To begin with, the grammatical

category a word belongs to can be seen as its grammatical priming. Instead of saying "This word is a noun" or "This word is an adjective" I would argue we should say "This word is primed for use as a noun". In other words, the word is loaded with the grammatical effects of our encounters with it in the same way as it is loaded with collocational effects. If the encounters all point the same way, we assume 100% identification of the word with a particular grammatical category; this happens occasionally with collocation also (e.g. *kith* with *kin*). Nevertheless such total identification is not as common as we might imagine (Hoey 2003). Words such as *estimated* (V, adj), *teaching* (V, N, adj), *human* (N, adj), and *real* (as in *real nice, get real, real world, the real and the unreal*) are the norm. How, for example, might one categorize *red* in *a red sunset, the colour red, he went red or he saw red*? If we agree that words are primed for grammatical category, the question must be regarded as inappropriate.

As with collocational priming, grammatical priming can change through an individual's lifetime. Anyone British and over 50 is likely to have had the word *program* shift in its priming from noun to verb in writing. (With the alternative spelling *programme* or from an American perspective the priming shift will have been different.)

As with collocation, grammatical priming is context specific. In the conversation of homophobes, *queer* is primed as adjective and noun, but in the writings of cinema theorists, *queer* is primed as adjective only (*queer cinema, queer theory*). This means that the priming must be tagged for domain, purpose and genre. Again, more controversially, the priming is a matter of weighting not requirement. Margaret Berry once wittily said that you can verb any noun. Strictly this is not true – it does not apply to nouns derived from verbs in the first place – but her observation encapsulates a real fluidity in the language. So routinely do we adjective our nouns that we see it as entirely normal and label the use as a noun modifier or classifier, rather than admitting the protean nature of language; the Oxford Dictionary of Collocations however treats such usage as adjectival. (After all, in *a red sunset*, we would traditionally treat the word *red* as an adjective, despite its nominal use in *the colour red*.)

This grammatical priming does not necessarily assume the prior existence of any grammatical category. Sinclair's masterly analysis (1991) of *of* as belonging to a grammatical category with just one member in it is a warning about the assumption that grammatical categories are givens in the language. It could indeed be argued that what we call grammatical categories may be *post hoc* generalisations derived from the myriad individual instances of lexical priming that we encounter and take on board in the course of our language development.

One last point needs to be made about both collocation and grammatical categories, and it is a point that equally applies to those categories of priming I have yet to explicate. This is that primings nest. Thus *wing* collocates with

west, west wing collocates with *the*, and *the west wing* collocates with *in*. Similarly, *face* is in the first place primed for use as a noun or verb. Put into the phrase *in your face*, however, *face* loses the verbal priming. Once *very* is added, the latent ambiguity of *in your face* disappears, and so does the nominal priming. In its place the phrase *in your face* (in *very in your face*) is primed for adjectival use.

If we accept, at least for the sake of argument, that words are collocationally and grammatically primed, in other words if we accept that the learning of a word involves learning what it tends to occur with and what grammar it tends to have, it opens the door to the possibility of other kinds of priming. The first of these is semantic association, which in earlier papers (e.g. 1997a, 1997b) I referred to as semantic prosody (following or rather mis-following Louw 1993, and Stubbs 1995, 1996), and which Sinclair (1996, 1999) refers to as semantic preference. I would use Sinclair's term if it were not for the fact that I want to avoid building the term "preference" into one of the types of priming, since one of the central features of priming is that it leads to a psychological preference on the part of the language user. Also, the use of "association" is designed to pick up on the familiar "company a word keeps" metaphor used to describe collocation. The change of term does not represent a difference of opinion.

Semantic association is defined as occurring when a word is associated for a language user with a semantic set or class, some members of which are also collocates for that user. The existence of the collocates in part explains the existence, and in part is explained by, the semantic set or class in question.

As an example of semantic association, consider the verb lemma *train*, analysed in considerable detail by Campanelli and Channell (1994) and cited by Stubbs (1996). *Train* (in my corpus) collocates with *as a* and the resultant combination of words has a semantic association with the notion "skilled role or occupation". The corpus has 292 instances of *train* as a*, of which 262 were followed by an occupation or related role. The data included the following (numbers of occurrences are given in brackets).

train* as a teacher (25)
train* as a doctor (12)
train* as a nurse (11)
train* as a lawyer (11)
train* as a painter (8)
train* as a dancer (7)
train* as a barrister (5)
train* as a chef (5)
train* as a social worker (5)
train* as a solicitor (5)
train* as a braille shorthand typist (1)

train* as a concentration camp guard (1)
train* as a kamikaze pilot (1)
train* as a boxing second (1)
train* as a cobbler (1)
train* as a train waiter (1)

The combination *train* as a* has some clear collocates (*teacher, nurse, doctor, lawyer, painter, dancer, barrister, chef, social worker* and *solicitor*), as the frequency figures suggest. But it is hard to imagine evidence ever being available to support the idea that *braille shorthand typist* is a collocation of *train as a* – except, importantly, in a specialist corpus of, say, minutes of the Royal Society for the Blind. But its occurrence is still accounted for because of the generalization inherent in the notion of "semantic association".

Many semantic associations such as the one just given seem to be grammatically restricted. Although there are plenty of instances of *train* with "skilled role or occupation" in other combinations in my data, particularly as *teacher training*, the relationship is in part constructed through the structure given in the column of data above. This suggests that for some lexical items there might be restrictions that are not simultaneously instances of semantic association. These can be covered under another type of priming – colligational priming. The term "colligation" was coined by Firth, who saw it as running parallel to collocation. He introduced it as follows:

> The statement of meaning at the grammatical level is in terms of word and sentence classes or of similar categories and of the interrelation of those categories in colligation. Grammatical relations should not be regarded as relations between words as such – between 'watched' and 'him' in 'I watched him' – but between a personal pronoun, first person singular nominative, the past tense of a transitive verb and the third person singular in the oblique or objective form.
>
> Firth (1957: 13)

As put, it is difficult to distinguish his notion from that of traditional grammar. Interestingly, though, Firth's student, M. A. K. Halliday, used colligation in an apparently different way, and it is to be assumed that his use followed Firth's intention. This is how Halliday introduces colligation:

> The sentence that is set up must be (as a category) larger than the piece, since certain forms which are final to the piece are not final to the sentence. Of the relation between the two we may say so far that: 1, a piece ending in *liau* or *fe* will normally be final in the sentence; 2, a piece ending in *ši*$_2$, *ŋa, heu* or *sanhŋgeu*$_2$ will normally

> be non-final in a sentence; 3, a piece ending in *lai* or *kiu* may be either final or non-final in a sentence.
>
> Halliday (1959:46; cited by Langendoen 1968, as an example of Halliday's use of colligation)

Halliday here uses colligation to mean the relation holding between a word and a grammatical pattern, and this is how the term is currently used. For several decades it disappeared from sight with only the most occasional of references, and returned into use in papers from Sinclair (1996, 1999) and myself (1997a, 1997b). (We were not aware of each other's work, but, as we were colleagues for many years, it is more than possible that I picked up the notion from conversations with him without realising that I had done so. In any case, the earlier of the papers in which he discusses colligation predates mine by a year, so the credit for resurrecting this valuable concept must rest with him.)

One point to note about Halliday's formulation is that he formulates the colligational relationship in terms of sentential position. Thus colligation covers not only grammatical relations as conventionally understood but also such matters as Theme/Rheme position – and, I shall later argue, textual positioning too. If one considers the conventional grammatical statements one might make about the first two words of a clause such as

(1) The cat sat on the mat [fabricated, as if you did not know]

they include the following:

a *cat* is head of the nominal group in which it appears
b *The cat* is Subject
c *The cat* is Theme of the sentence.

In other words, we are capable of talking about a word's place in its group, the function that the group plays in the clause and the textual implications of its position. It should be no surprise therefore that colligations can take any of these forms.

I define *colligation* as

a the grammatical company a word keeps (or avoids keeping) either within its own group or at a higher rank.
b the grammatical functions that the word's group prefers (or avoids).
c the place in a sequence that a word prefers (or avoids).

My claim is that every word is primed to occur in certain grammatical contexts with certain grammatical functions and in certain textual positions, and this priming is as fundamental as its priming for collocation or semantic

association. I see connections between colligation as I am here describing it and the notion of "emergent grammar" referred to by Farr and McCarthy (2002). There are also clear parallels between the position here formulated and Hunston and Francis's pattern grammar (2000).

As an instance of the first type of colligation, the grammatical company a word keeps (or avoids keeping) either within its own group or at a higher rank, consider the word *tea*, which characteristically is strongly primed to occur as premodification to another noun, e.g.

tea chest
tea pot
tea bag
tea urn
tea break
tea party

It is also typically primed to occur as part of a postmodifying prepositional phrase, usually with *of*, e.g.

time for tea
a cup of tea
a pot of tea
a packet of tea
her glass of tea
nine blends of tea

Even the *Guardian* newspaper with which I work provides evidence of this despite the low occurrence of such items as *tea pot* and *tea set* (presumably because the mechanics of tea-making are rarely newsworthy). In my data *tea* occurs as premodification over a quarter of the time (29%) and as part of a postmodifying prepositional phrase just under 19% of the time. When it does not occur as premodification or as part of a prepositional phrase, it is often coordinated or part of a list, this accounting for almost 20% of such cases:

green tea and melon
tea and coffee
tea and sandwiches
tea and refreshments
tea and digestives
tea and toast
tea and scones
tea and sympathy
tea and salvation

It will be noted above that colligations can be negative as well as positive, and one of *tea*'s most obvious colligations is negative: it is typically primed to avoid co-occurrence with markers of indefiniteness (*a, another*, etc.). Just as with other primings, this is a tendency, not an absolute. There are 52 instances of *a tea* or *a . . . tea* in my data, just over 1% of instances, and one instance with *another*. Examples are:

(2) a lemonade Snapple and a tea, milk no sugar
(3) a tea made from the blossoms and leaves
(4) there was never a tea or a bun at Downing Street
(5) a Ceylon tea with a fine citrus flavour
(6) to enjoy a cream tea or a double brandy
(7) Oh I'll have a tea, two sugars, thank you very much for asking
(8) Another tea and I start dealing with the day's twaddle

Notice that four of the above examples are not interpretable as "a type of tea". So *tea* can occur with indefinite markers – it is not a matter of grammatical impossibility, nor a matter of a specific type of usage – but typically it is primed to avoid them.

It is worth noting, too, that this aversion to indefinite markers is not the result of its being a drink. In my data there are 390 occurrences of the word *Coke*, referring to the cola rather than the drug or the fuel. Of the 314 instances which refer to the drink, as opposed to the company that markets the drink, 10% occur with *a*. (I have no instances with *another*, though there are three occurrences along the lines of *a rum and coke*).

All of the above illustrate the first type of colligation. As an instance of the second kind of colligation, consider the following data (Table 1), where the clausal distribution of *consequence* is compared with that of four other abstract nouns.

It will be seen that there is a clear negative colligation between *consequence* and the grammatical function of Object. The other nouns occur as part of Object between a sixth and a third of the time. *Consequence* on

Table 1 Distribution of consequence across the four main clause functions in comparison with that of other abstract nouns.

	Part of subject	*Part of object*	*Part of complement*	*Part of adjunct*	*Other*
Consequence	24% (383)	4% (62)	24% (395)	43% (701)	5% (74)
Question	26% (79)	27% (82)	20% (60)	22% (66)	4% (13)
Preference	21% (63)	38% (113)	7% (21)	30% (90)	4% (13)
Aversion	23% (47)	38% (77)	8% (16)	22% (45)	8% (17)
Use	22% (67)	34% (103)	6% (17)	36% (107)	2% (6)

the other hand occurs within Object in less than one in twenty cases. To compensate, there is a positive colligation between *consequence* and the Complement function. Only one of the other nouns – *question* – comes close to the frequency found for *consequence*. The others occur within Complement four times less often than *consequence*. There is also a positive colligation between *consequence* and the function of Adjunct, *consequence* occurring here nearly half the time. The other nouns in our sample occur between around a quarter and a third of the time. I would conclude that *consequence* is characteristically positively primed for Complement and Adjunct functions and negatively primed for Object function. (Interestingly, this is not true for the plural form *consequences*, which routinely occurs as part of Object, supporting the argument of Sinclair and Renouf (1988) and Stubbs (1996) against too ready an adoption of the lemma as the locus of analysis.)

Consequence also illustrates the third type of colligation, in that 49% of instances in my data occur as part of Theme. Given that one would expect, on the basis of random distribution, that around 33% of instances would occur in Theme, this suggests that *consequence* is typically primed for this textual position.

The position we have reached is that lexis is primed for each language user, either at the word or phrase level, for collocations, grammatical categories, semantic associations and colligations. I do not however believe that there is any necessity to assume that priming stops at the sentence boundary. After all, the third kind of colligation, concerned with textual positioning, is an overt claim that priming has a textual dimension, in that choice of Theme is in part affected by the textual surround and therefore we are primed to use *consequence* to encapsulate the previous text, whether as Adjunct or Subject. We can take this point considerably further. In Hoey (2004) I argue that words may be primed to appear in (or avoid) paragraph initial position. So *consequences*, for example, is primed to begin paragraphs, but *consequence* is primed to avoid paragraph-initial position. I secretly hope that you respond to this information with the feeling that I am spelling out the obvious, in that it is obvious that we might start a paragraph with mention of a multiplicity of consequences and then spend the rest of the paragraph itemising and elaborating on those consequences and equally it is obvious that if there is only a single consequence it will be tied closely to what ever was the cause. If you were to so react, then that would be evidence that it is part of your knowledge of the words *consequence* and *consequences* that they behave in these textual ways, in short, that you were primed to use them in particular textual positions. It may be objected that *consequence* and *consequences* are exceptional in that they have long been recognized to have special text organising functions (e.g. Winter 1977). But the evidence suggests that textual colligation is not limited to a special class of words, nor is priming for positioning only operative at sentence and paragraph boundaries or in the written word only. As an example of spoken

priming, *the* has an aversion to appearing at the beginning of conversational turns (McCarthy, personal communication). As an example of textual priming at a level higher than the paragraph, take *sixty* which, when the group in which it appears is sentence-initial, is positively primed for text-initial position. In my newspaper data, 14% of thematized instances of *sixty* are text-initial. Given that the average length of texts beginning with *sixty* is 20 sentences, this means that *sixty* begins a text three times more often than would occur on the basis of random distribution.

Sixty begins newspaper texts for a variety of reasons, all of which are specific to the goal of newspaper production. In the first place, sixty is a majority in terms of percentage and therefore potentially newsworthy; a number of such texts begin *Sixty per cent of . . .* Newspapers are conscious of time and their place in time; a number of articles begin *Sixty years ago . . .* If an event affects sixty people, it may be a significant event; a number of articles begin with phrases such *Sixty spectators . . .* Examples are the following:

(9) Sixty per cent of adults support the automatic removal of organs for transplant from those killed in accidents unless the donor has registered an objection, according to a survey published yesterday.
(10) Sixty years ago Florida was the holiday home of the super-rich and the flamboyant.
(11) Sixty baffled teachers from 24 countries yesterday began learning how to speak Geordie as part of a three-week course run by the British Council on the banks of the Tyne.

The explanations I have given for sentences such as these, which are after all not lexical but discoursal in nature, might be thought at first sight to challenge the notion of priming, in that the choice of *sixty* would appear to be the product of external factors. This would however be to misunderstand the relationship being posited between lexical choice and discoursal purpose. In the first place, the text-initial priming of *sixty* does not extend to *60* (nor do many of its other primings – there is no association of *60* with vagueness, for example). So the choice of *sixty* over *60* is made simultaneously with one of the discoursal choices described above. Secondly, there is no externally driven obligation on a writer to place the phrase of which *sixty* is a part in sentence-initial position. In theory (rather than practice), news articles and stories could begin:

(9a) A clear majority of adults support the automatic removal of organs for transplant from those killed in accidents unless the donor has registered an objection, according to a survey published yesterday.
(10a) Florida was, six decades ago, the holiday home of the super-rich and the flamboyant.

(11a) Five dozen baffled teachers from 24 countries yesterday began learning how to speak Geordie as part of a three-week course run by the British Council on the banks of the Tyne.

Thirdly, and most importantly, I would argue that the text-initial priming of *sixty* for journalists and *Guardian* readers is the result of their having encountered numerous previous examples of *sixty* in this position. Consequently, *Guardian* readers do not expect, and journalists do not provide, articles that focus on the views of twenty-two per cent of a sample of interviewees, despite the fact that these views may be original or thought-provoking; readers and journalists do not concern themselves with what happened twenty-two years ago, even though time divisions are arbitrary as a way of talking about changes in the world and what happened twenty-years ago is, from some perspectives, as interesting as what happened sixty years ago. The possible effects of primings on the way we view the world are perhaps matters for critical discourse analysts to consider.

Even more than was the case with collocation, grammatical category, semantic association and colligation (of the non-textual kind), claims about textual colligation have to be domain- and genre-specific. The claim just made for *sixty* is palpably false for academic articles, for example; on the other hand, I would speculate that for the latter genre the word *recent* might be positively primed for text-initial position – *Recent research has shown . . .*, *Recent papers . . .* etc. For some purposes – lexicography, dictionaries of collocations, thesauri, comparable corpora – huge corpora representing a wide range of linguistic genres and styles are extremely useful. Resources like the Bank of English and the British National Corpus have huge value. But I believe, and what I am saying here and in the remainder of this paper provides grounds for believing, that homogenized corpora iron out and render invisible important generalizations – truths even – about the language they sample. For the purposes of identifying primings, specialized corpora are likely to be more productive. Gledhill (2000) shows that no corpus is too specialized: a mini-corpus of the introductions to cancer research papers revealed distinct differences from that of results sections.

Before we leave textual colligation, it is worth noting that it is sometimes the case that one priming only becomes operative when another is overridden. An instance of this phenomenon is the combination of *in* and an abstract noun, which has a strong negative priming for sentence-initial position. Once, though, the negative priming is overridden, a strong positive priming for paragraph-initial position becomes operative.

Textual position is not the only supra-sentential feature for which lexis appears to be primed. I want to argue that lexical items are also primed for cohesion. Certain words (e.g., *Blair*, *planet*, *gay*, and *genetic*) tend to appear as part of readily cohesive chains, whereas others (e.g., *elusive* or *wobble*)

Table 2 Cohesive tendencies of eight lexical items from *The Invisible Influence of Planet X*.

	Frequency in original text	*No of instances participating in cohesive chains across 50 texts*	*No of occurrences in single cohesive links not forming chains across 50 texts*
planet	23	36%	13%
Uranus	11	68%	6%
Pluto	10	84%	3%
Planets	8	66%	8%
week	1	32%	12%
wobble	1	10%	6%
wide	1	0%	8%
weakest	1	0%	2%

form single ties at best, and rarely if ever participate in extensive chains. In order to test this claim, I took a text (*The Invisible Influence of Planet X*) that I had previously analysed with respect to its cohesion (Hoey 1995) and identified the four lexical items that contribute most to the cohesion of the text. I then selected four items that appear only once in the text. For each of these eight items, I examined 50 lines of a concordance, moving in each case into the text from which the line was drawn and analysing the text in terms of the cohesiveness of the item under investigation. The results of this investigation are presented in Table 2.

It will be seen that there is a close correlation between the cohesiveness (or otherwise) of the items in the *Planet X* text and their cohesiveness (or otherwise) across a range of texts. All four of the items forming strong cohesive chains in *The Invisible Influence of Planet X* participate strongly in cohesive chains in other texts, such that between 36% and 84% of instances in the concordance were participating in such chains. Three of the four words that were not cohesive in the *Planet X* text also never or rarely participated in cohesive chains. The exception is of course *week*, which is only slightly less cohesive than *planet* in the corpus; it is of course predictable from the statistics for the four highly cohesive items that their priming for cohesion is on occasion overridden and this appears to be the case with *week* also.

Obviously the more cohesive an item is, the fewer the texts represented in the corpus (because an item that is repeated twenty times in a single text will generate twenty concordance lines), so we cannot simply read the results off the table, without further investigation, but the correlation is strong for all that. I hypothesize that when we read or listen we bring our knowledge of cohesive priming to bear and attend to those items that are most likely to participate in the creation of the texture of the text.

Furthermore, it is part of our knowledge of every lexical item that we know what type of cohesion is likely to be associated with it. So *Blair*, for

example, tends to attract pro-forms – *he, his, him* etc. – and co-referents – *the Prime Minister, the Labour Party leader*, while *planet* tends to attract hyponyms – *Mars, Venus, Pluto* – and *gay* favours simple lexical repetition. Grosz and Sidner (1986) and Emmott (1989, 1997) argue that cohesion is better treated as prospective rather than retrospective; the position presented here is in accordance with their view, in that encountering a named person such as Tony Blair in a discourse immediately creates in the reader/listener an expectation that the pronoun *he* and the co-referent *the Prime Minister* will follow (as well as simple repetitions of the name); Yule (1981) discusses the conditions under which one rather than the other might be chosen, and Sinclair (1993) discusses the mechanisms of prospection. In the terms presented here, we could say that *Tony Blair* is characteristically primed to create cohesive chains making use of some or all of pronouns, co-reference and simple lexical repetition (Hoey 1991).

The claim that some items are characteristically primed to be cohesive and others are characteristically primed to avoid participation in cohesion is supported by Morley and Partington's finding (2002) that the phrase *at the heart of* is non-cohesive. They found 29 instances of the phrase, each one from a different text. Again, as so often in this paper, the observation seems obvious, but it is the obviousness of the observation that most supports my case.

With the first kind of textual priming, we associated certain lexical items with certain textual positions (e.g. beginning of the sentence, beginning of the speaking turn, beginning of the paragraph). This was seen as a textual extension of colligation. The kind of textual priming we have just been examining – the cohesive priming of lexis – could likewise be seen as a textual extension of collocation, in that the characteristic cohesion of a word could be seen as an extension of "the company a word keeps". Analogy suggests that there should be a third kind of textual priming of lexis, associated with semantic association, and preliminary investigation supports the suggestion. In addition to being primed for textual position and cohesion, lexical items are, I argue, primed for textual relations. What I mean by this is that the semantic relations that organize the texts we encounter are anticipated in the lexis that comprises these texts. So, for example, *ago* is typically primed to occur in contrast relations, occurring in such relations in my data 55% of the time, and *discovered* occurs with (or in) temporal clauses 86% of the time. The word *hunt* is associated with a shift within a Problem-Solution pattern (Winter 1977; Hoey 1983, 2001; Jordan 1984) or a Gap in Knowledge-Filling pattern (Hoey 2001) 60% of the time; it is also associated with a move from past to present in 67% of cases. I hypothesize that this aspect of textual priming accounts for the average reader/listener's enormous competence at following and making sense of text in very little time. Earlier work on textual signals (Winter 1977, Hoey 1979) only scratches the surface of the signalling that the average text supplies, if this hypothesis proves to be correct; it is possible that evoked

and provoked appraisal (Martin 2000; on appraisal see also Flowerdew 2004) and its textual reflex (Hoey 2001) are also accounted for by this feature of priming.

It will be noticed that I talk of items being "typically" or "characteristically" primed. This is of course partly because priming belongs to the individual, not to the language, and so no blanket claim can be made about any word. It is also because, as noted earlier, all claims about priming are domain and genre specific. A claim that a particular lexical item is primed to occur text-initially or form cohesive relations is only valid for a particular narrowly-defined situation. Since my corpus overwhelmingly comprises *Guardian* newspaper, the claims made above about lexical priming are true of those (kinds of) data, but carry no weight, until verified, in any other situations. While Biber *et al.*'s notion of the "bundle" may be oversimplified, he and his colleagues are certainly right in saying that they occur in, and are true of, text types. I want to claim that the types of textual colligation I have been describing occur in all kinds of texts but the actualisation of these colligations varies from text type to text type.

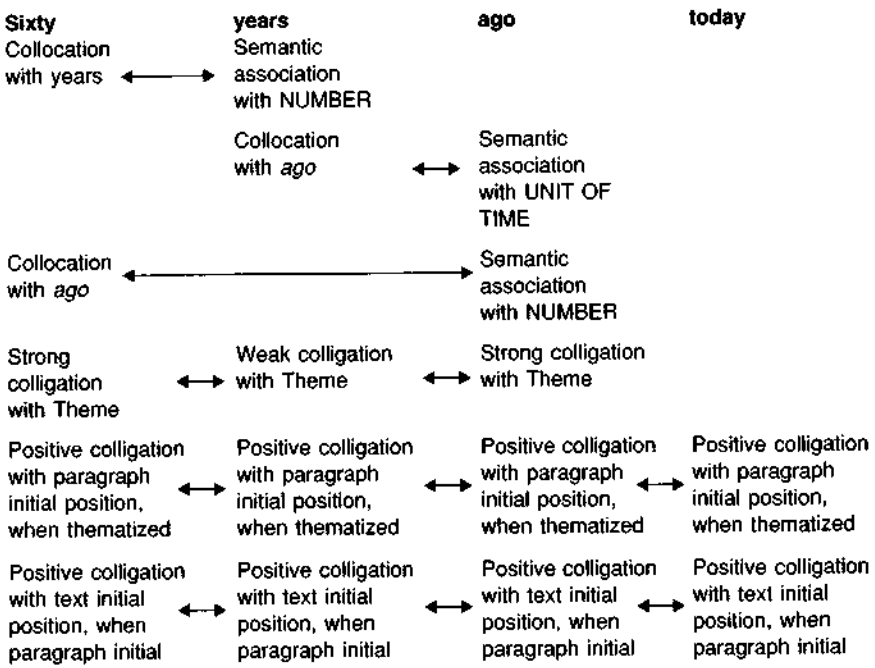
If we accept the notion that lexical items are primed for collocation, semantic association and colligation (textual or otherwise), there are two possible implications. The first is that this priming accounts for our ready ability to distinguish polysemous uses of a word. Where it can be shown that a common sense of a polysemous word favours certain collocations, semantic associations and/or colligations, the rarer sense of that word will, I would argue, avoid those collocations, semantic associations and colligations (see Hoey 2005).

The second implication is that in continuous text the primings of lexical items may combine. Thus the words that make up the phrase *Sixty years ago today*, which begins the text we considered earlier with regard to cohesion, have the primings on page 38 (amongst others) in newspaper data.

What we have here is *colligational prosody*, where the primings reinforce each other (or not), the naturalness of the phrase in large part deriving from the non-conflictual nature of the separate primings when combined. I would want to suggest that some of the work currently undertaken by grammar might be absorbed into, or superseded by, colligational prosody.

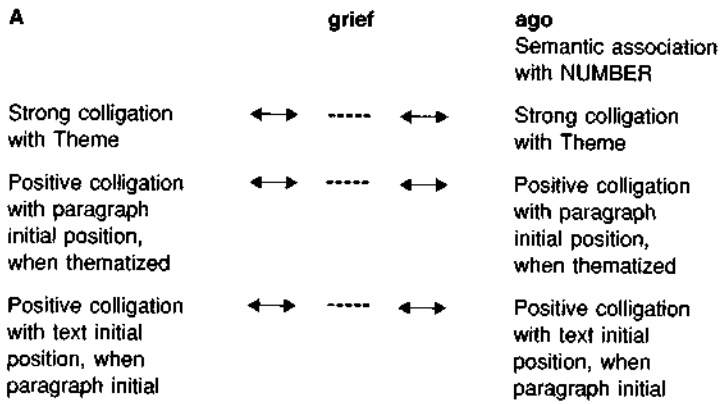
Two questions naturally arise from the preceding discussion. The first is practical in nature: what are the implications of all this for the language learner? The other is theoretical: what place is left in this theory for creativity?

To tackle the practical question first, if the notion of priming is correct, the role of the FL classroom is to ensure that the learner encounters the lexis in such a way that it is properly and correctly primed. This can only be a gradual matter; nevertheless, there are grave dangers in teachers or teaching materials incorrectly priming the lexis such that the learner is blocked, sometimes permanently, from correctly priming the lexical items.



Furthermore, certain learning practices must be inappropriate, such as the learning of vocabulary in lists (i.e. stripped of all its primings), while others (e.g. exposure to authentic data) are apparently endorsed. Authentic data, however, are usually inauthentically encountered in the classroom, in that they are read or heard for reasons remote from those that gave rise to the data in the first place. On the other hand, only authentic data can preserve the collocations, colligations, semantic associations of the language and only complete texts and conversations can preserve the textual associations and colligations.

To turn now to the theoretical question, there is of course ample room for the production of original utterances through semantic association, but semantic association will not by itself account either for the ability of the ordinary speaker to utter something s/he has never heard before or for the ability of the more self-conscious creative writer to produce sentences that are recognisably English but have never been encountered before. I would argue that, when speakers go along with the primings of the lexis they use, we produce utterances that seem idiomatic. This is the norm in conversation and writing. If they choose to override those primings, they produce acceptable sentences of the language that might strike one with their freshness or with their oddness but will not seem idiomatic. *Crucially, though, even these sentences will conform to more primings than they override.*



So when Dylan Thomas, a poet famous for his highly creative (and sometimes obscure) use of language, begins one of his poems with *A grief ago*, he rejects the collocations and semantic associations of *sixty* and *years* but conforms to the primings of *ago*, such that the phrase functions textually in similar ways to *sixty years ago*.

Priming is therefore something that may be partly overridden but not completely overridden. Complete overriding would result in instances of non-language. Thus the task that Chomsky set himself of accounting for all and only the acceptable sentences of the language requires priming as (part of) its answer. Indeed, what we think of as grammar may be better regarded as a generalisation out of the multitude of primings of the vocabulary of the language; it may alternatively be seen usefully as an account of the primings of the commonest words of the language (such as *the*, *of* and *is*). Either way, I hope I have done enough to demonstrate that a new theory of language might need to place priming at the heart of it.[1]

Note

1 Note that this sentence conforms to the priming for non-cohesion of *at the heart of* discussed earlier, and in that respect is idiomatic. In so far, however, as this endnote draws attention to the possibility of cohesion, it has created it and thereby demonstrated my ability to override one priming of the phrase while conforming to its other primings – an essential feature of a theory of priming.

References

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finnegan 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Brazil, D. 1995. *The Grammar of Speech*. Oxford: Oxford University Press.

Campanelli, P. and Channell, J. M. 1994. *Training: An Exploration of the Word and the Concept with an Analysis of the Implications for Survey Design*. London: Employment Department.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.
Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge (MA): MIT Press.
Emmott, C. 1989. Reading between the lines: Building a comprehensive model of participant reference in real narrative. Ph.D. thesis, University of Birmingham.
Emmott, C. 1997. *Narrative Comprehension: A Discourse Perspective*. Oxford: Clarendon Press.
Farr, F. and McCarthy, M. 2002. "Expressing hypothetical meaning in context: Theory versus practice in spoken interaction". Paper presented at the TaLC 5 Conference, Bertinoro, 26–31 July 2002.
Firth, J. R. 1957. "A synopsis of linguistic theory, 1930–1955" in *Studies in Linguistic Analysis*, 1–32, reprinted in *Selected Papers of J R Firth 1952–59*, F. Palmer (ed.), 168–205. London: Longman.
Flowerdew, L. 2004. "The problem-solution pattern in apprentice vs. professional technical writing" in G. Aston, S. Bernardini, and D. Stewart (eds.) *Corpora and Language Learners*. Amsterdam: John Benjamins.
Gledhill, C. J. 2000. *Collocations in Science Writing*. Tübingen: Gunter Narr Verlag Tübingen.
Goldberg, A. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.
Grosz, B. J. and Sidner, C. L. 1986. "Attention, intentions, and the structure of discourse". *Computational Linguistics*, 12(3): 175–204.
Jordan, M. P. 1984. *Rhetoric of Everyday English Texts*. London: Allen & Unwin.
Halliday, M. A. K. 1959. *The Language of the Chinese 'Secret History of the Mongols'*. Oxford: Blackwell [Publication 17 of the Philological Society].
Halliday, M. A. K. 1967–8. "Notes on transitivity and theme in English" (parts 1, 2 and 3), *Journal of Linguistics* 3.1, 3.2 and 4.2.
Halliday, M. A. K. and Hasan, R. 1976. *Cohesion in English*. London: Longman.
Hoey, M. 1983. *On the Surface of Discourse*. London: Allen & Unwin.
Hoey, M. 1995. "The lexical nature of intertextuality: A preliminary study" in *Organization in Discourse: Proceedings from the Turku Conference*, B. Wårvik, S.-K. Tanskanen and R. Hiltunen (eds), 73–94. Turku: University of Turku [Anglicana Turkuensia 14].
Hoey, M. 1997a. "Lexical problems for the language learner (and the hint of a textual solution)", in *Proceedings of the 5th Latin American ESP Colloquium*, Merida, Venezuela.
Hoey, M. 1997b. "From concordance to text structure: New uses for computer corpora", in *PALC '97: Proceedings of Practical Applications in Language Corpora Conference*, B. Lewandowska-Tomaszczyk and P. J. Melia (eds), 2–23. Łódź: University of Łódź.
Hoey, M. 2001. *Textual Interaction: An Introduction to Written Discourse Analysis*. London: Routledge.
Hoey, M. 2003. "Why grammar is beyond belief' in *Beyond: New Perspectives in Language, Literature and ELT*. Special issue of *Belgian Journal of English Language and Literatures*, J. P. van Noppen, C. den Tandt and I. Tudor (eds), Ghent: Academia press.
Hoey, M. 2004. "Textual colligation – A special kind of lexical priming", in *Proceedings of ICAME 2002, Göteborg*, K. Aijmer and B. Altenberg (eds). Amsterdam: Rodopi.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
Hunston, S. and Francis, G. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.
Langendoen, T. 1968. *The London School of Linguistics: A Study of the Linguistic Contributions of B. Malinowski and J. R. Firth*. Cambridge (MA): MIT Press.
Louw, B. 1993. "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies' " in M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology*, 157–76. Amsterdam: John Benjamins.
Martin, J. R. 2000. "Beyond Exchange: APPRAISAL systems in English". In *Evaluation in Text: Authorial Stance and the Construction of Discours*, S. Hunston and G. Thompson (eds), 142–75. Oxford: Oxford University Press.
Morley, J. and Partington, A. 2002. "From frequency to ideology: Comparing word and cluster frequencies in political debate", Paper presented at the TaLC 5 Conference, Bertinoro, 26–31 July 2002.
Murray, J. A. H. *et al.* (ed) 1884–1928. *A New English Dictionary on Historical Principles* (reprinted with supplement, 1933, as *Oxford English Dictionary*) Oxford: Oxford University Press.
Roget, Peter M. 1852. *Thesaurus of English Words and Phrases*. Harlow: Longman.
Sinclair, J. McH 1991. *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.
Sinclair, J. McH 1993. "Written discourse structure" in *Techniques of Description*, J. McH Sinclair, M. Hoey and G. Fox (eds), 6–31. London: Routledge.
Sinclair, J. McH 1996. "The search for units of meaning". *Textus* 9:75–106.
Sinclair, J. McH 1999. "The lexical item" in *Contrastive Lexical Semantics*, E. Weigand (ed.), 1–24. Amsterdam: John Benjamins.
Sinclair, J. McH and Renouf, A. 1988. "Lexical syllabus for language learning" in *Vocabulary and Language Teaching*, R. Carter and M. McCarthy (eds), 197–206. Harlow: Longman.
Sripicharn, P. 2002. "Examining native speakers' and learners' investigation of the same concordance data: a proposed method of assessing the learners' performance on concordance-based tasks". Paper presented at the TaLC 5 Conference, Bertinoro, 26–31 July 2002.
Stubbs, M. 1995. "Corpus evidence for norms of lexical collocation" in *Principle and Practice in Applied Linguistics*, G. Cook and B. Seidlhofer (eds), 245–256. Oxford: Oxford University Press.
Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.
Winter, E. O. 1977. "A clause-relational approach to English texts" *Instructional Science* 6: 1–92.
Yule, G. 1981. "New, current and displaced entity reference". *Lingua* 55: 42–52.