Gledhill, Christopher. 1999a. Towards a description of English and French phraseology. In Chris Beedham (réd.), *Langue and Parole in Synchronic and Diachronic Perspective*, 221-237. Selected Proceedings of the XXXIst Annual Meeting of the Societas Linguistica Europaea, St. Andrews, 1998. Oxford : Pergamon. ISBN 0-0804-3581-5. **(PDF)**

Towards a Description of English and French Phraseology.

Christopher Gledhill.

University of St. Andrews.

Abstract.

This paper sets out to establish the relationship between *idioms* and *collocations* and to explain how these terms relate to a general model of phraseology. We see phraseology as the relationship between a specialist language and the general or core language. We further argue that it is essential to ground notions such as phraseology in terms of discourse, an approach which emphasises the pragmatic and rhetorical functions of fixed expressions within the phraseological system, rather than simply their syntactic or semantic features. We argue that phraseology depends on the interplay between pragmatically marked expressions on the one hand (idioms), and their unmarked core equivalents on the other (collocations).

Keywords.

Phraseology, collocation, idiom, corpus linguistics, discourse analysis, science writing.

Introduction.

Linguists and non-linguists alike use a wide number of terms to express what are commonly thought of as chunks or strings in language. The following sample of terms express different aspects of this basic idea:

> …cliché, collocation, compound word, dictum, fixed expression, formula, formulaic expression, idiom, lexical phrase, lexical unit, locution, phrase, phraseme, polyword, prefabricated expression, proverb, turn of phrase, word complex… (compare with the French variants: *phrase toute faite, parlure* (in Canadian French), *tour de phrase, tournure,* etc.)

These terms essentially capture the intuitive idea that speakers select sequences of words as a whole. While the internal structure of expressions may obey the usual principles of grammar, they are also recognised as cultural artefacts rather than simply sequences or syntagms. Fillmore, Kay and O'Connor encapsulate the lexical nature of fixed expressions when they describe them as:

> "…phenomena larger than words, which are like words in that they have to be learned separately as individual facts about pieces of the language, but which also have grammatical structure [and] interact in important ways with the rest of the language." (Fillmore *et al.* 1988:501)

The idea that lexical sequences behave like words has been widely propagated by linguists such as Firth (1957) and Makkai (1992) and can be seen to have spread beyond the confines of lexicology and lexicography. For example, 'prefabs' are referred to in language acquisition theory (Granger 1994) and represent the type of expression that language learners can expect to use safely with little mastery of a language. Psychologists in turn refer to 'formulae' to describe the extent to which speakers access and predict sequences of words (Clark 1985). And the 'lexical phrase', an expression with a specific rhetorical function, is now seen as an important unit in text and discourse analysis (McCarthy and Carter 1994).

In addition, there is much evidence in mainstream linguistics to suggest that multiword items behave as single words. For example, Firth (1957) proposed that grammatical features and categories form predictable sequences (colligations) in

much the same way that single words form collocations. The idea has been pursued recently by van der Wouten (1997) in his discussion of long-range collocations and colligations extending beyond the boundary of the phrase (such as the negative associated with certain moods and verb forms). The theory of grammaticalization in Creole studies similarly emphasises the evolutionary conversion of lexical items into fully grammatical forms (Schwegler 1990). Many of the studies cited above imply that the fixedness of certain expressions eventually leads to word formation, as can be seen in *because, parce que, of course, d'accord, maybe, peut-être*, *today, aujourd'hui* etc. and in the existence of well-known historical fusions (*lord* derived from *loaf* + *ward*, *vinaigre* from *vin* + *aigre*) (Gross 1996). Idioms, collocations and other expressions therefore exist on a different linguistic level than the simple word, although in time they are used and become recognised as though they were simple lexical items, a process known as lexicalisation (Picoche 1992).

Idiomatic expressions thus embody the Saussurian principle of arbitrariness, whereby each expression is a sign composed of more than one word form with either a conventional meaning (its semantic component) or conventional formulation (a syntactic component). These conventional expressions can be said to belong to a general system of phraseology, which we define here as 'the preferred way of saying things in a particular discourse'. Within this broad definition, the concept of phraseology extends from the basic terminology belonging to a specific field (such as the fixed compounds and jargon of genetics: *gene expression, l'expression du gène*) to longer stretches of language which are typical of expository prose, or in our specific area, that of research articles (*Results have shown that + X, ces observations ont démontré que + Y*). Whereas terms such as idiom, cliché, lexical phrase and collocation often refer to discrete entities, phraseology denotes a broad system of expression. We argue here that phraseology is a continuum along which various types of expression are situated. For the purposes of this paper, we identify collocations and idioms as the opposite ends of this continuum. We attempt in the first instance to establish the continuum *idiom – collocation* according to pragmatic and rhetorical criteria. We then describe the role of collocations in specialist language, and attempt to demonstrate that 'collocational shift' is key to our understanding of core and periphery in language.

Idioms and Idiomaticity.

It is often noted that idiomatic and multi-word expressions are lexically special, or have a special grammatical status. Idioms traditionally involve at least one central lexical item (usually a 'dead' or fixed metaphor: *it's raining cats and dogs, il pleut des cordes*), or an unusual formulation, for example *to dress up to the nines* ('to dress in one's best clothes') or its French equivalent *être sur son trente et un* (literally 'to be on one's thirty one'). At times the idiom is somewhat more motivated (i.e. transparent or easily interpreted) than its foreign counterpart which appears relatively opaque and arbitrary. For example, (1) *a fat lot of good that'll do me* (meaning ironically and informally 'that is completely useless') is relatively predictable compared with the French equivalent (2) *cela me fait une belle jambe,* literally 'that gives me a nice leg'. According to traditional accounts, idioms resist changes in word choice or differ in the extent to which the expression can be transformed. Thus because one cannot say: ?*I am done a fat load of good by that,* or ?*a great load of good that might do me,* example (1) can be said to be idiomatic on lexical and syntactic grounds. Cruse (1982) summarizes the distinction between idioms and collocations very simply: collocations are grammatically simple but semantically complex (i.e. syntagmatic units, such as '*to take a break, faire une pause*'), while idioms are semantically simple but grammatically complex (i.e. semantic units *'to kick the bucket, casser sa pipe*'). The statistical analysis of collocations by Smadja (1993), or the criteria for idiomaticity set down by Fernando (1996), for example, are also based on purely semantic and syntactic grounds.

Idioms, collocations and the other terms we mentioned in the Introduction often appear to be *ad hoc* lexical items with little relation to the rest of the language system (or to each other), and have often been seen as an interesting but marginal topic in grammatical theory. Fernando (1996) makes the point that while idioms have been widely analyzed in terms of their syntactic transformability, they are seen by definition as marginal to the general principles of syntax, whereas collocations are seen by grammarians as merely syntactic restraints (most usually on verbs, as in the principle of lexical projection). Furthermore, the relation between different types of expression has been obscured, and collocations are at times presented as subcategories

of idioms (as in Fernando 1996) or at other times the other way round (as in Gross 1996). Idioms and collocations are seldom considered in terms of the 'norm' or varieties of language. For example, collocations such as *rancid butter, du beurre ranci* are presented as equivalent and 'bound collocations' involving restricted lexical items. However, while the English form is generally recognised not all French speakers recognise *ranci,* and it appears that *du beurre ranci* is rather technical and belongs to the category of LSP collocations ('specialised' collocations, a concept in the field of terminology). We return to this fundamental mismatch and notion of 'typicality' below.

Lexical and semantic properties or a sense of 'uniqueness' are not the only defining features of these expressions. Expressions, as implied by Fillmore *et al*. have a life of their own in the language, and can survive even when truncated or reformulated. While the two grammatically similar expressions (3) 'I have had it' and (4) 'I have done it' have the same grammatical structure, only one of these is recognised as an expression as such: (3) 'I have had it' can be taken to mean 'I have had enough', and this intention can not be translated literally into another language with the same effect, so not for example ?*Je l'ai eu,* but *Ça me suffit* or *J'en ai marre*. Example (3) specifically demonstrates that even formulations composed of grammatical items can be idiomatic as Mochet (1997) shows for collocations involving the French word *ça* (e.g. *ça va, ça y est, on ne fait que ça, il n'y a que ça à faire…*). But example (3) is also part of a longer expression 'I have had it up to here' (again, a long sequence of closed class items), although native speakers do not have to access the whole expression to realise its rhetorical potential. Clearly in order to have this effect 'I have had it' can not be changed radically in terms of its word order or vocabulary (although the longer version can exist in various truncated parts: "I have had it', 'I've had it', and (accompanied by an appropriate gesture) 'Up to here'). But this is not the most salient feature of the idiomatic expression. The main difference between (3) and (4) is that utterance (3) has a conventional rhetorical meaning. The move from word sequence (as a sentence) to rhetorical unit (as an utterance) is a central tenet of speech act theory, and certain linguists have claimed that this property, sometimes termed idiomaticity, is more central to a concept of native-like language use than the principles of grammatical competence proposed in mainstream linguistics (proponents include Yorio 1980, Pawley and Syder 1983, Sinclair 1991, Makkai 1992). Sinclair's

idiom principle, for example, posits that language is in constant flux (synchronically and diachronically), in a cycle between the compositional 'open choice' of single words, to the automatic and 'idiomatic' use of whole expressions.

While traditional accounts of idiom concentrate on semantic transparency or lexico-syntactic variation, others have explored the role of idioms in discourse. Makkai (1972), for example, emphasized the distinction between lexemes (compounds with predictable semantics, such as *fly off the handle*) and sememes (expressions with some rhetorical force, such as *not a mouse stirred*). The correct interpretation of either (3) or (4) above, equally depends on the extent to which they obey the general Gricean principles of conversation. As Moon (1992) points out, when utterances such as (3) or (4) appear to contravene the principles of relevance, the reader or interlocutor is justified in searching an alternative interpretation. This shift in emphasis has the advantage of making the concept of idiom less categorical. It means that sentence (3) is 'typically' interpreted as 'I have had enough' unless the literal sense 'I have had it' may make sense in context (for example as a response to (3a) Have you had your measles injection?). Utterance (4) 'I have done it' is typically interpreted as literal, if no relevant interpretation is forthcoming, and indeed it is difficult to invent a context in which (2) may be interpreted as a rhetorical utterance with some indirect meaning, and which is not the response the question (2a) 'Have you done it?'. The idea of 'authenticity' and 'naturalness', as with 'typicality' is a principle enshrined by empirical linguists such as Sinclair, and we return to them below.

According to Moon (1992), idioms play a vital role in encoding modality not only as potential speech acts, but as alternative and marked formulations in a system of choices of expression. For example, *'to walk slowly'* can be encoded subjectively as *'to walk at a snail's pace'*, where the use of the idiom can be interpreted as an additional, subjective evaluation of the proposition. For Moon, the paradigmatic choice of expression by an idiom as opposed to a more literal expression always implies some rhetorical force, and this explains the large number of idioms used as euphemisms or intensifying expressions (one thinks here of idioms for taboo subjects such as death *to shuffle off the mortal coil / manger son bulletin de naissance*, emotional states *to live it up / péter le feu*, relative success *to bark up the wrong tree / faire fausse route* and conversational gambits, *do you come here often? / tu habites chez tes parents?*).

Moon (1992) and Fernando (1996) further classify idioms according to Halliday's three 'metafunctions' (expressions which convey (a) ideational or conceptual information such as *down in the dumps*, *broyer du noir*, (b) interpersonal or dialogic information as in the ironic expression *cause toujours, tu m'intéresses* ('tell me about it') and (c) textual or relational information *at the end of the day, en fin de compte*). From this perspective, dictums, clichés and 'turns of phrase' can be seen to be archetypal idioms. Dictums and proverbs differ from other expressions in that while they share the same complex semantics of idioms, they are often seen as purely rhetorical devices where their function is to exhort or provide a metacomment (*more haste less speed* meaning roughly 'take your time' or *il faut semer le bon grain* meaning roughly 'spread the good word'). Clichés in turn have negative pejorative connotations attached to their context of utterance, and are often avoided or reformulated sometimes for humourous effect. Similarly, proverbs have a marked rhetorical function of 'advice'.

Although both Moon and Fernando point out the rhetorical role of idioms, they nevertheless stick to the traditional criteria for inclusion into the category: syntactic and semantic uniqueness. We would argue that it is equally valid to see rhetorical function and pragmatic force as determining factors in the classification of idioms, although this point can only be clarified in the light of our discussion of the related concept of collocation. For the moment, it is sufficient to point out that traditional accounts using syntactic and semantic criteria (and even not more radical accounts, such as Makkai's) fail to include as 'idiomatic' such expressions as 3) 'I have had it' and those suggested by Mochet such as *ça y est*.

Collocations and collocability.

Collocations (such as *ask a question, poser une question, high winds, vents forts, on foot, à pied*) are similar to idioms in that they involve relatively fixed sequences of words, but differ in that they are not recognised culturally or stylistically as expressions in themselves. We stated above that some linguists prefer to distinguish collocations and idioms on syntactic and semantic grounds. According to Cruse (1982), Benson *et al*. (1986:252) and others, collocations are syntactic units which

can be broken down intuitively into smaller recognisable and independent semantic units (*ask + a question, ask + the price*). Other linguists refuse to distinguish between idioms and collocations, on the grounds that they often see one form as a subordinate category of another (e.g. Moon 1992, Fernando 1996, Gross 1996). Van der Wouden (1997), for example, states:

> "…I will use the term collocation as the most general term to refer to all types of fixed combinations of lexical items; in this view, idioms are a special subclass of collocations, to wit, those collocations with a non-compositional, or opaque semantics." (van der Wouden 1997: 9)

Van der Wouden does point out that this entails problems. He cites the example of commonly considered collocations such as *a murder of crows* which happen to be opaque (i.e. interpretable as 'the slaughter of crows' and thus idiomatic). Similarly, the formulation *ask for money* is considered to be a collocation, although it is not completely compositional. The expression can not be broken down further than: *ask for + money*, so that '*for'* appears to be stuck, morpheme-like, to the verb.

By seeing idioms as essentially 'marked' expressions and collocations as 'unmarked' or normal means of expressing a concept, we are trying to make a distinction that is not categorical or binary and which lends itself to the notion of a continuum. Very common collocations such as Wierzbecka's examples of prepositional phrases (*in April, on Thursday, at ten o'clock*), are clearly unique formulations in that the prepositions are obligatory for each formulation, but are also unmarked, standard ways of expressing those concepts. Bound collocations are a little more unusual (*blond hair, nez aquilin*) and yet these represent the preferred way of saying things in general discourse. To take Moon's (1992) examples, *out of the blue, to call the shots, foot the bill*: all of these are of course semantically opaque, but they are also marked forms of more prosaic formulations, namely: 'unexpected', 'to take command', 'to pay the bill'. These phrases are idioms, because they bring some rhetorical force to the basic expression (usually by the use of explicit metaphors: the first two expressions increase the intensity of the expression, while *to foot the bill* also implies a reluctance to pay). At times it may also be the case that there are a cluster of related core statements with no really central phrase (such as *'finally, in summary, at last, enfin, en somme, pour terminer*) which coexist with more idiomatic expressions (*'at the end of*

the day, when all is said and done, all's well that ends well, en fin de compte, tout compte fait, tout est bien qui finit bien'). It is also perhaps worth noting that collocations appear to be neutral in terms of style, whereas idioms can be seen as at times inappropriate in terms of formality. We have also discussed in the previous section the possibility of analysing as idioms for the same reasons more 'transparent' expressions such as *I've had it up to here - J'en ai eu assez* = I have had enough, *Here we go again - Nous y voilà* = We must start from the beginning, *It's fine by me, Ça me va très bien* = I agree. It should be noted that we intend these judgements to be relative: the 'unmarked' forms at times coincide with even less marked forms, and will certainly change in status from one discourse or register to another.

In the following discussion, we attempt to establish the extent to which the concept of collocation can be applied to various features of language, and assume that all of the examples of collocation cited are unmarked in the general language, although we argue later that the norm will change according to context (thus scientific formulations will appear marked in general contexts, and informal forms will appear marked in formal registers). In addition, while some linguists see collocation as a rather restricted category (perhaps extending only to lexical compounds and formulations such as *addled brains*, *le cerveau fêlé*), we argue below that collocation extends to grammatical items as well as grammatical categories. The 'strong hypothesis' of collocation, especially stated by Sinclair (1991), is that every syntagm is a collocation, including even those formulations which display 'weak' collocational restrictions (such as *he forgot his keys, il a oublié ses clés*) and can be seen to enter into a default schema or colligation (S V O). In fact, Sinclair predicts that these so-called open expressions are more likely to occur in running text than canonical or stereotypical idioms (such as *it's raining cats and dogs*). According to this view, collocations are not seen as 'items' or units in the traditional sense, but underlying patterns of lexical attraction, a concept evoked by Mel'◊uk's (1984) 'lexical functions'.

Firth (1957) promoted the concept of collocation in order to relate the combinatory features of words to the rest of the language system. This idea was pursued by his students who were later to develop the systemic model of language (Halliday 1985 and Sinclair 1991). As Moon (1992), Fernando (1996) and Gross (1996) point out,

collocation is simply a restriction of expression. For example, 'to bake a cake' and 'to curry favour' are verbal phrases with varying degrees of lexical restriction ('bake' is relatively free as a verb, 'curry' as a verb is extremely limited). Collocations are important to the contrastive analysis of languages, since they reveal fundamental mismatches between lexical systems, largely in relation to these differing ranges of lexical restriction. The English and French expressions *to hammer a nail, enfoncer un clou* are 'free' in that both the verbs and their complements can be used with other words. However, *to hammer* and *enfoncer* are also used with different sets of collocates in the rest of the language system and the French verb, for example, differs markedly from other English equivalents: *enfoncer la porte* (smash a door down), *enfoncer un bouchon* (to push a cork in tight), *enfoncer un poignard* (to plunge a dagger in). Similarly the expression *to jog one's memory* is relatively restricted in English as there are few other possible complements of the verb *jog*. The French equivalent *rafraîchir la mémoire* is not as restricted (*rafraîchir* simply signifying 'to refresh'). *Jog one's memory* and *curry favour* are therefore known as a bound collocations, whereas *rafraîchir la mémoire* and *bake a cake* are said to be 'free' collocations.

While purely lexical collocations such as verb + noun, adjective + noun etc. have long been recognised, especially in the fields of terminology and lexicography, it has only recently been possible to gauge the extent to which collocation has a role to play in the grammatical system with the advent of corpus linguistics. The computational analysis of text corpora has enabled linguists to search very large text archives systematically, and according to Stubbs (1996) the computer has afforded linguistics the same degree of data processing potential as the telescope did for astronomy. There are two assumed advantages of computer-based corpus analysis: (i) as with the astronomer, the linguist can test theoretical 'armchair' hypotheses by examining authentic data, and (ii) the size of the data base can provide insights into language that had not been previously envisaged. In particular, corpus analysis has shifted the emphasis in lexicological studies away from the analysis of idioms and the transformational or semantic properties of fixed expressions, towards the analysis of collocations and the distribution patterns of stereotypical phrases. For example, dictionaries now rely on corpus evidence, not only for the existence of words and phrases but for their use and distribution patterns (Mel'◊uk 1984, Sinclair 1991,

Corréard and Grundy 1994). In addition, corpus evidence on the distribution of idioms suggests that idioms are much less widespread and more variable than previously thought. Moon (1987) has found that of 2265 idioms (including a mixed category of metaphors) identified in the 323 million word Bank of English (held at Birmingham University, England), 47% occur less than once per 4 million words. Of all the idioms examined, only 135 occur more than twice per million words (among these items Moon includes *out of the blue, call the shots, foot the bill*). This can be compared with the collocation *of course,* which occurs over 240 times per million words. Moon concludes that 'pure' idioms are somewhat marginal in nature, but are likely to be reformulated for stylistic effect (thus *to be a penny short of a sixpence* 'to be mentally deficient' is virtually always reformulated, e.g. *to be a trunk short of a tree*). The original idiom becomes obscured, and all that is left is a framework: *to be an X short of a Y*.

On the basis of such corpus evidence, Sinclair and his co-workers (Moon 1987, Renouf and Sinclair 1991, Francis 1993, Fernando 1996, Francis and Hunston 1998 *inter alia*) have demonstrated that collocations are more systematically organised in language than had previously been thought and have argued that collocation is more central to the 'idiom principle' than traditional idioms themselves. Instead of seeing collocations such as *rancid butter, du beurre ranci* as items, the collocation has been extended to a more abstract notion of the lexico-syntactic sequence. For example, Sinclair (1991) found systematic patterns of co-occurrence for prepositional verbs such as *to set in*. *To set in* typically has negative NP subjects: [*bad weather, disease, depression, gangrene*] + *sets in*. In French, we find a similar negative semantic set emerges for *essuyer* + [*une défaite, un affront, une crise, des pertes*] ('to undergo, to soak up' + 'defeat, an affront, crisis, losses' etc.). The negative semantic context of both of these terms is known as 'semantic prosody' (Lowe 1991). A semantic prosody is the net effect of an accumulation of collocations, and the lexico-grammar of the expression is inseparable from the semantics of its constituents. Any new constituents are interpreted in the light of the established collocational pattern (thus *The Tory Party had set in* has the intended implication that *Tory Party* is a negative item). New elements in the prosody can be seen to integrate but also to impregnate the pattern: thus collocations provide a framework for gradual language change.

The idea that idiomatic expressions carry with them a skeletal grammatical structure together with the notion that collocation extends beyond individual collocates to a more abstract semantic prosody leads us to the concept of the 'collocational framework' (Renouf and Sinclair 1991): an interrupted sequence of grammatical items (such as *an X of, the X-er the Y-er*) where the gap is filled by a restricted set of possible lexical words. In their 'pattern grammar' Hunston and Francis (1998) have similarly posited that most grammatical items become fixed to specific schemata and that these serve as a redundant frames for the intervening lexical items, while the intervening lexical items often belong to restricted semantic sets. Thus the frame *NP is X-ed as being* only allows verbs with similar semantics such as *considered, viewed, seen*, while the collocational framework *the X that NP has/have to V* only permits sentiments as the main metalanguage noun (X): *the wish, the desire, the need* (Winter 1997) and is usually followed by verbs such as *to succeed, to enjoy, to be loved* (etc.). Hunston and Francis point out that in the potential patterns they examine in the corpus, many are 'fulfilled' by a small number of probable formulations. Thus N *in* N is usually *increase in N*, N *into* N is usually *insight into N* and so on. If the principle of collocational frameworks is applied to longer stretches than phrases and sentences, it should be possible to arrive at some meaningful reading of a text with no lexical items available (as in a cloze-test), or to interpret foreign text or texts where the lexical items are obscured by nonsense words (the 'Jabberwocky' phenomenon, described by Hoey 1991). It the following extracts, for example, it should be difficult to guess many of the missing lexical items, but it is more likely that the reader will be able to assign a genre or text-type label to the extract (each lexical item has been replaced by an X, the morphology has been left more or less intact):

(5) X the Xs with a X X X. X the X with X and X to X, and X each X in it. X the X, X with a X X or X and X over each X. X the Xs with Xs, X on Xly and X in X X until Xly Xed.

(6) X. An X of X X which has been Xen off or Xed from the X of an X or from a X X, and is X in the X. When X enters the X, the X is Xed up by the X, and a X of the X is easily Xen off and Xs away. A X X is X in X; a X from an X X is X in X, often very X, and is X of the X.

(7) Most X and X Xed Xs X X Xs and X Xs have Xed that X might be Xed by the X of Xs such as Xs that X the X of these X Xs. The X Xs Xed were X and X, but X has Xed their X in X. A X X of X X Xs have been Xed, among which X is Xly X since it is Xly Xed by X Xs in X.

The native or fluent speaker of English should be able to identify that (5) is some form of instructional text, (6) is a definitional text and (7) an technical introduction [The original texts are presented in Appendix 1]. Other genres, such as narrative and persuasion are equally identifiable, although it is usually impossible to guess the degree of lexical specificity of the texts. Once these extracts have been checked against the full originals in the Appendix, it becomes clear that certain clues (such as lexical repetition of items) are also missing, and play an important role in the formulation of these texts (Hoey's original point). What is interesting from the collocational point of view, however is that the grammatical items provide a series of coherent links, almost establishing a rhythm within each extract, and they allow the reader to predict certain key features phrases. For example, in (5), *X the Xs with a X X X*, the first X must be grammatically an imperative verb i.e. we must obtain the pattern grammar: V the N with a (Adj. or N) N. In (6) similarly, we interpret the final two missing items as prepositional verbs which must have to do with breaking and splitting off: *and a X of the X is easily Xen off and Xs away.* The rest of the text can then be reconstructed on this basis, and of course it can be seen that given one or two lexical cues, it would be possible to build a coherent text as one expression leads to another. The final text extract (7) is very highly technical: so much so that having access to the lexical items themselves would hardly make much difference to the un-initiated. It is the case, however that non-specialists can read very specialist scientific prose (as in (7)) as though the lexical items were obscured in this way, and we are still able to build coherent interpretations on the basis of recognisable collocational frameworks. Most of the typical grammatical features of scientific text can be seen in this short extract: defining relative clauses, hedging (use of 'might'), complex nominals, passives, and so these together with a sense of some collocational frameworks leads the reader to impose a coherent reading on the text.

To summarise, although collocations and corpus evidence have mainly been exploited in lexicography, Sinclair and his colleagues have put forward a grammatical theory of

collocation which attempts to reassign the relationship between lexis and grammar in the language. Francis conceives of collocational frameworks as integral to utterances:

> "As communicators we do not proceed by selecting syntactic structures and independently choosing lexis to slot into them. Instead we have concepts to convey and communicative choices to make which require central lexical items, and these choices find themselves syntactic structures in which they can be said comfortably and grammatically." (Francis 1993:138).

Furthermore, corpus evidence has been able to challenge the general observation that grammatical items do not have general collocational properties. Even Halliday and Hasan (ironically, having proposed a theory of lexico-grammar) at one time claimed that a fundamental property of grammatical items was that they have few collocational restrictions and have little to contribute to the cohesion of text (1976:290). Yet it is possible to demonstrate that even grammatical classes such as the preposition have highly idiosyncratic collocational properties, especially for very high frequency items such as *of,* which has often been seen as atypical. Again, Francis advances this hypothesis:

> "If we take any one of a huge range of the most frequent words in English, and examine its citations *en masse*, it will emerge that it, too, has a unique grammatical profile, which certainly cannot be encapsulated by calling the word in question an adjective or a preposition." (Francis 1993: 147).

Halliday in turn points out that no one feature can characterise a register, and that a register is simply a set of statistically probable features:

> "… In fact lexis and grammar are not different phenomena; they are the same phenomenon looked at from different ends. There is no reason therefore to reject the concept of the overall probability of terms in grammatical systems, on the grounds of register variation. On the contrary; it is the probabilistic model of lexicogrammar that enables us to explain register variation. Register variation can be defined as the skewing of (some of) these overall probabilities, in the environment of some specific configuration of field, tenor and mode. It is variation in the tendency to select certain meanings than others, realising variation in the situation type." (Halliday 1991 :57)

Having raised some of the implications of a collocational approach to language, we can now turn to some of our own corpus data, and while we can not hope to establish general phraseological differences in French and English, then at least we may show

ways in which a phraseological account could be used to describe differences in the general language and specific genres.

Phraseology and genre.

In the cloze-text above, we hypothesised that a text-type or genre can be identified solely by recognising key collocational frameworks. We have previously published data on grammatical items in science writing, with the suggestion that grammatical items have radically different behaviour in different registers, even in different sub-genres (Gledhill 1995). These data show that grammatical items are not equally distributed across the language, and that the collocational patterns of grammatical items correspond systematically to register or text type, as Biber has argued from the perspective of more general grammatical categories (1996). Since the emphasis has until very recently been on the phraseology of the general language, very little work has been done on the comparison of collocational patterns between more specialised language varieties, especially those in different languages, and so we set out here some sample analyses to demonstrate the principles of our methodology.

As mentioned above, for *enfoncer* and *to hammer*, the differences between cognate terms in different languages can be particularly unpredictable. We set out to compare similar patterns from our English language corpus of scientific texts (the Pharmaceutical Sciences Corpus, PSC 500 000 words) and a recently designed French corpus (Corpus de l'Institut Pasteur, CIP 250 000 words). From both corpora, it was possible to identify phraseologies on the basis of simple computer-generated concordances (we underline the lexical collocations and artificially limit the number of concordances to five examples per word):

CIP) <démontrer>
nous avons pu démontrer l'existence d'autoanticorps
ces résultats démontrent l'existence de compétition cellulaire
sa découverte a permis de démontrer l'existence d'une nouvelle famille de gènes
nos expériences ont démontré que plusieurs mutations de cx32… entraînent une perte totale de fonction
les résultats de cette étude ont permis de démontrer les propriétés hypolipidémiaires des huiles

PSC) <demonstrate>
the present study failed to demonstrate a sustained cell proliferation

we could in no case demonstrate expression of the papillomavirus
the high optical absorption spectra demonstrated that HUM does not directly decay
the fact that we cannot demonstrate this change might be due to insufficient sensitivity of our method
we have been unable to demonstrate methylene chloride adduction to heptocyte DNA

Both verbs split very distinctly in terms of their lexico-grammar and semantic prosodies. In the English corpus, as can be seen, 'demonstrate' is almost uniquely used in negative contexts, usually where the researchers failed to demonstrate some spread of data. On the other hand, the French use of 'démontrer' shows a pattern with a strong lexical collocation: *démontrer + évidence*. Since we are dealing with a highly specialized form of writing (the research article in the biomedical sciences), negative data and failure are not perceived as necessarily bad. The expression 'failed to demonstrate' is thus very frequent in this type of English discourse, but may not carry into other fields (such as linguistics).

We suggested above that collocations and idioms exist in relation to or in competition to clusters of related expressions. In the case of 'démontrer', we can examine a number of related verbs in the French CIP corpus all relating to the biomedical preoccupation with empirical demonstrations of evidence (préciser 'to point out', étudier 'to study', montrer 'to show', indiquer 'to indicate'). The collocational patterns for these words are also divergent:

CIP <préciser>
Afin de **préciser** le rôle de phénomènes d'amplification
L'objectif des travaux menés …est de **préciser** le potentiel offert par un arcomycète
Plusieurs travaux ont permis de **préciser** le mode d'activation de ces protéines
Nous **précisons** actuellement leur rôle dans la pathogénie
analyses d'ARN ribosomique 16s ont été réalisées permettant de **préciser** les relations phylogéniques

CIP <étudier>
Un premier groupe **étudie** les bactéries fixatrices
nous avons également **étudié** la régulation des gènes
nous avons **étudié** la réponse des lyphocytes
…nous fournit des marqueurs intéressants pour **étudier** la morphogénèse
le diabète insulo-dépendant est **étudié** à travers la souche NOD

CIP <montrer>
nos résultats **montrent** l'importance de la structure …de la régulation
l'analyse génétique **montre** que sap1 est essentiel à la vie de la cellule
ces résultats… **montrent** de plus qu'il devrait être possible de vacciner contre le choléra
cette observation **montre** l'importance de pax-6 dans la formation des yeux
des recherches en région endémique **montrent** un polymorphisme important dans les …parasites

CIP <indiquer>
l'ensemble des données cliniques **indique** qu'il s'agit d'une anomalie
l'ensemble de ces données **indique** donc qu'il existe une régulation.

l'ensemble des données dont nous disposons **indique** que l'anticorps sélectionne un confomère
les données épidémiologiques **indiquent** que ce type de cancer est très fréquent
la comparaison de ces données génétiques … a **indiqué** qu'un gène unique devrait être en cause

These words appear to have found their own small but significant collocational niche in French science writing, each with different degrees of collocational fixedness and grammatical role. 'Préciser' has no English equivalent expression in the PSC corpus, but is used in French to state the aims of the research institute (usually in terms of its main collocation, its role). 'Etudier' is used with technical biochemical entities, and with less abstract, research-oriented words than 'préciser', while 'montrer' collocates lexically with 'important' but relates specifically to emphasizing the importance of a new model. Conversely, 'indiquer' is almost exclusively introduced by 'the data set' (l'ensemble des données) and is followed by a projecting V-complement clause.

It is interesting to return to the English technical corpus to examine the equivalent expressions, to see whether they occupy the same phraseological space. Perhaps predictably, some share the same niche as the French expressions while others do not. Of the two that do, 'show' appears to have the same role as 'montrer' in terms of reporting results, but does not have the same collocates, and the same is also true of 'indicate':

PSC <show> (c.f. 'montrer')
The studies reproducing elevated TNFa induction **showed** no correlation
HPC was **shown** to be topically active
It was **shown** to inhibit 12-0-tetrachloryl… compounds
PHEPC does not show any get-to-liquid planar transition about 0 degrees C.
These results **show** a dramatically reduced resistance to N,N-dimethylated antracyklines

PSC <indicate> (c.f. 'indiquer')
This result may **indicate** that AJ-1 is a very distant exon
Combined with present data, this would **indicate** that about 50% of the compound is present
these findings **indicate** that it is extremely difficult to immobilize named human cells
these results **indicate** that distinct metastasis is significantly associated
Data from other investigators …may also **indicate** the occurrence of some microciculatory events

We claim that the similarities between these expressions should be seen as significant evidence of a coherent discourse of science. Both specialist corpora involve texts by multiple authors, and texts on a wide number of issues (within the specialism of cancer research or biomedical sciences). Thus such similar phraseologies for what are known as 'semi technical' lexical items indicate an established way of writing which

appears, somehow, to have been propagated within the discourse community. However, it is necessary to distinguish between phraseological systems which appear to become established in a genre or a small specialism, and the possible regularities of the general language. To what extent are these systematic patterns (within French and within English, not necessarily between the two) different to those of the general language?

This question is unfortunately difficult to answer, not least because there is currently no French equivalent of the accessible Bank of English (ex-Cobuild corpus). We have built however a control corpus for the purposes of comparison from *Le Monde* (one million words). The following concordances, this time focusing on 'suggérer' (suggest), show that there are significant differences in French:

CLM) <suggérer> journalistique:
tes ". L'Académie de médecine **suggère** en troisième lieu une révision des resp
timisation. Le simple bon sens **suggère** d'agir le plus possible durant la pério
La pratique médiatique actuelle **suggère** une autre réflexion sur elle-même, ne s
6 ues d'un " Munich social ". Il **suggère**, désormais, l'organisation d'un référendum
22 t qu'aujourd'hui. Les mesures **suggérées** ici pourraient permettre de redistribuer...

CIP) <suggérer> scientifique:
S/BvgA. Une observation récente **suggère** qu'il existe une autre voie de régulation
Montevideo, Uruguay). Le modèle **suggère** que la spécificité anti-Tn est associée
Ces résultats **suggèrent** que l'expression de CD26 joue un rôle important
75 l'infection. Cette observation **suggère** que le système immunitaire joue un rôle
96 gillus fumigatus. Ces résultats **suggèrent** que plusieurs facteurs sont nécessaires...

Our choice of examples is of course selective (we have taken the most frequent patterns to show the typical collocates). However, journalistic 'suggérer' clearly requires nominal complements (semantically: political acts and decisions), whereas scientific 'suggérer' overwhelmingly (there are virtually no exceptions in its 25 occurrences) requires active clause complements, placing the onus on the act of suggestion (suggestion is an important part of academic hedging or modality).

Similar patterns can be seen for other cognate pairs (including nouns), lexical phrases (such as *au cours de*) and collocational frameworks (Gledhill 1997). From this research, the collocational patterns of grammatical items and lexical items appear to be more stable and fixed as we observe more specialized genres (leading to the notion of sublanguage) which suggests that when we need to reconsider the relationship

between the periphery and the core of language: the core, the *langue* can be seen as determined by the idiosyncrasies of the periphery, *parole*. It would however be a mistake to conclude that the general effect of collocational studies emphasizes the repetitive, stereotypical nature of language and also the extreme conventionality and conformism of specific genres. The pressures to conform in speech and writing styles are well known in discourse communities, as Swales (1990) points out. But the collocational patterns we have been exploring should be seen as a backdrop on which novel writing and innovation are able to develop: clearly, by our own definition, everything that is not phraseological is not the 'preferred way of saying things'.

Conclusion.

While lexicologists use the term 'phraseology' to refer to lexical co-occurrences (Thoiron, Hartley), we refer to the phraseology of a word or single expression as its rhetorical effect or pragmatic application of use. While phraseology refers to the rhetorical or pragmatic use of an expression, the term lexico-grammar, a central term in Hallidayan grammar (Halliday 1985) indicates the strict interrelationship between lexical form and syntactic formulation. This in turn allows us to distinguish the phraseology and lexico-grammar of an expression from its semantic prosody, its typical semantic context, as discussed by Lowe (1991) and exemplified in our discussion above. Our motivation for revising the distinction between idiom and collocation lies in the recent development of corpus linguistics. By attempting to fix the applications of the terms phraseology, lexicogrammar and semantic prosody in relation to each other, we envisage a model of language in which phraseology embodies a continuum of expressions from pragmatically marked forms (idioms) to pragmatically unmarked or normal expressions (collocations). Unlike other models of idiomatic expressions therefore, we use discourse criteria to determine the idiomatic status of an expression. This model presupposes that there are two forms of expression: a norm and a variant. The underlying assumptions are that a norm can be established and that the speaker has available to him or her a variety of expressions, of which many can be identified as pragmatically marked. While we have had the space to enumerate only a small number of authentic corpus examples of collocation, we hope to have shown that collocational norms (and therefore phraseological systems)

are dependent on the context of situation in which they are produced. Any concept of core language (a concept of *langue* that assumes that peripheral forms are marked or 'special' as in the term Languages for Special Purposes) must contend with the fact that in any particular discourse, new norms are forged and become effectively the new core for that particular register or genre.

References.

Benson. M., Benson., E. and Ilson R. (1986) *The Lexicographic Description of English.* John Benjamins, London.

Biber D. (1986) *Variation across Speech and Writing.* Cambridge University Press, Cambridge.

Clark E. (1985) 'The acquisition of Romance, with special reference to French.' In D. I. Slobin (ed.) *The Crosslinguistic study of language acquisition*. Vol.1 *The data*. Hillsdale, NJ: Erlbaum.

Corréard M. H. and Grundy V. (eds.) (1994) *Oxford Hachette French Dictionary*. Oxford University Press.

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

Fernando C. (1996) *Idioms and Idiomaticity*. Oxford University Press, Oxford.

Fillmore C. J., Kay P. and O'connor M.C. (1988) 'Regularity and idiomaticity in grammatical constructions.' in *Language* Vol. 64 :501-538

Firth J.R. (1957) *Papers in Linguistics 1934-1951*.Oxford University Press, Oxford.

Francis G. (1993) 'A Corpus-driven approach to grammar'. In Baker M., Francis G. and Tognini-Bonelli E. (eds.) (1993) *Text and Technology*. :137-156. John Benjamins, Amsterdam.

Gledhill C. (1995) "Collocation and genre analysis. The discourse function of collocation in cancer research abstracts and articles." In *Zeitschrift für Anglistik und Amerikanistik.* Vol. 1/1995:1-26.

Gledhill C. (1997) "Les collocations et la construction du savoir." In *Anglais de Spécialité.* GERAS: Presses de l'Université Victor-Segalen, Bordeaux No. 15-18:85-104.

Granger S. (1993) 'The International Corpus of Learner English.' in J. Aarts, P. de Haan and N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation*. :56-69. Rodopi, Amsterdam.

Granger S. (1996a) 'Romance words in English: From history to pedagogy.' in J. Svartvik (ed.) *Words*. Alanqviar and Wikall: Stockholm.

Granger S. (1996b) 'Prefabricated patterns of advanced EFL writing: collocations and lexical phrases.' in Cowie A. (ed.) *Phraseology*. Oxford University Press, Oxford.

Gross G. (1996) *Les Expressions figées en français.* Ophrys, Paris.

Halliday M.A.K. (1985) *Introduction to Functional Grammar* London: Edward Arnold

Halliday M.A.K. (1991) "Towards probabilistic interpretations." in E. Ventola (ed.) 1991 *Functional and Systemic Linguistics: Approaches and Uses* :39-61. Mouton de Gruyter, Den Haag.

Halliday M. A. K. (1993) 'Quantitative studies and probabilities in grammar' in M. Hoey (ed.) 1993 *Data, Description, Discourse: Papers on the English Language in Honour of John McH Sinclair*. HarperCollins, London.

Halliday M.A.K. and Hasan R. (1976) *Cohesion in English* Longman, London.
Hoey M. (1991) *Patterns of Lexis in Text* Oxford: Oxford University Press

Hunston S. and Francis G. (1998) 'Verbs observed: A Corpus-driven pedagogic grammar.' in *Applied Linguistics* Vol. 19/1:45-72.

Jacobi D. (1994) 'Lexique et reformulation intradiscursive dans les documents de vulgarisation scientifique.' In Candel D. (ed.) (1994) *Français scientifique et technique et dictionnaires de langue*. Didier Erudition, Paris.

Makkai A. (1972) *Idiom Structure and Idiomaticity*. Mouton, Paris.

Makkai A. (1992) 'Idiomaticity as the essence of language'. In *Actes du XVème congrès international des linguistes.* :361-365. C. I. L, Presses universitaires de l'Université laval, Québec.

McCarthy M. and Carter R. (1994) *Language as Discourse*. Longman, London.

Mel'◊uk I. (1984) *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques.* Vol. 1. Presses de l'Université de Montréal, Montréal.

Mochet M-A. (1997) 'Expressions et groupements discursifs de l'oral: questions d'inventaire et propositions didactiques.' Unpublished paper presented at *Colloque Triangle*. ENS Fontenay / St. Cloud, 7-8th March 1997.

Moon R. (1987) 'The analysis of meaning'. In Sinclair J. McH. (ed.) (1987) *Looking up: an Account of the Collins COBUILD Project*. :86-103. Collins ELT, London.

Moon R. (1992) 'There is reason in the roasting of eggs. A comparison of fixed expressions in native speaker dictionaries.' in *Euralex '92 Proceedings* Oxford University Press :493-502

Nattinger J., and De Carrico R. (1992) *Lexical Phrases and Language Teaching*. Oxford University Press, Oxford.

Pavel S. (1993) 'Neology and phraseology as terminology-in-the-making.' In H. B. Sonneveld and K. L. Loening (eds.) *Terminology: Applications in Interdisciplinary Communication*. :21-34. John Benjamins, Amsterdam.

Pawley A. and Syder F. H. (1989) 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency'. In J. C. Richards and R. W. Schmidt (eds.) (1989) *Language and Communication*. :191-226. Longman, London.

Picoche J. (1992) *Précis de léxicologie française. L'étude et l'enseignement du vocabulaire*. Nathan, Paris.

Renouf A. and Sinclair J. McH. (1991) 'Collocational frameworks in English.' in K. Aijmer and B. Altenberg (eds.) (1991) *English Corpus Linguistics* :128-144. Longman, London.

Rey A. (1977) *Le Lexique. Images et Modèles.* Armand Colin, Paris.

Schwegler A (1990) *Analyticity and Syntheticity: A Diachronic Perspective with Special Reference to Romance Languages.* Empirical Approaches to Language Typology, 6. Mouton de Gruyter, Berlin/New York.

Sinclair J. McH. (1991) *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Sinclair J. McH., Fox G. and Hoey M. (eds.) (1993) *Techniques of Description*. Routledge, London.

Smadja F. (1993) 'Retrieving collocations from text: Xtract.' in *Computational Linguistics* Vol. 19/1 :143-177.

Stubbs M. (1996) *Text and Corpus Analysis*. Routledge, London.

Svartvik J. (ed.) (1992) *Directions in Corpus Linguistics*. Proceedings of the Nobel Symposium 82, 4-8 August 1991, Stockholm.

Swales J. (1990) *Genre analysis. English in Academic and Research Settings*. Cambridge University Press, Cambridge.

Titone D. A. and Connie C. M. (1994) 'The comprehension of idiomatic expressions: Effects of predictability and literality.' *Journal of Experimental Psychology: learning, Memory and Cognition*, No. 20 :1126-1138

van der Wouden T. (1997) *Negative Contexts. Collocation, Polarity and Multiple Negation.* Routledge: London.

Yorio C. (1980) 'Conventionalized language forms and the development of communicative competence.' *TESOL Quarterly*. Vol. 14/4:433-422.

Wierzbecka A. (1993) 'Why do we say *in April, on Thursday, at ten o'clock*? In search of an explanation.' In *Studies of Language* 1993 Vol. 17/2:437-454.

Winter E. (1996) 'Metalanguage nouns of clause relations'. unpublished paper presented at *Corpus Research: Sharing Interpretations* 20th Sept. 1996, University of Birmingham.

Appendix.

Cloze Texts and Collocational Frameworks

The reader is invited to attempt to find the lexical items (each replaced by one X) and attempt to assign a genre or text-type label to the following extracts. The full texts follow these amended extracts (reference numbers 5 - 7 refer to the examples discussed in the paper):

(5) X the Xs with a X X X. X the X with X and X to X, and X each X in it. X the X, X with a X X or X and X over each X. X the Xs with Xs, X on Xly and X in X X until Xly Xed.

(6) X. An X of X X which has been Xen off or Xed from the X of an X or from a X X, and is X in the X. When X enters the X, the X is Xed up by the X, and a X of the X is easily Xen off and Xs away. A X X is X in X; a X from an X X is X in X, often very X, and is X of the X.

(7) Most X and X Xed Xs X X Xs and X Xs have Xed that X might be Xed by the X of Xs such as Xs that X the X of these X Xs. The X Xs Xed were X and X, but X has Xed their X in X. A X X of X X Xs have been Xed, among which X is Xly X since it is Xly Xed by X Xs in X.

(5) Wipe the fillets with a clean dry cloth. Season the flour with salt and pepper to taste, and dip each fillet in it. Beat the egg, mix with a little milk or water and brush over each fillet. Coat the fillets with breadcrumbs, press on firmly and fry in hot fat until nicely browned. (Mrs Beeton's Cookery Book)

(6) Iceberg. A mass of land ice which has been broken off or carved from the end of a glacier or from an ice barrier, and is afloat in the sea. When a glacier enters the sea, the ice is buoyed up by the water, and a portion of the glacier is easily broken off and floats away. A glacier berg is irregular in shape; a berg from an ice barrier is rectangular in shape, often very large, and is characteristic of the Antarctic. (W. G. Moore's Dictionary of Geography)

(7) Most rodent and human xenografted tumours contain hypoxic cells and clinical studies have suggested that radiotherapy might be improved by the use of agents such as nitroimidazoles that increase the radiosensitivity of these hypoxic cells. The first agents evaluated were metronidazole and misonidazole, but neurotoxicity has limited their use in radiotherapy. A second generation of hypoxic cell sensitisers have been developed, among which pimonidazole (PIMO) is particularly interesting since it is preferentially accumulated by tumour cells in vitro. (Cancer research article introduction from the Pharmaceutical Sciences Corpus: Gledhill (1995)).