Frath, Pierre & Gledhill, Christopher 2005a. Free-Range Clusters or Frozen Chunks? Reference as a Defining Criterion for Linguistic Units. *Recherches anglaises et Nord-américaines*. (38). 25-43. ISSN 0557-6989.

Free-Range Clusters or Frozen Chunks? Reference as a defining criterion for linguistic units

Pierre Frath	***Christopher Gledhill***
Département d'anglais *UFR des Langues Vivantes* *EA1339 Linguistique et Didactique des Langues* *Université Marc Bloch* *22 rue Descartes* *67084 Strasbourg* *France* *Email : frath@umb.u-strasbg.fr*	*Langues Etrangères Appliquées* *UFR des Sciences Humaines Appliquées* *EA1339 Linguistique et Didactique des Langues* *Université Marc Bloch* *22 rue Descartes* *67084 Strasbourg* *France* *Email : gledhill@umb.u-strasbg.fr*

Résumé

Des collocations telles que *strong tea* ou *spill the beans* sont généralement expliquées selon l'un ou l'autre de deux points de vue. Pour le premier, leur existence découle simplement de l'usage ; pour le second, il existerait une sorte de 'colle' grammaticale ou logique sous-jacente, qui serait à l'origine de leur formation. Le premier point de vue ne propose pas d'explication, en fait, puisqu'invoquer l'usage revient simplement à dire que c'est ainsi que nous faisons; le second s'empêtre, comme il sera vu, dans des problèmes métaphysiques et dans la contradiction. Dans cet article, nous développons une autre possibilité explicative, celle de la référence comme critère de définition des unités linguistiques. Suivant Peirce (1978), nous avançons que les mots complexes, les collocations et les expressions idiomatiques qui réfèrent à des 'objets' sont des 'dénominations', qui se développent au sein d'énoncés discursifs que Peirce appelle des 'signes interprétants'. Notre approche propose une conception unifiée des items polylexicaux et montre comment ils fonctionnent dans des portions de texte plus importantes.

Abstract

Collocations such as *strong tea* and *spill the beans* have usually been accounted for by one of two points of view. The first explains their existence simply by usage, while the second posits an underlying grammatical or logical 'glue' which is supposed to account for their formation. The first does not provide any explanation at all, since invoking usage boils down to saying that this is just the way we do things, whereas the second is bogged down, as we shall see, in metaphysics and contradiction. In this paper, we develop another possibility: a reference-based definition of linguistic units. Following Peirce (1978), we argue that complex words, collocations and idioms which refer to socially accepted 'objects' are 'denominators', which are distinguished from discursive utterances known as 'interpretants' in Peirce's work. This approach proposes a unified view of polylexical items and explains how they work in larger chunks of text.

Introduction

In the field of lexicology, restrictions on the combination of words are usually discussed in terms of **collocation**. A collocation can be defined as a complex of words which functions like a single lexical item, as in *merry-go-round, blow the gaff* and *stark naked*. Collocations are related to a variety of different types of lexical expression, including "catch-phrases, clichés, fixed expressions, formulae, free and bound collocations, idioms, lexical phrases, turns-of-phrase and so on" (Gledhill 2000b:7). The distinctions between these categories have caused much debate and any definition of collocation varies according to the observer's particular standpoint. A handbook for EFL learners might class several types of expression as 'idiomatic', whereas a dictionary might distinguish between collocations, lexical phrases, proverbs, and so on. Research in this area is very much alive and kicking, as many linguists feel that collocations and other phraseological units constitute an unsolved problem. A case in point is the recent *Proceedings of the European Society for the Study of English* (Hamm, Frath & Rissanen 2003), where the majority of papers submitted dealt with phraseology, despite the fact that the range of subjects was open and contributors were free to choose their own topic.

1. Three views on collocation

The methodological and theoretical approaches to collocation can be divided into three general types (Gledhill 2000b:7-18), according to the way in which collocations are defined:

1. statistical / textual clusters
2. semantic / syntactic chunks
3. discoursal / rhetorical chains

The first two approaches conceive of collocations as phraseological units, either as statistically determined 'clusters' or relatively fixed 'chunks'. The third approach is somewhat different, in that it considers collocation to be a chain in the development of discourse. It can be seen that while the first two approaches have methodological advantages, they also have theoretical problems. In the second half of the article, we argue that the third approach may offer a more fruitful theoretical perspective, especially when we consider the notion of reference as a defining criterion.

1.1 Statistical / textual clusters

The statistical / textual view of collocation became established with the advent of large-scale computer-based corpus studies of texts, as exemplified by the work of Sinclair (1966), Kjellmer (1982), Smadja (1989) and others. Such studies have shown that the vicinity of a given lexical item is not hap-hazard. As Van Roey puts it:

> [collocation is] that linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its 'synonyms' because of constraints which are not on the level of syntax or conceptual meaning but on that of usage. (1990:46 [our emphasis]).

Such 'lexical preferences', or collocations, can be defined as statistically significant co-occurrences of items within a span of arbitrarily fixed length. The collection of items which commonly co-occur near to a given lexical item form what are known as **clusters**. According

to the statistical / textual definition, *blow the gaff* and *stark naked* qualify as collocations simply on the grounds that they are statistically more likely to cluster next to each other than expected.

A particular advantage of the statistical / textual approach, as set out in Sinclair (1987), is that it strictly avoids treating linguistic units as though they are derived from *a priori* grammatical units. Sinclair argues that far from depending on 'open choice' or compositional meaning, our default system of interpretation is governed by the 'idiom principle'. Supporters of this approach point to psycholinguistic studies which have shown that the meanings of idioms can be directly accessed without passing through a stage of literal interpretation (Swinney & Carter 1979, Gibbs 1985).

The proponents of the statistical / textual perspective have convincingly argued that quantitative studies reveal patterns of use that are not obvious when analysts examine texts individually or rely on introspection alone. In a well-known example, Sinclair showed that *set in* as an intransitive prepositional verb collocates with negative subjects such as *rot* or *decay* (Sinclair 1987:150-159). By contrast, *happiness is setting in* would seem rather awkward, because *happiness* seems to be endowed with a negative quality at odds with its generally accepted positive meaning. It is noticeable that this constraint on *set in* is not mentioned in traditional, non-corpus-based dictionaries, probably because intuition alone does not reveal it.

So there is a strong case for the quantitative observation of machine-readable texts, and corpus-based studies have become a staple in lexicography. However, apart from work on pattern grammar (Hunston and Francis 2000), this type of approach has not had much impact on grammatical theory. Indeed, the statistical / textual approach offers no explanation about why lexical items should 'prefer' (as van Roey says) the company of other words. Observation and the idiom principle do not therefore appear to be enough. The linguist needs a more principled framework capable of accounting for the choice of companions lexical items make.

1.2 Semantic / syntactic chunks

Such a principled approach seems to be offered by the semantic / syntactic perspective. According to this point of view, the reason why words coalesce into a collocation is that they obey inherent semantic and syntactic patterns of particular lexical items. Collocations can thus be categorised by the extent to which they form increasingly fixed units or **chunks**. They range from free collocations, whose meanings are said to be totally compositional (*blow a trumpet*), to restricted collocations, which are less compositional (*blow a fuse*), figurative idioms (*blow your own trumpet*) and pure idioms (*blow the gaff*) (Howarth 1996: 32-33). This way of classifying expressions relates increasing syntactic invariability or 'frozenness' to increasing semantic particularity or 'opacity'. A frozen (or fixed) expression is one that cannot be transformed in the usual generative sense of the term. For example the idiom *blow the gaff* is supposed to resist passivisation (?*the gaff was blown by us*) as well as clefting (?*it's the gaff that we've blown*). From a semantic point of view, *blow the gaff* is opaque since its meaning ('to ineptly reveal a secret') cannot be predicted from its individual words. The status of *blow the gaff* as a 'pure idiom' is enhanced by the fact that *gaff* is a lexical fossil; i.e. it is lexically unproductive, and its figurative meaning is unrelated to other homonyms which have come into English (Provençal *gaf* 'hook' or French *gaffe* 'mistake'). By contrast, the figurative idiom *blow your own trumpet* is relatively transparent, in that it can be 'decoded' and glossed as 'to promote yourself' (Makkai 1972).

The semantic / syntactic view thus provides a reasonably objective way of classifying different types of phraseological units. But it has its problems. It is difficult to argue that increasing syntactic frozenness is systematically related to increasing semantic opacity. For

example, it appears that *blow the gaff* can in fact be 'defrosted', since we find one counter-example from the British National Corpus: *…her mother was anyway exceptionally good at not acknowledging that a gaff had been blown*.... In fact functional linguists such as Abeillé (1995) have argued that no expression can escape from the general rules of syntax, and it would seem that all idioms, even the most frozen ones, can undergo at least some syntactic transformation:

> Idioms cannot be divided into two sets: fixed idioms (not subject to any syntactic rules) and flexible idioms (presumably subject to all rules) (Abeillé 1995: 15).

A further problem with the semantic / syntactic approach is that it cannot distinguish between ungrammatical and original expressions. Presumably, a neologism or new phrase enters the continuum as a 'free collocation' such as ?*blow a smile* (by stretching the analogy with *blow a kiss*). The proponents of this approach are therefore obliged to refer to a notion of 'acceptability', which is clearly a very different criterion from that of semantic opacity or syntactic frozenness.

So far we have seen the semantic / syntactic view as a 'top-down' approach, taking the expression as a whole chunk. A similar set of problems besets those approaches which take a 'bottom-up' approach, that is classifications which gauge the extent to which individual words can be chunked together in terms of 'compositionality'. A recent example of this appears in Nesselhauf (2003). Following Cowie, Nesselhauf posits a distinction between 'free combinations', 'collocations' and 'idioms'. Each can be distinguished by what Nesselhauf calls 'restricted sense', which corresponds to increasingly arbitrary restrictions on lexical compositionality. For example, she argues that:

> "… *want* can be combined with a great number of nouns (*want toys, a child, a drink, a car, truth* etc.) and there are no arbitrary constraints on its combinability… *perform* (as in to *perform a task*) on the other hand, would be considered as having restricted senses, because […] some nouns that seem to be possible from a semantic point of view are not possible (e.g. **perform a survey*, c.f. Cowie 1994: 3169)" (Nesselhauf, 2003: 225-226).

This kind of opposition appears to be a standard approach in lexicology and semantics. Unfortunately, the restrictions posited by Cowie and Nesselhauf turn out to be false when we submit them to corpus analysis. If we consult the BNC, we discover at least one counter-example: *Rocco apparently performed a survey and determined that 9 out of 10 skaters are street skaters*... Many more examples can be found on the web using a text-browser such as *alltheweb.com*. Not even looking for inflected forms or variants, we found over 3000 examples of *perform a survey*, of which many are valid native-speaker uses. This is rather interesting, because Cowie's claim about ?*perform a survey* is contained in a major encyclopaedia of linguistics, and appears to have been reproduced without question.

A comparable problem emerges when we consider Nesselhauf's claim about *want*. It is true that the verb *want* takes many thousands of different types of complements. But Nesselhauf suggests there are no 'arbitrary constraints' on its complementation. Do we imply from this that *want* does not collocate with anything, or collocates with everything? In fact, three main types of nominal complement for *want* are listed in the Cobuild dictionary (based on the *Bank of English* corpus, Sinclair 1987). Starting with the most frequent usage, these include Noun Group complements expressing bald demands to a second person (*I want you, I want an explanation from you Jeremy, What do you want?*), resultatives expressing a goal (*I want my boy alive, I want my car this colour, They began to want their father to be the same as other daddies*) and very specifically a wish to have children (*I want this baby very much*). These are clearly very different but consistent collocational clusters. It would be unwise therefore to categorise the complements of such a frequently used verb as 'free combinations',

and we are led to the conclusion that most other verbs, even high frequency ones, can display a similarly restricted set.

If there is no clear distinction between 'free combinations' and 'restricted collocations', we are left with the task of explaining why words chunk together in the first place. Some authors put forward an explanation in terms of abstract relations which govern patterns, for example intensification as in *sleep like a log, stark naked* or *piping hot*, where the underlined items are intensifiers. In his influential study, Mel'cuk has attempted to gather these relations together in terms of 53 general lexical functions (Mel'cuk 1988a, 1988b). *A speck of dust* is made possible because one of these functions governs 'quantity'. Similarly, *to lend support* or *deal a blow* is possible thanks to an 'operational' function, and so on. In this view, collocations occur when abstract universal entities trigger lexical and grammatical patterns into the creation of more or less rigid set phrases, their degree of frozenness and opacity depending on the degree of lexicalisation. Of course this raises the question why these abstract relations are put into motion in the first place.

1.3 Discoursal / rhetorical chains

The third element in our typology, the discoursal / rhetorical point of view, bypasses the question of structure and causality altogether. A number of functional grammarians and discourse analysts (including Nattinger and De Carrico 1992, Fernando 1996 and Moon 1998) have taken this line, dispensing with the traditional debate about compositionality, and instead examining the textual and pragmatic functions of collocations (sometimes termed 'sentence stems' or 'lexical phrases'), their role in language acquisition or their use in texts. Halliday and Hasan's (1976) work on text structure can be seen to inform this approach, in that for them, collocation is a cohesive feature of textual reference rather than a unit of grammar or phraseology:

> …because they lie outside the bounds of structure, and are not constrained by structural relationships, … lexical patterns serve to transform a series of structures into a unified, coherent, whole (Halliday and Hasan, 1976: 320)

Hoey (1983) discusses the textual function of lexical items in terms of lexical **chains**: these are key organising expressions in a text which not only have local significance, but also serve to establish lexico-grammatical dependencies that emerge across clause boundaries within texts. According to this approach, phraseological units are not defined by formal criteria, but are seen as more or less stylistically marked members of a family of expressions, such as *to get the sack, to be fired*, *to be dismissed, to lose one's job*. Since the members of these families are not related by form but by function, they can be treated at the paradigmatic level of single-word units. The question of relative frozenness and opacity thus becomes irrelevant. As Gledhill has put it:

> … the discourse / rhetorical approach is not concerned with lexis and grammar as such. Instead, the suggestion is that collocations and idioms can be distinguished on the basis of a rhetorical or textual function […] or pragmatic marking […] (Gledhill 2000b:14).

We would argue that what all of these expressions have in common (lexical phrases, sentence stems, chains) is that they are neither defined in terms of local co-occurrences (as clusters) or in terms of a unitary expression (as a chunk), but in terms of discourse. The discourse perspective on phraseology leads us to distinguish between expressions that have a significant role to play in textual reference, and those that have a role to play in the incremental or 'instantial' development of a text or discourse. Putting this more simply, some linguistic units

refer as wholes, as **denominators**, regardless of their grammatical size or structure, while other combinations (which we call **interpretants,** following Peirce) refer as constructs. This distinction is explained in the following section.

2. Reference *vs.* concept

Thus far we have argued that the statistical / textual perspective is essentially a descriptive heuristic and does not give any principled explanation about why lexical items cluster or express lexical 'preferences'. In contrast, the semantic / syntactic point of view does provide some kind of explanation, but faces problems of classification and the ontological nature of the hypothetical functions which are supposed to govern the formation of chunks. Our third discoursal / rhetorical point of view ultimately considers polylexical units to be on a par with single words and therefore does not deem any distinction necessary. The first two points of view share a concern about lexical preference and form, because they regard collocations and their clusters of co-occurrence as constructs, while the third approach clearly takes a holistic view. In this section, we develop our reasons for preferring this 'third way'. In particular, we take insights from the quantitative observation of corpora as well as Peirce's theory of reference.

2.1 The cognitivist viewpoint

We have argued that it is impossible to divide between compositional expressions ('free collocations') on the one hand and fixed expressions (ranging from 'restricted collocations' to 'pure idioms') on the other. We have argued instead that all words enter into collocational relations (as can be seen with the verb *want*) and that even high frequency grammatical items have a collocational lexico-grammar (as demonstrated in Gledhill 2000a). This position is clearly very different to the viewpoint generally adopted by many cognitivists. The cognitivist approach posits that all usage is based on composition (in their terms, all surface structure is the product of underlying deeper patterns). Any exceptions, as discussed by Pinker (1994: 148) are seen as irregular forms or idioms which must be classified as **listemes**. A listeme, as defined by Di Scullio and Williams (1987), is a lexical root or phrase which cannot be produced mechanically by rules and has consequently to be memorized as part of a list. While the use of the term might imply a concession to a more phraseological point of view in generative theory, it in fact infers a clear division of labour between the lexical module on the one hand and the syntax on the other. If we examine Jackendoff's attempt to account for idioms in generative theory, we can see that his argument for a unified treatment of words and fixed expressions is couched in exactly these terms:

> In productive syntactic composition, the meaning of a phrase is a rule-governed function of the meanings of its parts. However, when a syntactic phrase is lexically listed, there is no need to build it up semantically from its parts – the meaning is already listed as well, so full linking of the parts is unnecessary.… [T]he lexical listing of idioms […] must override in whole or in part the meanings of the constituent words and the way they are combined in the literal meaning. (Jackendoff 1995: 148, 152).

The cognitivist assumption is that an utterance is generated at an abstract level by logical, semantic and syntactic entities whose job it is to match pre-existing thought to a resulting linguistic string. Thought is somehow transformed into an abstract and organised set of entities which are in turn translated into language. The linguist's job is then one of formulating the rules and mechanisms which permit such a feat, to discover which sort of syntactic and semantic 'glue' will hold the bits and pieces together. This view is commonly

shared by many linguists, who assume that a non-idiomatic expression is a 'free combination' to be interpreted compositionally as the sum of its parts .

The main problem with the cognitivist view is that even 'free combinations' appear ultimately to be conventional. In fact, it is arguable that any so called 'free combination' can turn out to be a phraseological unit. To use J.R. Firth's example, why can we not accept *powerful tea* and *strong car,* when both *powerful argument* and *strong argument* are acceptable? Is it possible to argue that the meaning of *powerful* is somehow more 'restricted' in combination with *car* than with *argument*? This sort of explanation seems too *ad hoc* to have any real explanatory value. Similarly, the verbs *end, finish* and *terminate* could certainly be considered as carrying the same meaning, yet one can *end* or *terminate a pregnancy,* but not *finish* one[1]. To cope with such difficulties, cognitivists tend to introduce a number of intermediate semantic entities, such as facets (Cruse 1995), active zones (Langacker 1984), *qualia* (Pustejovsky 1991), or an array of more or less *ad hoc* mechanisms. The alternative, of course, is to appeal to a general principle of phraseology, such as Sinclair's 'idiom principle', as we saw above, which simply amounts to asserting that this is the way we do things.

What sort of an explanation can we formulate, then, if we exclude the cognitivist school for being too *ad hoc*, and usage-based explanations such as Sinclair's for being too general and too powerful? Is it possible to view language from another perspective and to reduce the problem under scrutiny to a more general one? We suggest that a more fruitful way of looking at the problem is to re-examine the notion of reference, and more generally the 'act of naming'.

2.2 Denominator, object and interpretant

Our solution to the problem of phraseological units is inspired by Charles S. Peirce's theory of semiotics. We shall not present Peirce's philosophy in this paper, for lack of space. Interested readers may refer to Peirce's work[2] or to Frath (2005) for a summary and discussion. Briefly, we posit three semiotic entities: **denominator**, **object** and **interpretant**. 'Denominator' is our term for the French *dénomination*, from Medieval Latin *denominatio*. We have coined it to avoid the religious connotations of 'denomination' and also because it echoes a key point we want to make about reference, i.e. 'the act of naming'. Reference has not been popular in linguistic theories over the last century, both in Continental and Anglo-Saxon linguistics. Saussure explicitly rejected reference in order to consider language solely as a system of signs, and cognitivism totally bypasses the question as it views language as the expression of concepts and other mental entities.

A denominator (symbolised here by N) is a word or a string of words which refers globally to elements of our experience which are lumped into a category by the N. Ns are not usually created by the individual, they are given to us by our community. They are what Merleau-Ponty (1945) called *parole instituée*, i.e. institutionalised language. Whenever we get acquainted with an N, we naturally suppose that it refers to an object (or O), even if we know nothing about it. For example, when we come across *pennon, prolegomena* or *propitiatory* for the first time, we know that these refer to objects which our linguistic community has named in this way. It is only after we have learned both the name and the object that we are able to acquire knowledge about the object. Such knowledge is put across to us, or created by us, within longer discursive strings, which Peirce calls 'interpretants'. A concept is an interpretant. It is produced linguistically after the object has been named. The reader can check for himself that we cannot talk about something if we do not know its name, be it a

[1] We thank David Allerton for alerting us to this example.

[2] French speakers may read Peirce (1978), a choice of basic Peircian texts, with a very good introduction by Gérard Delledalle.

colour, a feeling, an illness, an ideology, etc. We can not even think about it in any structured way. Therefore the cognitivist claim that the concept is the semantic content of lexical items is dubious, because it posits knowledge before it posits words and does not make clear how concepts can actually refer to the external world.

Words are not the linguistic shape of concepts that each of us produce in our minds. If this were the case, it would be very difficult to ensure that we all produce the same concepts and name them the same way[3]. We would have to posit a very strict set of primitives together with semantic and syntactic rules that we would be endowed with from birth, genetically. Even then, it would still be difficult to explain why a given concept should be split up differently or take different shapes in different languages. One might then posit a universal grammar filtered by existing languages, and this is of course Chomsky's generative point of view. Yet the mind boggles at the number of hypothesised entities posited by cognitivist theory and at the endless recess of new explanatory entities each new observation seems to generate. Linguistics probably needs explanations that do not strain our credulity to such insufferable levels.

We could simply accept the common sense view that words are given by our linguistic community and that they point globally to elements of our existing common experience. They are not in our brains from birth in the shape of primes and universals which some cognitive mechanisms manage to put together into meaningful linguistic strings. We have to learn them. Objects and words come first, everything else follows when they are the subjects of interpretants. For example, before we get acquainted with *pennon, prolegomena* or *propitiatory*, we do not know that the objects they refer to exist. It is only after we have heard or read the Ns that we know they have a social existence. We then suppose that our community has already compiled knowledge about them in the shape of interpretants, i.e. other more discursive signs, which a more knowledgeable person might know or which might be stored in books or suchlike. To understand them, we may be able to infer the objects they refer to from the discursive context and the real-world situation, as is usually the case when we learn a new word. If we are unable to do so, we can inquire by asking a question or by looking up the words in the dictionary. In both cases, we will be given interpretants, i.e. an explanation or a definition. When the objects are finally linked to their denominators, we can start using them in interpretant discursive utterances.

In essence, words name our habitual experience and their acquisition is on a par with the acquisition of other habits such as using knives and forks or chop sticks, or smiling when we meet friends. When Ns are completely new, as *denominator* and *interpretant* may be to the reader of this paragraph, they are usually uncomfortable and difficult to understand. To overcome the difficulty and be able to use them successfully, the reader will have to acquire a new linguistic *habit*, derived from exposition to various social uses of the words, beginning with this text. He will find *denominator* and *interpretant* in other contexts, but the mathematical meaning of *denominator* will not be of great help. The use of the word in linguistics will gain usage only if the English-speaking community of linguists accepts it as referring to an object and endows it with social existence. If it does not, the linguistic use of *denominator* will remain an idiosyncrasy of these two authors, despite the fact that *dénomination*, being a denominator in French, enjoys social existence in French linguistics.

In the following sections, we apply this approach to a variety of lexical units, namely monolexical words, polylexical phrases, so-called 'free combinations' and larger chunks of text.

3. Applying the theory

[3] This view of language was developed by Ludwig Wittgenstein in his *Philosophical Investigations* in what is known as the *Private Language Argument* (paragraphs 243 and following).

3.1 Reference and lexemes

Monolexical words are often constructed from other words, for example *psychoanalysis* or *retrovirus*. Yet *psychoanalysis* is not just the analysis of the psyche; it refers globally to a theory of the mind and a therapy, independently of its constituents. On hearing *psychoanalysis*, we do not break it down into bits and pieces to understand it. We know the object has social existence, and this means we do not *have to* know anything about Freud, Jung, Transfer or the Oedipus complex to be able to use the word or understand it when we hear it. If we did, it would mean that we could only speak about objects about which we have intimate knowledge. This is obviously not the case. One of the authors of this paper has recently taken to bird-watching. He was dismayed when he realised he had been talking of *chaffinches*, *blue tits* and other birds for decades without the slightest idea of what they look like. The other author is a confirmed city slicker and has no clue at all about any of these creatures. Our minimal understanding of birds does not involve knowledge; it consists in knowing that they have a social existence materialised by a denominator.

If we want to go beyond this minimum, constituent analysis will not help us much in understanding what *psychoanalysis* is about, even the first time we hear the word. We certainly understand that it has something to do with mind and analysis but this is not enough. *Psychoanalysis* is not synonymous with *analysis of the mind*. The former is an opaque and timeless denominator which refers synthetically; the latter is a discursive and transparent NG which happens to be used *hic et nunc* to refer analytically. The difference between the two is not syntactic or semantic. What distinguishes them is the nature of their referential capability: one of them is an N referring to an ontologically existing object, the other one is an NG whose object needs to be spelled out in an interpretant, in other words as the instantial development of a text. Suppose someone mentions *psychoanalysis* as an isolated occurrence. We might probably ask "*What about it?*", meaning that we know what is being referred to, but we wonder why it has been uttered and that we expect an interpretant explanation, for example "*Psychoanalysis is the theory of the mind that was first formulated by Freud*". If we hear an isolated occurrence of *analysis of the mind*, we shall probably wonder which object or denominator the speaker is talking about. We might say something like "*What are you talking about*?", meaning that we do not know which N or O has been developed by the speaker into this interpretant. We might then expect something like "*Psychoanalysis is not the only theory that deals with the analysis of the mind*".

3.2 Reference and phrases

What is true for compounds such as *psychoanalysis* is also true for collocations. Since we have argued that every lexical item enters into collocational relations, it is no longer relevant to discuss syntactic frozenness or semantic opacity as the defining features of phraseological units, or for that matter to distinguish usefully between collocations and a variety of other expressions (idioms, lexical phrases and so on). Instead, what is of interest is the notion of reference. The test should be: does our expression refer globally to a social object or is it related to other denominators in an on-going discourse? If the latter is true, it is likely that our expression is an instantial, discursive feature of a text, i.e. an interpretant. The collocations *strong tea* and *powerful car* refer globally to socially existing complex objects, they are denominators. *Powerful tea* and *strong car* do not refer to socially existing objects, and so can only be seen as one-off mistakes or literary creations. Such expressions as *blow the gaff* and

kick the bucket on the other hand opaquely refer to the social objects of disclosing secrets and dying[4].

Spill the beans is a well-studied expression, because in lexicological circles there has been much debate about whether it is a collocation or an idiom (see for example van der Linden 1989). *Spill the beans* is a phraseological unit in our view because it refers to a specific social object: the disclosure of some secret. A query of *spill* in the *British National Corpus* produces 81 occurrences under the part-of-speech tag VVB (conjugated verb). *Spill the beans* occurs 11 times, *spill it* with the meaning of *tell me what you know* occurs twice, in the imperative only; the other occurrences referring ergatively to liquids or other entities spreading beyond certain limits (*the audience spill out on the road, the effects of alcohol misuse spill over from private life to the workplace*, etc.). *Spill* thus has the meaning of disclosure when occurring with the anaphoric expressions *it* and *the beans*, which refer to whatever is considered as the secret to be disclosed. There is also an occurrence of *spill the details*, which suggests that there is a consistent lexical cluster to the right of the verb. The expression *spill* {*it / the beans / the details*} is clearly more of an identifiable 'expression' than *spill blood, spill your drink* or *spill guts*, where *blood, drink* and *guts* are Ns which refer specifically to known objects. If one understands *spill* and *blood*, one understands *spill blood* on the basis of the normal lexico-grammar of English. This is a discursive construction or interpretant, and since it is often used, it is similar to the *want* collocation of the sort we discussed in section 1. Notice that our distinction starts off from the point of view of reference: there is no need to examine he semantic / syntactic properties of these expressions where one might be tempted to first posit that somehow *spill* or *beans* in *spill the beans* have a more 'restricted' sense than in *spill blood*. The semantic / syntactic search for restricted senses does not of course provide an explanation, and begs the question why some senses should be more 'restricted' than others. In our view, what matters is reference. *Spill the beans* has to be learned as a globally referring unit and when we have done so, there is limited scope for paradigmatic substitution, for example *spill it* and *spill the details*.

To conclude, expressions such as *kick the bucket* resemble the lexeme *psychoanalysis*, since both expressions refer globally. *Spill the beans* also refers globally, but accepts limited paradigmatic substitution. In this respect it displays similar semantic and syntactic features to the denominators *strong tea* and *powerful car*. It follows that reference is not achieved by fixed-size lexical units. It seems language provides overlapping pre-existing chains that we use within the bounds of their referential capacity.

3.3 Overlapping referential chains

We argued above that *strong tea* is a collocation. The expression is clearly not a creative piece of discourse. When we utter it we do not create an original linguistic segment out of the blue, we make use of an existing lexicalised, some would say conventionalised, unit. A search of *tea* in the BNC shows that *strong* is the most frequent adjective occurring with it (28 occurrences). Other adjectives include *weak* (16), *rich* (8), *ordinary* (6), *proper* (4), and a range of less frequent ones (*over-sweet, old, scalding, robust, rotten, Russian, real, quick, pleasant, ...*). There are even three occurrences of *powerful tea* in a text about semantics and vocabulary, explaining that this collocation is not acceptable.

Our point is that both *strong* and *tea* are denominators in their own right: they are able to refer independently in a variety of discursive contexts. This is also the case with *psyche* and *analysis,* which have an independent life of their own and may apply to a range of objects

[4] It is debatable whether *kick the bucket*, this arch-example of an idiom, has any productive use in English other than as a piece of metalinguistic evidence: the BNC produces only 7 occurrences, all in an explanatory context : two extracted from D.A. Cruse's *Lexical Semantics* and five from a university lecture.

within creative and original sentences. *Strong tea* is a denominator too. It is a piece of vocabulary endowed with a social existence, and used when tea is viewed specifically with reference to its concentration in tannins and caffeine. The difference between *strong tea* and *psychoanalysis* is that the latter does not allow for changes in its components, whereas *strong tea* opens a paradigm to the left, where a small set of grading adjectives may be used to reformulate the variable level of strength. If we refer to *strong tea*, we also acknowledge that *tea* may not possess this attribute much (it is *weak, ordinary*, …) or that it possesses it in some other way (*rich, robust, real, ...*). *Strong tea* resembles *spill the beans*: they both consist of a fixed lexical item (*tea, spill*) and a slot typically occupied by an archetypal lexical item (*strong, the beans*) which is the seed for a restricted reformulatory paradigm. The difference between them is that the *strong* paradigm is less restricted than *the beans* paradigm. Also, *the beans* functions like an anaphoric device, referring to a category of objects which have nothing to do with legumes, whereas *strong* is used within the boundary of its normal reference.

Why can we not use *powerful* with respect to *tea*? The simple answer is: why should we? We do not wonder about the reason why we do not use the idioms *spill the peas, spill the string beans* or *spill the legumes*, so why should we about *powerful*? The fact is that *powerful* and *strong* do not have the same usage. If they had we would not distinguish between them, and one of them would be doomed to extinction. It is entirely possible that in the history of the English language the expression *powerful tea* could have been adopted, as did thousands of expressions which do not exist today, but it did not. The question only arises because of a view of language as the product of a process that translates thought into language according to which the brain 'chooses' *strong* with *tea* and *powerful* with *cars*. In that case, what motivates this 'choice'? Could it be that these apparently synonymous words have different meanings after all? We are then prone to explain the difference in terms of a general semantic feature, for example that *powerful* and *strong* are only synonymous when used with abstract words such as *argument.*

But is speaking a calculation? The denominator for tea with a lot of tannins and caffeine is *strong tea*. When we think of *strong tea* our thought consists of the words *strong* and *tea*. Language is the substance of thought, not an overcoat flung over concepts. In that case we can abandon the commonly accepted view that the mind operates in two steps, with first the thought of tea and strength, and then the matching of the words *tea* and *strong* because of some semantic feature *strong* has in relation to *tea*.

So, once we have accepted a referential view of language, we find similarities between all Ns, whether they are mono- or polylexical units, whether they are fixed or whether they possess a slot where limited substitution is possible. According to such a view, even proverbs are Ns, as they refer globally to a category of objects of our experience (Kleiber 1994). *The early bird catches the worm* does not refer to some aviary behaviour. It applies to a whole category of human actions which are deemed beneficial when they take place. The reader may easily convince himself that such is the case. The sentence *Chandeliers hang from the middle of the ceiling* is a discursive utterance which refers to some feature of chandeliers. Yet, it is quite conceivable that it could become a proverb meaning for example that the brightest people are always the centre of attention. In that case its reference would be achieved by the whole sentence and not by the concatenation of the parts. As a consequence, the parts of a referential lexical item are not essentially material for the elucidation of the meaning of the whole. Their main role seems to act as prompts for recognition. They are the historical trace of the neological decisions that were made when the expression was first coined. As such, they may provide clues the first time we are exposed to it, but this is not essential. What matters is that the expression as a whole is recognised as a socially existing referential capacity.

3.4 Referential chains in text

A study of the following paragraph in Marcel Proust's *Du côté de chez Swan*, the first of the book, showed that Proust makes extensive use of existing collocations (see Frath 2005):

> Longtemps, je me suis couché de bonne heure. Parfois, à peine ma bougie éteinte, mes yeux se fermaient si vite que je n'avais pas le temps de me dire : «Je m'endors». Et, une demi-heure après, la pensée qu'il était temps de chercher le sommeil m'éveillait ; je voulais poser le volume que je croyais avoir encore dans les mains et souffler ma lumière ; je n'avais pas cessé en dormant de faire des réflexions sur ce que je venais de lire, mais ces réflexions avaient pris un tour un peu particulier ; il me semblait que j'étais moi-même ce dont parlait l'ouvrage : une église, un quatuor, la rivalité de François Ier et de Charles Quint.

Each of the words in this passage has been checked against a corpus of 35 mega-bytes of literary texts dating back to Proust's time. By way of illustration, let us examine the first two sentences, *Longtemps, je me suis couché de bonne heure (For a long time, I went to bed early),* and *Parfois, à peine ma bougie éteinte, mes yeux se fermaient si vite que je n'avais pas le temps de me dire : «Je m'endors»* (*Sometimes, as soon as my candle was out, my eyes would close so fast that I would not have the time to think: "I am falling asleep"*.). The corpus evidence shows that the first sentence consists of two overlapping collocations *de bonne heure* and *se coucher de bonne heure*, and one colligation, the initial position of *Longtemps*, which is a frequent place for temporal adverbs. In *Parfois, à peine ma bougie éteinte, mes yeux se fermaient si vite que je n'avais pas le temps de me dire : «Je m'endors»,* it was shown that this consists of the same temporal colligation as in the first sentence (the initial position of *Parfois*), and three overlapping collocations with paradigmatic substitutions: *à peine [+ Noun + Past Participle], si vite que [+ Clause], [avoir / laisser / donner / laisser +] le temps de [+ Verb]* as well as one co-occurrence: *les yeux ... se fermer.*

To understand collocations such as *se coucher de bonne heure*, we do not have to calculate their meaning from their components. We recognise them as existing denominators which refer to existing objects. We follow the text by activating the Ns and their reference in our own memory and we construct a complex object as we go. This object is a description of what happens when we fall asleep, also an element of our common experience. When a reader first gets acquainted with this passage, he may think, as was the case with the authors of this paper, that this is the first time he has read something about this very common phenomenon, falling asleep, which we experience every night. Proust's creativity does not come from his ability to translate thought into language - it consists of using existing denominators within interpretants to refer to a complex object that we recognise when the words have been uttered.

A similar study was carried out by Gledhill and Frath (forthcoming), in which the authors compared a phrase from Irvine Welsh's novel *Trainspotting* ("*The most wretched, servile, miserable, pathetic trash that was ever shat intae creation*") with usage in the British National Corpus. This NP is made up of three interrelated and overlapping structures: a series of stacked modifiers, a relative clause of comparison and a resultative construction. Although it looks novel, the originality of this expression does not come from the juxtaposition of new syntactic or lexical elements, nor from the translation into language of a pre-conceived thought. Irvine Welsh did not proceed by thinking that the Scots were the product of the excremental activity of some unspecified entity, a thought that he then translated into English. Instead, he made use of pre-existing lexical sequences which he integrated creatively. He in effect coined a new denominator, i.e. a public sign which refers to a new social object, the new found self-assertiveness and confidence of the Scots about their country, which paradoxically allows them to engage in self-deprecating humour. A search on the Internet showed that *shit into creation* has been used by other speakers to refer to this new cultural

fact, although sometimes imprecisely (*the most wretched miserable servile pathetic trash that was ever shat on civilisation / shagged into civilisation*). The expression has become a denominator of its own, a lexical chain which is part of an on-going discourse. The quest for the compositional 'glue' between *shit* and *creation* would be a hopeless task. These words were not put together to create a concept or item of meaning. They were juxtaposed in a strikingly effective expression that refers globally, i.e. to a Scottish political and literary reality. The engine of creativity is reference, i.e. our desire to speak of our experience, making use of existing vocabulary and collocations and of our habit of using them.

Conclusion

To sum up the arguments developed in this paper, polylexical units are not delineated by hidden underlying syntactic or semantic features endowed to particular lexical items. In this paper, we have defended a referential view of language. The boundaries of linguistic units are determined, not by an inner syntactic, semantic or logical 'glue', but by a link arbitrarily established between them and objects of our common experience. This link is not created on the spot as we speak, it is given to us by our community and has to be learned. When it is established, the linguistic unit (whether a complex word, phrase or clause) becomes a denominator, capable of a synthetic reference which does not involve knowledge. Knowledge is constructed in a second stage when denominators are used within interpretants. The components of such units are then only cues for recognition, not elements which have to be coded into a unit by the speaker and decoded by the listener. It could be argued that we have no direct contact with the world at large anyway, that whatever we perceive has to be somehow stored inside our brains for processing and that therefore it is a valid assumption to study language with respect to its mental aspects only. Yet, we do make a difference between outside objects and inner thoughts. When we speak about *dogs* for example, we do know there are such things as *dogs* in the outside world, independently of what we know about them. Indeed, reference should be a basic concern in linguistic research, and we believe the Peircian view we develop here could be particularly fruitful.

We have quite deliberately extended our discussion of phraseological units to units of texts. We believe that the approach adopted by many discourse analysts (following Halliday) is more realistic than the attempts of lexicographers to divide between 'free combinations' on one hand and 'fixed expressions' on the other. In our view, all lexical items 'collocate' and enter into lexico-grammatical relations. The only distinction that needs to be made is between denominators on the one hand and interpretants on the other.

References

Abeillé, A. (1995). 'The Flexibility of French Idioms: A Representation with Lexicalized Tree Adjoining Grammar'. In Everaert M., van der Linde, E-J., Schenk A. and R. Schreuder (eds.) *Idioms: Structural and Psychological Perspectives*. Hove, Lawrence Erlbaum Associates pp.15-42.

Cowie, Anthony (1994) 'Phraseology' in R. E. Asher (ed.) *The Encyclopedia of Language and Linguistics*. Oxford: Pergamon, pp 3168-71.

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

Cruse, D.A (1995) «Polysemy and related phenomena from a cognitive linguistic viewpoint». In *Computational Lexical Semantics*, Patrick Saint-Dizier & Evelyne Viegas eds. Studies in NLP. Cambridge University Press.

Di Scullio, A. M. & Williams, E. 1987. *On the Definition of Word*. Cambridge, Massachusetts, MIT Press.

Fernando, C. (1996) *Idioms and Idiomaticity*, Oxford, Oxford University Press.
Frath, Pierre (2005) : *Signe, référence et usage*. A paraître.
Gledhill, Christopher (2000a) 'The Discourse Function of Collocation in Research Article Introductions'. In *English for Specific Purposes.* Volume 19/2:115-135.
Gledhill Christopher (2000b) *Collocations in Science Writing*. Gunter Narr Verlag, Tübingen.
Gledhill, C. & Frath, P. (forthcoming). 'Une tournure peut en cacher une autre : L'innovation phraseologique dans *trainspotting*.'
Gibbs, R. (1985) 'On the Process of Understanding Idioms' in *Journal of Psycholinguistic Research* 14 : 465-77.
Halliday, M.A.K. & Hasan, Ruqaiya. (1976). *Cohesion in English*. London : Longman.
Hamm Albert, Frath Pierre & Rissanen Matti, eds. (2003): *Proceedings of the European Society for the Study of English (ESSE) Conference,* September 2002, Université Marc Bloch, Strasbourg.
Hoey, Mike. (1983). *On the Surface of Discourse*. London: George Allen and Unwin.
Howarth, P. (1996) : *A phraseological Approach to Academic Writing. Some Implication for Language Learning and Dictionary Making*. Tübingen: Max Niemeyer Verlag
Hunston, Susan & Francis, Gill (2000). *Pattern Grammar- A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam, John Benjamins.
Jackendoff, R. (1995) 'The Boundaries of the Lexicon'. In Everaert M., van der Linde, E-J., Schenk A. and R. Schreuder (eds.) *Idioms: Structural and Psychological Perspectives*. Hove, Lawrence Erlbaum Associates pp.133-165.
Kjellmer, G 1982 'Some Problems Relating to the Study of Collocations in the Brown Corpus' in S. Johansson (ed.) *Computer Corpora in English Language Research*, Bergen, Norwegian Centre for the Humanities.
Kleiber, Georges (1994) : «Sur la définition du proverbe». In *Nominales. Essais de sémantique référentielle*. Colin, Paris.
Langacker, R.W. (1984) : «Active zones», *Proceedings of the Annual Meeting of the Berkeley Linguistic Society*, 10, 172-188.
Makkai, Adam (1972) *Idiom Structure in English*. The Hague, Mouton.
Mel'cuk, Igor (1988a) *Dictionnaire Explicatif et Combinatoire du français contemporain*. Recherches lexicographiques II. Presses de l'Université de Montréal, Montréal.
Mel'cuk, Igor (1988b) : «The Explanatory Combinatorial Dictionary». In *Relational Models of the Lexicon*, Martha Walton Evans, ed. Cambridge University Press.
Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Paris, Gallimard.
Moon, R. (1998) *Fixed Expressions and Idioms: A Corpus-Based Approach.* Oxford, Oxford University Press.
Nattinger, J. & De Carrico J. (1992) *Lexical Phrases and Language Teaching,* Oxford, Oxford University Press.
Nesselhauf, Nadja (2003) 'The Use of Collocations by Advanced Learners of English and Some Implications for Teaching' in *Applied Linguistics* 42/2 : 223-42.
Peirce, Charles Sanders (1978) *: Ecrits sur le signe*, rassemblés, traduits et commentées par Gérard Deledalle, Seuil.
Pinker, Steven. 1994. *The Language Instinct*. London, Penguin.
Pustejovsky, James (1991) : *The Generative Lexicon* . Computational Linguistics 17(4).
Sinclair, John (1966) 'Beginning the Study of Lexis' in C.E. Bazell, J.C. Catford, M.A.K. Halliday and R.H. Robins (eds), *In Memory of J.R. Firth* London, Longman : 148-62.
Sinclair, John (1987) : *Looking up : An Account of the COBUILD Project in Lexical Computing*. John Sinclair, *ed*. Collins Cobuild.

Smadja, Frank (1989) 'Co-occurrence: the Missing Link', in *Literary and Linguistic Computing* 4/3:163-68.
Swinney, D. & Cutler, A. (1979) 'The Access and Processing of Idiomatic Expressions' in *Journal of Verbal Learning and Verbal Behaviour* 18: 523-34.
Van der Linden, E-J., (1989) 'Idioms and Flexible Categorial Grammar' in M. Everaert and E-J. van der Linden (eds.) *Proceedings of the First Tilburg Workshop on Idios*. ITK, Tilbrg.
Van Roey, J. (1990) : *French-English Contrastive Lexicology: An Introduction*. Louvain-la -Neuve: Peeters.
Wittgenstein, Ludwig (1963): *Philosophical Investigations*. Basil Blackwell, Oxford