Version de travail

Paru dans :
Léon J. 2007. " Meaning by collocation. The Firthian filiation of Corpus Linguistics » Proceedings of *ICHoLS X, 10th International Conference on the History of Language Sciences*, (D. Kibbee ed.), John Benjamins Publishing Company :404-415

JACQUELINE LÉON
CNRS, Université Paris 7

MEANING BY COLLOCATION. THE FIRTHIAN FILIATION OF CORPUS LINGUISTICS

The works which refer to John Rupert Firth can be classified in two series. First, shortly after Firth's death in 1960 (Robins 1961), the publication of *In Memory of Firth* gathering essays by his colleagues and pupils (Bazell and al. 1966), was immediately followed by the publication in 1968 of the second part of Firth's works by Frank Palmer (Palmer 1968). Then monographies on Firthian or Neo-Firthian linguistics were published in the 1970s (Langendoen 1968 ; Mitchell 1975 ; Monaghan 1979).

The second series of works referring to Firth corresponds to the revival of so-called Corpus Linguistics in the 1990s, that is computerized corpora based studies[1]. Two stances can be observed within British Corpus Linguistics regarding Firth's work, although both have the London School as a common background. John Sinclair and his followers have never stopped referring to Firth's work, while the Randolph Quirk-Geoffrey Leech line of development completely ignored Firth's legacy and chose the American Brown Corpus as a pioneer instead (Stubbs 1993 ; Léon 2005).

While the works of the late 1960s referred both to Firth's main contributions, phonology and semantics, Corpus Linguistics only adressed collocations referring to Firth by quoting very short excerpts from one of Firth's last papers written in 1957 (Firth [1957f] 1968): « You shall know a word by the company it keeps » and « collocation as actual words in habitual company » which have been repeated from papers to papers (see for example Mackin 1978 ; Sinclair 1991 ; Stubbs 1992 ; Hanks 1996 ; Kennedy 1998 ; Tognini-Bonelli 2001) even in corpus linguists's papers coming from completely different backgrounds, such as computational linguistics or natural language processing (Church & Mercer 1993 ; Manning & Schütze 2002).

1 See also more recent theoretical surveys on Firthian linguistics : Beaugrande (1991), Butt (2001), Henderson (1987), Palmer (1994), Robins (1998) and Stubbs (1993).

In my paper, I would like to address the issue of collocation works, starting from Firth's view on meaning by collocation, in order to see how it has been worked out by Corpus Linguistics works. I will especially focalize on the issue of corpus itself, which had been hardly addressed by Firth himself, and is still a tricky issue, even in an empiricist view of linguistics.

1. ***Meaning by collocation***

Meaning by collocation was first conceived in Firth's 1935 paper « The Technique of Semantics » as lexical meaning, one of his five dimensions of meaning (phonetic, lexical, morphological, syntactic and semantic). Later lexical meaning was developed under the name of 'meaning by collocation' in three papers essentially, « Modes of meaning » of 1951 was published in 1957 by Firth himself in his Papers in Linguistics. The second paper « Linguistic analysis as a study of meaning » written in 1952 was never published during Firth's life and was later published in 1968 by F.R. Palmer so as the third paper « A synopsis of linguistic theory 1930-55 » first published in 1957.

Meaning by collocation is closely related to Firth's main theoretical principles such as his polysystemic approach of meaning[2]: « The basic principle, first stated in the Technique of Semantics, is a dispersion of meaning at a series of congruent levels of analysis, at each one of which statements of meaning are made in linguistic terms. » (1957:xi) ; the notion of context of situation[3], borrowed from Malinowsky, Wegener and Gardiner ; his concern for linguistic applications, and finally the centrality of language use and attested texts for linguistic analysis.

When considering Firth's view on collocation, it should be said that it may seem contradictory. Even in the same text, collocation may appear as a mere word phenomenon, or may be connected to any level of language.

(1)
« Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of night is its collocability with dark, and of dark, of course, collocation with night » (Firth [1951] 1957 :196).

---

2 Firth's polysystemic analysis is a criticism of the structuralist view, especially Meillet's, of language as a one system whole. 'Polysystemic' refers to 'multilevel' and to 'multistructural' and is associated with restricted languages: « Linguistic analysis must be polysystemic. For any given language there is no coherent system (où tout se tient) which can handle and state all the facts. » ([1952] 1968 :24).
« It is unnecessary, indeed perhaps inadvisable, to attempt a structural and systemic account of a language as a whole. Any given or selected restricted language, i.e. the language under description is, from the present point of view, multi-structural and polysystemic. » ([1957] 1968 :200).

3 In Firth's view, context of situation is an « abstract construct, a set of categories » which assert attested texts as theoretical objects.

What appears in excerpt (1), which is most often quoted as the definition of collocation, is that meaning by collocation implies mutual expectancy and that Firth's argument is based on words : one of the meanings of 'night' is its collocability with 'dark', and of 'dark', its collocation with 'night'. Meaning by collocation is a syntagmatic phenomenon, which means that it should be analyzed at text level. However collocations are not merely the juxtaposition of words : « The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of mutual expectancy. » ([1957f] 1968 : 181). Finally as an abstraction it must be inductively inferred from data, in accordance with Firth's strong attachment to empiricism.

2. *Collocations and stylistics*

It should be said that Firth's own and single field of application for meaning by collocation was stylistics and that it is through stylistics that he introduced the notion in 1951. In « Modes of Meaning », Firth studies the idiosyncratic language of Swinburne's poems and the stylistics of what persists in common usage over long periods through the study of English letters written in the 18th and 19th centuries. In «Linguistic analysis as a study of meaning » ([1952], 1968), he studies Edith Sitwell's poems. This type of application presents two issues which are crucial to understand Firth's conception of collocation : in his stylistic studies it appears that collocations cannot be reduced to « word company ». Besides the crucial role of text is emphasized.

Firth introduced his notion of collocation from the application of the phonoaesthetic function and phonologic meaning (or prosodic mode) to Edward Lear's limericks. Limericks are forms of nonsense verses the rhymes of which can be expected by the initiated at the prosodic, grammatical, stylistic and social levels[4].

(2)
« At this point in my argument, still confining our references to the language of limericks, I propose to bring forward as a technical term, meaning by 'collocation' and to apply the test of 'collocability' (Firth [1951] 1957 :194).

The phonoaesthetic function of sounds, coined in his first works *The Tongue of men* and *Speech*, is the « association of sounds (alliterations) and personal and social attitudes ». The phonoaesthetic function allows him to avoid « the misleading implications of onomatopoeia and the fallacy of sound symbolism». It is illustrated by the alliterative use of 'str-' in Swinburne's poems 'straightening', 'streamers', 'straining' :

(3)
« Ah the banner-poles, the stretch of straightening streamers

4 « Once started on a limerick, there are modal expectancies for the initiated at all these levels, at the grammatical, stylistic, and indeed at a variety of social levels. » ([1951] 1957 :194).

Straining their full reach out ! » (Firth [1951] 1957 :194).

Considering Firth's early interests in phonetics, it is not surprising that collocation was first devised from phonetical issues. Thus, in his *Papers in Linguistics*, ten papers out of sixteen are explicitely about sounds, from experimental phonetics to phonology.

Collocation plays at several levels. Expectations are not only at the phonetic or prosodic levels, but are also at grammatical, lexical and stylistic levels. At the grammatical level, Firth gives the example of the repetition of exclamatory verses beginning by 'Ah the' and ending with a mark of exclamation in Swinburne's poems ; in the letters, he notes the recurrent structure of participial construction in '-ing' and '-ed' preceded by a personal pronoun such as 'my using it' in 'would there be any harm in my using it?'.

Collocations do not only concern words but phrases, compounds, turns of phrases, or any expressions. They also concern inferior units such as morphemes, continuous or discontinuous. It could be said that, at the syntagmatic level, the meaning of any stretch of text results from the collocation with any other stretch in the same text. See in particular what Firth tells us about two verses of one of Edith Sitwell's poem, where two noun phrases 'Emily-coloured primulas' and 'Bob in their pinafores on the grass' collocate only because they occur in the same verse :

(4)
« According to my analysis part of the meaning of *Emily-coloured primulas* is collocation with *Bob in their pinafores on the grass*. This level I have termed meaning by collocation, which may be personal and idiosyncratic, or normal. » (Firth [1952] 1968 :18)[5].

In this case, Firth's use of 'collocation with' seems very close to the common use of collocation dating back to the beginning of the 17th century, that is « a disposition or arrangement with, or in relation to others »:

(5)
Collocation : 1605
[ad L. collocationionem] The action of setting in a place or position; disposition or arrangement with, or in relation to, others ; the state of being so placed. (*The Shorter Oxford English Dictionary* 1962)

3. *You shall know a word by the company it keeps*

5 « For spring is here, the auriculas
And the Emily-coloured primulas
Bob in their pinafores on the grass. ».

In order to emphasize the crucial role of text, Firth had recourse to quotations from Wittgenstein's *Philosophical Investigations* (1953)[6]. See quotations (6) to (9) :

(6)
« The *placing* of a *text* as a constituent in a context of situation contributes to the statement of meaning since situations are set up to recognize *use*. As Wittgenstein says, 'the meaning of words lies in their use.' (Phil. Investigations, 80, 109). The day-to-day practice of playing language games recognizes customs and rules. It follows that a text in such established usage may contain sentences such as 'Don't be such an ass !', 'You silly ass !', 'What an ass he is !' In these examples, the word *ass* is in familiar and habitual company, commonly collocated *with you silly-, he is a silly-, don't be such an-*. **You shall know a word by the company it keeps !** One of the meanings of *ass* is its habitual collocation with such other words as those above quoted. Though Wittgenstein was dealing with another problem, he also recognizes the plain face-value, the physiognomy of words. They look at us ! 'The sentence is composed of words and that is enough' [note 46 : Wittgenstein,1953 :181]. » ([1952], [1957f] 1968 : 179).

(7)
« **Collocations are actual words in habitual company.** A word in a usual collocation stares you in the face just as it is. Colligations cannot be of words as such. Colligations of grammatical categories related in a given structure do not necessarily follow word divisions or even sub-divisions of words. » (Firth [1957f] 1968 : 182).

(8)
«The elements of style can be stated in linguistic terms. They are formally presented in the text which can be said to have a physiognomy. » (Firth [1957f] 1968 : 195)

(9)
« Words stare you in the face from the text, and that is enough; and as Wittgenstein said, a word in company may be said to have a physiognomy. The elements of style can be stated in linguistic terms. » (Firth 1957 : xii).

I will not comment on the relevance of these quotations with respect to Wittgenstein's theory, in so far as Firth himself acknowledged that their respective aims, his and Wittgenstein's were quite different (see quotation 6). One need only note that Firth uses Wittgenstein's excerpts in order to

6 « The meaning of words lies in their use » (Wittgenstein 1953 :80). « One cannot guess how a word functions. One has to look at its use, and learn from that » (Wittgenstein 1953 : 109). He likens the practice of various types of languages in speech behaviour to games with rules. « A language is a set of games with rules or customs » (Wittgenstein 1953 : 47 and 81). Actually these quotations occur in two of Firth's last papers ([1957e] 1968 :138) and ([1957f] 1968 :179, 182, 195).

legitimize the study of language in use, the study of collocations of words at the syntagmatic level, and the settlement of stylistics on linguistic basis.

Moreover, reference to Wittgenstein leads him to distinguish between collocation and colligation, that is word collocations from grammatical collocations (quotation 7). The meaning at the grammatical level is in terms of word classes and of the interrelation of those categories in colligations. Colligations may include discontinuous or cumulative morphemes, such as negation, or case + person+ plural + tense :

> (10)
> Grammatical relations should not be regarded as relations between words as such - between *watched* and *him* in 'I watched him' – but between a personal pronoun, first person singular nominative, the past tense of a transitive verb and the third person pronoun singular in the oblique or objective form. These grammatical abstractions state some of the interrelated categories within an affirmative sentence. (Firth [1957f] 1968 :181, note 49).

He advocated to treat each word form separately and not as a single lemma :

> (11)
> « The collocations of *light* (n.s.) separate it from *lights* (n.s.) and *light* (n.adj.) from *lighter* and *lightest*. Then there are the specific contrastive collocations for *light / dark* and *light /heavy*. » ([1957f] 1968 :180).

Finally he relies upon these quotations to settle stylistics on linguistic basis and assert the crucial role of text. Firth repeatedly stated in his work that « the text is the main concern of the linguist »[7]. Thus, by taking up Wittgenstein's quotations, Firth gives text a specific place : « Words stare you in the face from the text » (quotation 9), and it is the text that has a physiognomy (quotation 8). Note the shift from word to text : Wittgenstein's physiognomy of words becomes the physiognomy of texts[8].

Embedded in these quotations can be recognized the excerpts coined as slogans by corpus linguists : « You shall know a word by the company it keeps » and « Collocations are actual words in habitual company ». It should be said that Mackin is the first to use this quotation in his 1978 article "On Collocations: 'Words shall be known by the company they keep'". Mackin is

---

7 See for example, « Processes and patterns of life in the environment can be generalized in contexts of situation, in which the text is the main concern of the linguist. » ([1952] 1968 :24) ; « language texts which are the linguist's main concern » Linguistics and translation ([1957] 1968 90); « the focal constituent for the linguist being the text. » (1968 [1957] : 173).

8 See Wittgenstein's text : « The meaning of a word is not the experience one has in hearing or saying it, and the sense of a sentence is not a complex of such experiences.- (How do the meanings of the individual words make up the sense of the sentence « I still haven't seen him yet » ?). The sentence is composed of the words, and that is enough.
Though - one would like to say—every word has a different character in different contexts, at the same time there is one character it alway has : a single physiognomy. It looks at us.-But a face in a painting looks at us too. » (Wittgenstein 1953 II-vi :181).

one of the authors of *the Oxford Dictionary of Current Idiomatic English Using* based on collocations. Using these quotations, dictionary makers and corpus linguists restrict collocations to words while, as we have seen, Firth's view encompassed any stretch of text.

4. ***Methodological issues : text, corpus, restricted languages***

In fact, this move was initiated by Firth himself when he put forward methodological indications to study collocations of words in order to achieve practical aims, such as dictionaries and translation:

> (12)
> An approach to the meaning of words, pieces, and sentences by the statement of characteristic collocations ensures that the isolate word or piece as such is attested in established texts. The characteristic collocations of 'key' or 'pivotal' words may be supported by reference to contexts of situation, and may constitute the material for syntactical analysis and provide citations in support of dictionary definitions. (Firth, 1957 :xi).

Note that Wittgenstein's quotations and methodological indications both appear in « A synopsis of linguistic theory 1930-55 » (1968, [1957]), one of Firth's last ones.

To focalize collocation on words may seem contradictory, because Firth did not agree to regard words as linguistic units. As was specified in quotation (1), he did not accept the equation of 'lexical' with 'semantic' and was utterly opposed to a conceptualist, logical or psychological approach « which treats words and sentences as if they have meanings by themselves » (1968, [1952] :19) just as he was opposed to the 'one morpheme one meaning' of the Neo-Bloomfieldian approach. In fact, Firth preferred pieces, 'combinations of words', to isolated words as linguistic units. Yet the method specifies that the collocations of selected 'key' or 'pivotal' words should be searched in whole attested texts. Once pinpointed, collocations should be grouped into sets where words are arranged in ordered series. A set of collocations is assumed to help establishing the meaning of a word. It is also a method to systematize the use of quotations in dictionaries.

In Firth's work, text and corpus are equivalent : he based his stylistic investigations on finite sets of texts : the complete works of an author (Swinburne's poems) or an homogeneous set of letters of the 18th and 19th centuries. No need of generalization then, since, as far as stylistics is concerned, idiosyncratic collocations are even more significant results than 'habitual collocations'.

The issue of corpus was nevertheless raised when Firth tackled other types of applications for collocation studies, such as lexicography and dictionary making. He put forward the idea of 'restricted languages' in order to circumscribe the field within which collocations should be studied.

One consequence of Firth's insistence on the polysystemic nature of language is his stress on its non-homogeneity. There is no such thing as 'one language' (see note 2). The task of descriptive linguistics is not to study the language as a whole, general language, but to study restricted languages, more manageable. A restricted language limits and circumscribes the field of linguistic investigation and is sufficient to state coherent grammatical structures: « A restricted language can be said to have a *micro-grammar* and a *micro-glossary*. » (Firth [1957c] 1968: 106).

The range of restricted languages is large : from scientific and technical languages to literary texts, and may even be limited to one author and sometimes to one manuscript. From a methodological point of view, it is rewarding to investigate collocations in restricted languages :

> (13)
> Statements of meaning at the collocational level may be made for the *pivotal* or *key words* of any *restricted language* being studied. Such collocations will often be found to be characteristic and help justify the restriction of the field. The words under study will be found in 'set' company and find their places in the 'ordered' collocations. (Firth [1957f] 1968 :180).

> (14)
> In the study of selected words, compounds and phrases in a restricted language for which there are restricted texts, an exhaustive collection of collocations must first be made. It will then be found that meaning by collocation will suggest a small number of groups of collocations for each word studied. The next step is the choice of definitions for meanings suggested by the groups. (Firth [1957f] 1968 :181).

Restricted languages are more adapted than general language to carry out practical objectives, encompassing language teaching, dictionaries, translation, international languages.

> (15)
> *The study of English* is a very vague expression referring to a whole universe of possibilities which must be reduced and circumscribed to make exact study and disciplined teaching possible. Hence the notion of a *restricted language*. Restricted languages function in situations or sets or series of situations proper to them, e.g. technical languages such as those operative in industry, aviation, military services, politics, commerce or, indeed, any form of speech or writing with specialized vocabulary, grammar and style. (Firth [1957c] 1968 : 112).

Here again comes the issue of text : the restricted language must be exemplified by texts. So that text will remain a key issue for the Firthian line of thought[9].

[9] « The linguist operates with *language* and *text*, the latter referring to all linguistic material, spoken or written, which we observe in order to study language. The linguist's object of study

(16)
The restricted language, which is also called the language under description (beschreibene Sprache) must be exemplified by texts constituting an adequate *corpus inscriptionum*. (Firth [1957c] 1968 : 112).

This is one of the very few places where Firth mentions the word 'corpus', always used in its Latin collocation *corpus inscriptionum*. He mentions corpus the first time when referring to the analysis of spoken languages:

(17)
In the study of spoken language … In support of any linguistic analysis formally presented, there should always be texts. It is perhaps never possible nor desirable to present the whole of the material collected during the observation period, but some sort of 'corpus inscriptionum' seems to me essential in almost all studies. (Firth [1957a] 1968: 32)

However he did not give any clue how to gather texts into corpora. Although one key issue in Firth's empiricist model is the idea that linguistic theoretical ideas should be tested repeatedly against real language, he never mentioned statistical counts nor lexical probabilities on which corpora-based studies are grounded. Even if the notions of collocability and mutual expectancy may imply some kind of probability or potentiality.

One last remark should be make concerning the use of machines. Firth never really addressed the issue of computers for the treatment of collocations. Although he wrote two papers on translation and linguistics published in Palmer (1968) and advocated the use of collocations and restricted languages for translation purposes, he only mentioned machine translation when discussing the relevance of interlingua based on 'naked ideas' for 'linguistic engineer' purposes (see section 5. below).

On the other hand, Firth was much interested in phonetic machinery, most notably in technological advances in kymography and palatography, and wrote several articles on this matter. Thus the machines mentioned in the first sentence of quotation (18) are phonetic machines, which is confirmed by the next sentence dealing with phonetic laboratories.

(18)
The use of machines in linguistic analysis is now established. The present approach prefers to take linguistics into the laboratory rather than to look into laboratories for linguistics. (Firth [1957f] 1968 :202)

Therefore it could be quite misleading, as some corpus linguists did, to head a paper on Corpus Linguistics by the first part of Firth's quotation followed by a quotation from Sinclair's referring to large computer corpora :

---

is the language and his object of observation is the text : he describes language, and relates it to situations in which it is operating ». (Halliday 1960 :l8)

(19)
The use of machines in linguistic analysis is now established. (Firth 1968)
It is my belief that a new understanding of the nature and structure of language will shortly be available as a result of the examination by computer of large collections of texts. (Sinclair 1991 :489)
(Stubbs 1993 :1)

It may induce the reader to believe that Firth's 'machines' were 'computers' and that he had an early interest in computational linguistics and in computer-based corpora, whereas only his followers, most notably John Sinclair, devised systematic computational studies of collocations.

To conclude this section, it should be said that Firth's followers using computers focalized on word collocations, and most of the time pairs of words, while, as seen before, Firth's collocation level refer to phonoaestic, prosody, turns of phrase as well as words. It should also be remembered that Firth spoke of pieces, that is combinations of words, more than single words as linguistic units.

5. *On British traditions of lexical studies*

There is a strong tradition of lexical studies in Britain since Cruden's work on co-occurrences in the Bible in the eighteenth century. Concerning collocations, some corpus linguists (Mitchell 1975 ; Kennedy 1998 ; Sinclair et al. 2004) regard Harold E. Palmer as a precursor although he was never quoted by Firth, Halliday or Sinclair in the1950-60s[10]. As an English language teaching specialist in Japan in the 1930s, Palmer undertook corpus-based research on recurrent combinations of English words, the outcome of which was a report (Palmer H.E. 1933) and books of English as a foreign language, such as « A Grammar of English Words. One thousand English words and their pronunciation, together with information concerning the several meanings of each word, its inflections and derivatives, and the *collocations* and phrases into which it enters » (Palmer H.E. 1938). Palmer's collocations are nothing else than combinations of words, and it is not surprising that he is quoted as a precursor by corpus linguists whose conception of collocation has been reduced to word collocations. Note that Palmer could also have been considered the precursor of lexicogrammar ; in his 1938 book «a grammar of English words », he distinguishes the grammar of forms treating of grammatical categories, from the grammar of words treating of grammatical usages specific to each word.

Another British lexical tradition at work in Firth's meaning by collocation is thesaurus models dating back to the 17th century universal language schemes.

10 It is not impossible that Palmer and Firth met at the University College of London where Palmer was assistant in the phonetic department in 1917, and where Firth was appointed senior lecturer of phonetics in 1928.

Halliday, while appointed as assistant lecturer in Chinese at Cambridge, joined the Cambridge Language Research Group and worked on the faisability of machine translation using thesaurus methods (see Léon, to be published). The CLRU was directed by one of Wittgenstein's pupils, Margaret Masterman, convinced that the logic unit for studying language should not be word nor proposition but word context, namely word use. The CLRU planned to create a new intermediary language based on the classification of words according to a set of contexts, and chose Roget's thesaurus[11]. When Firth was invited in 1955 to the first meeting of the Cambridge Language Research Group, he discussed Richens's interlingua *Nude* based on semantic primitives called 'naked ideas' (CLRU 1956). Opposed to any form of *a priori* semantics and universal languages, he advocated a polysystemic approach of meaning:

> (20)
> The problems of stating meaning in linguistic terms are more manageable if we distinguish between two of many possible methods of approach. First the approach of the 'linguistic engineer' who hopes to arrive a the mechanism of rendering material in one language, the source language, into a second language, the target language. It may well happen that recourse to some theory of 'naked ideas' will at first prove attractive. Where are the materials for the bridge to be found? Presumably in some sort of analytical segmentalized dictionary based on units of meaning whatever they might be … The second method of approach is by linguistic analysis. This proceeds on the assumption that language is polysystemic, and that multiple statements of meaning in linguistic terms can be made at a series of congruent levels. (Firth [1956] 1968 :81)

Halliday (1958) discussed thesaurus as a way to describe and systematize the lexis for machine translation, showing that in Roget's *Thesaurus of English Words and Phrases*, divisions (sub-paragraph, paragraph, section etc.) correspond to different ranges of context, but that collocations have advantages over thesaurus lists. In Firthian linguistics, the meaning of a lexical item includes the set of lexical items with which it habitually collocates, or co-occurs, while Roget's thesaurus lists of words are predetermined semantic notions.

Later, Sinclair (1966) also addressed the difference between thesaurus lists and lexical sets based on collocations. Among the three words, 'tome', 'paperback' and 'cruelty', the first two share a non-linguistic notional similarity, and they occur in the same paragraph in Roget. But they may show no special tendency to cooccur, while 'tome' and 'paperback' may share collocations with other words like 'edition', 'bookshop', 'paper' or 'print'.

6. *Pioneer computer-based studies of collocations*

[11] note that Peter Mark Roget (1779-1869) quotes Wilkins' *Real Character* and is considered as one of his continuators.

Now let us examine how meaning by collocation has been handled by Firth's followers. McIntosh (1961) and Halliday and McIntosh (1966) compared grammatical and lexical patterning, stating that, as lexical items have only a certain potential of collocability, a lexical set - the set of lexical items which have collocations in common - is the nearest lexical equivalent of a grammatical pattern.

However, Halliday was the one who really set up the theoretical conditions of studying Firth's meaning by collocation systematically. In his early works, in « Categories of the theory of grammar » published in *Word* in 1961, and in his 1966 « Lexis as a Linguistic Level », Halliday included meaning by collocation into his own model, so that Firth's collocational level became lexical level again, as it was stated first in 1935[12]. Halliday proposed to study lexical patterning in language in the light of lexicogrammar. The assumption was that there were no strict boundaries, as was claimed at that time, but a continuum between lexis and grammar. Lexis was conceived in his first full scale model, as 'most delicate grammar'. In particular, he put forward the category of lexicalness to parallel that of Chomsky's grammaticalness. More generally, the notion of collocation allowed him to question the Chomskyan exclusive opposition grammatical / ungrammatical in order to introduce the idea of degree of acceptability[13]. The argumentation presented in his 1966 paper is quite enlightening on his view on collocation, lexicalness, and especially predictibility as a language feature. As he will state later, the linguistic system is inherently probabilistic. He opposed the distribution of 'strong' and 'powerful'. Even if one can say 'a strong argument' or 'a powerful argument', 'strong' does not always stand in this same relation to 'powerful' :

(21)
he drives a powerful car
* he drives a strong car
this tea is too strong
* this tea is too powerful

(22)
To put it another way, 'strong car' and 'powerful tea' will either be rejected as ungrammatical (or unlexical) or shown to be in some sort of marked contrast with a 'powerful car' and 'strong tea' ; in either case the paradigmatic relation of 'strong' to 'powerful' is not a constant but depends on the syntagmatic relation into which each enters, here with argument, car or tea. …
What is abstracted is an item 'strong', having the scatter 'strong, strongly, strength, strengthened', which collocates with items 'argue' (argument) and

12 Halliday mentions collocations as early as 1957 in one of his first papers « Some Aspects of Systematic Description and Comparison in Grammatical Analysis ».
13 During the debate which followed Chomsky's talk at the 9th Congress of linguists in 1962, Halliday acknowledged the interest of grammaticalness on condition that it was expressed in terms of degree and not of exclusivity between well-formed and ill-formed sentences, and that it was completed by lexicalness (see Chomsky, 1964 : 989).

> 'tea' ; and an item 'power' (powerful, powerfully) which collocates with 'argue' and 'car'. It can be predicted that, if 'a high-powered car' is acceptable, this will be matched by 'a high-powered argument' but not by 'high-powered tea'. It might also be predicted, though with less assurance, that 'a weak argument' and 'weak tea' are acceptable, but that 'a weak ca'r is not.[14] (Halliday 1966 :150f.)

Halliday adds a paradigmatic dimension to Firth's only syntagmatic view on collocations. Especially he introduces a probabilistic turn which is crucial for computational corpora studies. The property of mutual expectancy belonging to Firth's definition of collocation has been reinterpreted in terms of lexical patterning, where the tendency of words to occur in the vicinity of each other is not predicted by chance.

Actually, Halliday himself never achieved computer research on collocations and no mention of collocation can ever be found in his work after 1966, though he sometimes joined Corpus Linguistics publications on probabilistic grammar in the 1990s (Svartvik 1992). However he encouraged his young colleague John Sinclair to develop collocation study using computer methods.

John Sinclair's 1966 paper was his first published paper on lexis when he was working on the OSTI project (Office for Scientific and Technical Information) initiated in 1963. The project was devised in consultation with Angus McIntosh, MAK Halliday, and another former member of the CRLU, Roger Needham, his elders at the University of Edimburgh who remained members of the steering committee. Sinclair wrote the final report in 1970 (Sinclair and al. [1970] 2004).

It should be said that, at that time, works on computational lexicography, and particularly the achievement of concordance tools, were undertaken in several countries In 1957, a conference about 'Lexicologie et lexicographie françaises et romanes' took place in Strasbourg to study the faisability of a Dictionary of Modern and Contemporary French (1789–1960) based on concordances from a computerized Thesaurus or 'Trésor de la Langue Française' (TLF), a corpus of 1350 literary or technical books written from 1789 to 1960. Subsequently, the Center of TLF was created by the CNRS (Centre National de la Recherche Scientifique) in 1960 (CNRS 1961). The British scholars P-J. Wexler and Stephen Ullman attended to the conference, and in 1963, Halliday and Sinclair visited the Laboratoire d'analyse lexicologique, created in 1959 by Bernard Quemada. In 1974, Sinclair

[14] It is worth mentioning that the definition of collocation given by Hornby and al. in their *Oxford Advanced Learner's Dictionary of Current English* (1974) already takes Halliday's considerations into account, especially the 'strong tea' example : « Collocation : Collocate (of words) combine in a way characteristic of language : 'weak' collocates with 'tea' but 'feeble' does not. Collocation coming together ; collocation of words ; 'strong tea' and 'heavy drinker' are English collocations. So are 'by accident' and 'so as to'».

published a paper on collocations in the *Cahiers de lexicologie* (John & Sinclair 1974).

In his 1966 paper, Sinclair essentially investigates basic methodological issues of computer-based collocation studies such as node, span and collocates. 'Node' refers to the lexical item under study ; 'span' refers to the number of lexical items on each side of a node[15]. Items in the environment set by the span are called 'collocates'. By specifying the definition of lexical items and collocations, he contributed to the discussion on word as a linguistic unit ; he defined a lexical item as a unit of language representing a particular area of meaning which has a unique pattern of co-occurrence with other lexical items. Significant collocation is regular collocation between two items, such that they co-occur more often than their respective frequencies. The aim is to provide a description of the lexical items of a text in terms of their collocational patterns. Sinclair extended Halliday's lexicogrammar by specifying the interdependence of lexis and syntax. For example, the word *lap* is more likely to occur in a prepositional phrase in adjunct position, than to occur as the subject or object of a clause (see Stubbs 1993 on this point).

In a recent interview (Sinclair and al. [1970] 2004), Sinclair acknowledges that, in 1966, he had no real conception of corpus, and contented himself with looking for collocations in some conversational English recorded for the purpose[16]. He notes that, at that time, the techniques of analysis were extremely crude, as they were 'perforce based on the poor capabilities of machines'. Thus he stopped working on computational lexicology after the publication of the OSTI report in 1970 in order to devote himself to discourse analysis. He resumed corpora and computational lexicology only in the 1980s with practical issues such as the making of dictionaries and grammars based on collocations, most notably the COBUILD project.

7. *Collocations and large corpora*

Computational lexicology gained new vitality in the 1990s when computer power and data availability increased, in the wake of signal and speech recognition using probabilistic methods. Although the famous definition of collocation as « actual words in habitual company » is used to rally corpus linguists who claim a Firthian inheritance, the term 'collocation' often refers to many different things such as 'habitually co-occurring lexical items', 'recurrent word combination', 'repeated word co-occurrences' or 'multiword units which have an independent existence' (Altenberg & Eeg-Olofsson 1990). In recent computational linguistic works (Manning & Schütze 2002), collocations are identified as idioms « any turn of phrase or accepted usage where somehow the

15 Since the 1980s, and still now on, the optimum extent has been settled as four words on either side of the node, in order to grasp discontinuous collocations. It is acknowledged that this criterion remains arbitrary, as it is often the case for Natural Language Processing features.
16 Strangely he does not mention Quirk's SEU although he used recordings from the UCL.

whole is perceived to have an existence beyond the sum of the parts ». Some researchers have investigated collocations under other names such as Biber (1996) claiming that his association patterns are an extension of Firth's notion of collocation.

Actually, many questions are raised by computer-based collocation research which still remain unsolved. Some of them have already been adressed by Firth, such as discontinuous collocations (words which frequently occur in each other's company are not necessarily contiguous) ; he also advocated not to lemmatize items in order to count flections as different collocations in a first step (see excerpt 11). Most significant questions have also been addressed by Kennedy (1998 :111):

How often does a combination have to recur to be 'habitual', and who decides what 'sounds natural' ? Does a combination have to be 'well-formed' or canonical to be a collocation (e.g. *Wannanother one* ?) ?

Can a sequence which occurs only once in a particular corpus but which is intuitively recognized by native speakers as a sequence they have heard before, be listed as a collocation nevertheless ?

How big does a corpus have to be in order to establish that a collocation does exist ?

Altenberg and Eeg-Olofsson (1990) pointed out that grammatically and semantically these expressions exhibit varying degrees of stability : some are lexicalized (*to spill the beans*), others rule-governed (*at the same time*) ; some are neither quite lexicalized nor formed by normal grammatical rules (*the sooner, the better*), others are partly memorized wholes and partly free constructions (*I am sorry, that's right*).

In spite of these numerous problems, Firthian corpus linguists consider that collocations have strong theoretical interests. In addition to Halliday's pioneering assumptions on loose boundaries between lexis and grammar, on lexicalness and on the probabilistic nature of language, corpus linguists claim that collocations question Chomskyan's views on language creativity and intuition.

Opposing the idiom principle to the 'open-choice principle' Sinclair (1991) assumes that speakers use ready-made linguistic forms, or prepackaged chunks, such as collocations, rather than isolated words in rule-governed sequences. Many corpus linguists (Kennedy 1998), Altenberg and Eeg-Olofsson (1990), Manning and Schütze (2002) took up Sinclair's view to argue that the use of high frequencies of preconstructed segments give new relevance to the role of memory in language learning and production. So far as most of language use is people reusing phrases and constructions that they have already heard, the Chomskyan focus on the creativity of language should be questioned. As Manning and Schütze put it, « this serves to de-emphasize the Chomskyan focus on the creativity of language use, and to give more strength to something

like a Hallidayan approach that considers language to be inseparable from its pragmatic and social context. » (Manning & Schütze 2002 :130).

A similar argument has been put forward by historians of linguistics, such as Joseph (2003), to show that Chomsky's conception of infinite linguistic creativity obliges him to reject any ' collocational ' model while for Sinclair and his followers, collocations do not involve a lack of creativity.

Concerning the opposition between intuition and use, the arguments are the following. According to Stubbs (1995), native speakers may be able to give examples of collocation or to judge their likelihood, but they cannot document them, that is give accurate estimates of their frequency. Native speakers are very poor at estimating large numbers. Actually, this argument cannot be said to infirm the recourse to intuition. Frequency counts should be considered part of the analysis, and, as such, they are not directly accessible to native speakers' intuition, just as grammatical categories are not accessible to them either.

8. ***Conclusion***

To conclude, let us resume the issue of corpora. Firth's legacy is very strong regarding the use of texts by Corpus Linguists: language should be studied as whole attested texts, not as isolated text fragments. That is why the Sinclair line rejects the sample method adopted by the Quirk-Leech line so that the choice of whole text against samples is one of the main features which distinguishes the two British Corpus Linguistics trends.

As to the notion of restricted languages, it has been adopted by corpus linguists under the name of « register » first devised by Halliday, McIntosh and Strevens in the 1960s (Halliday et al. 1964).

Still remains the question of generalization from restricted languages (or registers) or any corpora result to general language. This issue had not been tackled by Firth really. In Firth's empiricist view, collocations are abstractions from attested texts, 'abstractions at the syntagmatic level' and restricted languages are « a scientific fiction required by linguistic analysis » and not a general term for any actual institutionalized form of language (Firth [1959] 1968). But here again, he did not give any real clue on the nature of these abstractions and the construction of facts.

Remember that, at the same period, Hockett and later Chomsky had recourse to the notion of projection to face the problem of generalization from a corpus to language. This notion was put forward by Nelson Goodman in order to avoid inductive methods. Conversely, corpus linguists advocate empiricist inductive methods and bottom-up procedures. They tackle the issue of generalisation through the question of statistics and probability within the context of information theory which allowed to think out probability and redundancy of language.

On this point, large-scale corpus methods seem well suited to deal with Halliday's early assumptions on the probabilistic and virtually open-ended

nature of lexis, lexicogrammar, and most notably collocations. But they are likely to make heavy demands on corpus size, computer capacity and statistical sophistication. Then the same questions arise : how big does a corpus have to be and how frequent a structure have to be in order to allow generalizations. This is of course one of the recurrent issues of low-range empiricism where very often only raw lists of word pairs are produced and the necessary stage of fact abstraction is absent. On the other hand, in which extent do these questions matter so far as the objective of corpus-based collocations is the study of language use in order to make grammars and dictionaries for language teaching.

References

Altenberg, Bengt & Mats Eeg-Olofsson. 1990. « Phraseology in Spoken English : Presentation of a Project». *Theory and Practice in Corpus Linguistics* ed. by Jan Aarts & Willem Meijs, 1-26. Amsterdam : Rodopi.

Baker, Mona, Gill Francis & Elena Tognini-Bonelli, eds. 1993. *Text and Technology, in Honour of John Sinclair*, Philadelphia, Amsterdam : John Benjamins.

Bazell, C.E., J.C. Catford, M.A.K Halliday & R.H. Robins, eds. 1966. *In memory of J.R. Firth*. London: Longmans

Beaugrande De, Robert. 1991. « J.R. Firth» *Linguistic Theory : The Discourse of Fundamuntal Works*. London & New York: Longmans.

Butt, David G. 2001. « Firth, Halliday and the development of systemic functional theory » *History of the Language Sciences* ed. by Sylvain Auroux et al. vol. II, 1806-1838. Berlin &New York : Walter de Gruyter.

Chomsky, Noam. 1964. « The Logical Basis of Linguistic Theory» *Proceedings of the 9th International Congress of Linguists 1962* ed. by Horace Lunt, 914-978. The Hague: Mouton.

Church, K. & R. L. Mercer. 1993. « Introduction to the special Issue on Computational Linguistics Using Large Corpora» *Computational Linguistics* 19:1.1-24.

CLRU. 1956. « Cambridge Language Research Group Issue. Abstracts, discussions and three papers in full » *Machine Translation* 3:1. 2-7.

Firth, John Rupert. 1930. *Speech*. London: Benn's Sixpenny Library.

_____ . 1937. *The Tongues of Men.* London: Watts & co.

_____ . 1957. *Papers in Linguistics (1934-1951).* Oxford: Oxford University Press.

_____ . [1935] 1957. « The technique of semantics ». Firth 1957. 7-33.

_____. [1951] 1957. « Modes of meaning ». Firth 1957. 190-215.

_____. [1952] 1968. « Linguistic analysis as a study of meaning ". Palmer F.R. 1968. 12-26.

_____. [1956], 1968, « Linguistic analysis and translation ». Palmer F.R. 1968. 74-83.

_____. [1957a] 1968. « The languages of linguistics ». Palmer F.R. 1968. 27-34.
_____. [1957b] 1968. « Linguistic and translation ». Palmer F.R. 1968. 84-95.
_____. [1957c] 1968. « Descriptive linguistics and the study of English ». Palmer F.R. 1968. 96-113.
_____. [1957d] 1968. « A new approach to ». Palmer F.R. 1968. 114-125.
_____. [1957e] 1968. « Ethnographic analysis and language with reference to Malinowski's views». Palmer F.R. 1968. 137-167.
_____. [1957f] 1968. « A synopsis of linguistic theory 1930-55 ». Palmer F.R. 1968. 168-205.
_____. [1959] 1968. « The treatment of language in general linguistics». Palmer F.R. 1968. 206-209.
Halliday, M.A.K. 1957. « Some Aspects of Systematic Description and Comparison in Grammatical Analysis » *Studies in Linguistic Analysis (Special Volume of the Philological Society)*. 54-67. Oxford : Blackwell.
_____. 1958. « Machine Translation » *Proceedings of the 8th International Congress of Linguists, Oslo, 5 - 9 August 1957* ed. by Eva Sivertsen. 527-533.
_____. 1961. « Categories of the theory of grammar» *Word* 17:3. 241-92.
_____. 1966. « Lexis as a Linguistic Level» *In memory of J.R. Firth*, ed. by C.E. Bazell, J.C. Catford, M.A.K Halliday & R.H. Robins, 148-162. London: Longmans.
Halliday, M.A.K., Angus McIntosh & Peter Strevens. 1964 *: The Linguistic Sciences and Language Teaching*, London: Longmans, Green & co ltd.
Hanks, Patrick. 1996. « Contextual Dependency and Lexical Sets» *International Journal of Corpus Linguistics* 1:1. 75-98.
Henderson, Eugénie J.A. 1987. « J.R. Firth in retrospect : a view from the eighties » *Language Topics : Essays in Honour of Michael Halliday* ed. by Ross Steele & Terry Threadgold, vol I, 57-68. Amsterdam & Philadelphia: John Benjamins.
Hornby, A.S., E.V. Gatenby & H. Wakefield. 1974. *The Advanced Learner's Dictionary of Current English, 3rd edition*. London :Oxford University Press.
Jones, C. & John McH. Sinclair. 1974. « English Lexical Collocations» *Cahiers de Lexicologie* 24. 15-61.
Joseph, John E. 2003. « Rethinking linguistic creativity» *Rethinking Linguistics* ed. by Haylay Davis & Talbot Taylor, 121-150. London & New York : Routledge Curzon.
Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London & New York : Longman.
Langendoen, D. Terence. 1968. *The London School of Linguistics: a study of*

the Linguistic Theories of B.Malinowski and J.R. Firth, *Research Monograph n°46*. Cambridge, Mass.: the MIT Press.

Léon, Jacqueline. 2005. « Claimed and unclaimed sources of Corpus Linguistics ». *The Henry Sweet Society Bulletin* 44. 34-48.

Léon, Jacqueline. « From universal languages to intermediary languages in Machine Translation : the work of the Cambridge Language Research Unit (1955-1970) » *Proceedings of ICHoLS 9, 9th International Conference on the History of Language Sciences,* Sao Paulo (Brasil), 27-30 août 2002 (to be published).

Mackin, Ronald. 1978. « On Collocations: 'Words shall be known by the company they keep'». *In Honour of A. S. Hornby* ed. by Peter Strevens, 149-165. Oxford: Oxford University Press.

Manning, Christopher D. & Hinrich Schütze. 2002. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: The MIT Press.

McIntosh, Angus.1961. « Patterns and ranges ». *Language* 37:3. 325-337.

McIntosh, Angus & M.A.K. Halliday. 1966. *Patterns of Language*. London: Longmans.

Mitchell, T.F. 1975. *Principes of Firthian Linguistics*. London: Longmans.

Monaghan, James. 1979. *The Neo-Firthian Tradition and its Contribution to General Linguistics*. Tübingen : Max Niemeyer Verlag.

Palmer, Harold E. 1933. *Second Interim Report on English Collocations.* Tokyo : Institute for Research in English Teaching.

Palmer, Harold E. 1938. *A grammar of English Words*. London : Longmans Green & co.

Palmer, F.R. ed. 1968. *Selected papers of J.R. Firth (1952-59)*. London: Longmans.

Palmer, F.R. 1994. « Firth and the London School». *Encyclopedy of Language and Linguistics* ed. by R.E. Asher, 1257-1260 Oxford : Pergamon Press.

Robins, R.H. 1961. « John Rupert Firth». *Language* 37:2. 191-200.

Robins, R.H. 1998. « The Contribution of John Rupert Firth to Linguistics in the First fifty years of Lingua» *Texts and Contexts. Selected papers on the History of Linguistics.* 285-310. Münster: Nodus Publications.

Sinclair, John McH. 1966. « Beginning the study of lexis ». *In memory of J.R. Firth*, ed. By C.E. Bazell, J.C. Catford, M.A.K Halliday & R.H. Robins, 410-30. London: Longman.

Sinclair, John McH. 1991. *Corpus, concordance, collocation*. Oxford : Oxford University Press.

Sinclair, John McH., S. Jones & R. Daley, 2004. *The OSTI Report (1970)* ed. by Ramesh Krishnamurthy, New York & London : Continuum.

Stubbs, Michael. 1992. « Institutional Linguistics : Language and Institutions, Linguistics and Sociology». *Thirty Years of Linguistic Evolution* ed. by M. Pütz, 189-211. Amsterdam & Philadelphia: John Benjamins.

Stubbs, Michael. 1993. « British Traditions in Text Analysis – From Firth to Sinclair» *Text and Technology. In Honour of John Sinclair,* ed. by Mona

Baker, Gill Francis, Elena Tognini-Bonelli, 1-36. Amsterdam & Philadelphia: John Benjamins.

Stubbs, Michael. 1995. « Collocations and semantic profiles : On the cause of the trouble with the Quantitative studies» *Functions of Language* 2:1.1-33.

Svartvik, Jan, ed. 1992. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 August 1991*. Berlin & New York: Mouton de Gruyter.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam & Philadelphia: John Benjamins.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. New York : Basil Blackwell.