Palmer, Firth and Internet: Drawing together collocational threads

Geoffrey Williams, Chrystel Millon

Palmer, Firth and Internet: Drawing together collocational threads

Geoffrey Williams and Chrystel Millon
HCTI-LiCoRN
Université de Bretagne-Sud
[williams|millon]@univ-ubs.fr

1 INTRODUCTION

In corpus linguistics, the contribution of Palmer to collocation studies is often overlooked. However, Palmer's Second Report on English Collocations published in Japan in 1933 has been the inspiration for many major threads in phraseology, even if few have actually had access to the work itself. When Firth started looking into collocation, he almost certainly had never heard of Palmer. Thus were the parallel worlds of overseas ELT and British academia at a time before Internet made data exchange and cross disciplinary exchange easy. The consequence is that traditional phraseology, and much pre-corpus lexicography, and corpus linguistics developed on parallel lines. Those lines were effectively drawn together in the COBUILD initiative, although the Palmer connection remained largely forgotten. Part of the cause lies in the fact that these two approaches are based on radically different visions of language. Phraseologists and lexicographers seek to tame language so as to list and classify for inclusion in published works. This requires an essentially static vision of collocation where phraseological units are treated as if created ex nihilo and are simply found and classified on purely linguistic grounds as to what may and what may not be termed as a collocation. The NeoFirthian approach developed by John Sinclair within the context of corpus linguistics is very different in that it places collocation at the heart of language as a essentially dynamic process in which meanings are created and exploited within textual contexts. This requires a much wider vision of collocation rather than simply reducing it to a series of part of speech groupings, with occasionally a smattering of pragmatic or linguistic considerations. The advantage of corpus linguistics is that it allows an analysis of dynamic collocation whilst providing the material for more reductive phraseological or computational exploitation of the data.

This paper intends to look at three issues; the development along parallel lines of phraseological and corpus linguist collocation, reinstating the place of Palmer whilst underlying the centrality of collocation as a language phenomenon, an overview of criteria for restricted collocation and finally how the threads can be drawn together in the extraction and analysis of collocational data from Internet.

2 Collocation as a parliament of words

Words tend to flock together. As a single item, a word is simply a string of characters, in association with other words, it gains meaning. This fact is the basis of all phraseological analysis and underlies the contextualist theory developed by Firth and enshrined in corpus linguistics and especially in collocation studies. Nevertheless, we are still told that words have meanings, and that a dictionary, in which words appear as single headwords, is the arbiter of correct meaning.

Well before the term collocation had come to get its current usage, the great lexicographer Samuel Johnson had noted the importance of linking words to a larger phraseological unit. The arbitrary nature of collocation, rooted as it is in use within a given community, he put forward as self-evident, hence:

> The syntax of this language is too inconstant to be reduced to rules, and can only be learned by the distinct consideration of particular words as they are used by the best authors. Thus, we say according to the present modes of speech, the soldier died of his wounds, and the sailor perished with hunger; and every man acquainted with our language would be offended with a change of these particles, which yet seem originally assigned by chance, there be no reason to be drawn from language why a man may not, with equal propriety, be said to die with a wound or perish of hunger. (Johnson 1747|2008) : 25)

Self evident to Johnson, but still not entirely taken up in more traditional circles. For a first attempt at classifying collocations, or more correctly in this case, what Moon (1998) has termed Fixed Idioms and Expressions (FEI) came only with the work of Palmer from the nineteen twenties onwards.

Based in pre-war Japan, Palmer was concerned with the problem of developing a basic vocabulary for teaching purposes. Linked to this was a project aiming at located and classifying repeated lexical units, what we would probably now call clusters. This was an immense task of locating and listing these items into what was an enormous ragbag of phraseological units. The next step was an attempt at classifying these units into different phraseological groups, amongst which was collocation, defined as being:

> a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts. (Palmer 1933: title page)

The classification systems was largely syntactic with categories as heterosemes, verb + object, verb + preposition, absence of article, “coined” collocations, collocations written without a break, construction patterns. This was a truly ground breaking piece of research, but although republished since in Japan, the most recent reprint being in 1966, few have actually

been able to consult it. This is the main reason why Firth is often designated as the father of collocation, as his writings continued to be discussed.

Whilst Palmer's report remains little read nowadays, its influence has been great. Hornby's groundbreaking 1942 *Advanced Learner's Dictionary of Current English*, the ancestor of the *Oxford Advanced Learner's Dictionary*, was aimed at both encoding and decoding, and thus made great use of collocational and phraseological information. This was such a break-through in dictionary making that traditional lexicographers refused it a place in the Oxford dictionary catalogue (Béjoint 2010: 165). The ideas have now become mainstream.

Not only did Palmer influence lexicography, but also phraseological research throughout Europe, thereby creating a phraseological tradition of collocation studies concerned by fixed or semi-fixed units. This tradition was developed by Haussman (1985) with his notions of free, trivial and bound collocations, as well as the accompanying terminology of base and collocate. It also bore fruit in numerous dictionaries of collocations such as the seminal Benson, Benson and Ilson's The *BBI Dictionary of English Word Combinations* (1986). It also inspired the great Russian traditions of phraseological studies, amongst which is Mel'čuk's Théorie Sens-Texte with its lexical functions. In most cases, the classifications are syntactic, the difference being that of Mel'čuk where collocations are based on semantic relations.

What typifies most this phraseological tradition is that it treats collocation as a product to be studied and listed in reference works. The traditions that stem from Firth see collocation more as a process leading to a more dynamic outlook.

Developed also in the 1930's, but in an academic setting rather than a teaching one, Firth's theory of collocations turns around the company words keep. The major influence here was that of Malinowski with his twin notions of context of culture and context of situation which based collocation firmly within usage with society and as a dynamic phenomenon. Most of Firth's discussion of the issue is found in papers published between 1934 and 1951 brought together in 1957 (Firth 1957). What brought this outlook on collocation to the forefront was its being taken up by John Sinclair and developed as a tool for lexical analysis (1966), then in the early corpus linguistics of the OSTI report (1970|2004) and then through the building of the COBUILD corpus with its associated dictionary. The COBUILD dictionary essentially brought together the process and product approaches in a new dynamic tool for language learning.

This corpus linguistic approach to collocation is best summed up in Sinclair's seminal Corpus Concordance Collocation of 1991. A disarmingly easy book to read, and still by far the best way into corpus linguistics. It is within this context that Sinclair developed the notion of the idiom principle, which holds such a central place in corpus linguistics.

The company words keep, as seen through the corpus, requires a much wider view of collocation than that imposed by syntactic classification criteria. This outlook has brought about the theory of semantic prosody, which demonstrates underlying connotational features of language use (Louw 1993), thematic structuring of language in collocational networks (Williams 1998), pattern grammars (Hunston & Francis 2000) and studies in language production through lexical priming (Hoey 2005). All of these seen collocation as a dynamic process in which the lexical environment plays the key role.

Thus two traditions have developed, a phraseological and a contextualist one. The COBUILD dictionary has shown that the parallel lines can converge. The scientific verb dictionary project (Williams and Millon 2010) is trying to take this even further by harnessing the combining power of words in collocational networks as a lexical selection and dictionary navigation tool. The following sections are not part of this project, but relate to previous work developed in a doctoral thesis (Millon, 2011) that endeavours to use web extracted collocations in a way that will bring together the two outlooks, process and product, on collocation and also feed into the dictionary project and also the web-based comparative corpus and terminology extraction tools being developed in the METRICC research programme (http://www.metricc.com/).

3 Automatic extraction of a semantic collocations database from the Web

The goal of the doctoral thesis (Millon, 2011) was to propose a protocol for building automatically a semantic database of French collocations for a given target noun. An endogenous method has been used for discriminating the different meanings of the target word that occur in the corpus. The 'corpus' is in fact a concordance of the target word gathered from the snippets returned by the commercial search engine *Yahoo Search!* in which the target word appears. The result of the meaning discrimination is a set of semantic classes, which are filled with significant ‘textual’ collocates of the target word. The semantic classes then serve for the automatic semantic categorization of ‘syntactic’ collocations that have been extracted from the corpus. The methodology of Millon (2011) will be illustrated with the French noun VAISSEAU (vessel), for which we give in appendix the senses found in the French on-line dictionary the *Trésor de la Langue Française Informatisé* (TLFI).

3.1 Corpus constitution

Two first Web corpora are constructed, one per each form the noun VAISSEAU. The Web is searched via the commercial search engine *Yahoo Search France!* using a simple query : the query vaisseaux for getting occurrences of the plural form *vaisseaux*, and the query vaisseau for getting occurrences of the singular form *vaisseau*. I should be noted that the API Yahoo is used, not the graphic interface of the search engine. The search engine returns a maximum of 1000 results per query. Below is reported an example of result returned by *Yahoo Search France!*.

> Vaisseau.
>
> Les mots du vivant. ... Vaisseau (anatomie, Planche anatomique : angéiologie). - Les **vaisseaux** (vasa) en anatomie sont les canaux dans lesquels circulent les liquides de l ...
> www.**cosmovisions.com**/vaisseau.htm

A Yahoo result is composed by three data : the name of the Web, a snippet of the Web page and the URL of the Web page. In each snippet, there is at least one occurrence of the word written in the Web query. Each snippet is split according the following three characters "...". For example, the snippet of Yahoo result given below contains two occurrences of these three characters. Consequently, the snippet is split in two textual elements, that are "Les mots du vivant." and "Vaisseau (anatomie, Planche anatomique : angéiologie). - Les vaisseaux (vasa) en anatomie sont les canaux dans lesquels circulent les liquides de l". The textual elements linked to the query vaisseaux are gathered, keeping out those in which *vaisseaux* does not occur. This is then the corpus of *vaisseaux*, in which each textual element constitute the textual context of this target word-form. The corpus of the singular form *vaisseau* is also build in this way.

Both corpora are small in size, so in order to get more occurrences of the target word-form, notably more occurrences of meaning(s) that have a low representation in the corpus, the Web is again searched but this time using binary queries composed by the target word-form and a significant collocate of it : vaisseaux +sanguins for instance. With this second search of the Web, two other corpora are built. The collocates of *vaisseaux* implied in the binary queries are extracted from the corpus of *vaisseaux*. A collocate is defined within textual and frequency dimensions. No fixed window is used for the extraction, because collocates are extracted simply within contexts ; the window is the context. Collocates that are listed in a stop-list or have less than three characters are kept out. A frequency filter is lastly applied to keep only ones that occur at least 8 times in the first corpus of *vaisseaux*.

From the list of the collocates, the binary queries for building the second corpus of *vaisseaux* are automatically generated, and launched to *Yahoo Search!*. The construction of the

second corpus follows the same process that the one described above. Duplications are eliminated, and corpus is cleaned in order to eliminate contexts that are specific to the structure of the Web site, such as tables contents, and those that contains less than five orthographic words. The second corpus of the singular form *vaisseau* is also build in this way.

The Web is exploited then as a huge textual source, that is free. As the size of the web pages database of *Yahoo Search!* is immense, it is more likely to find different meanings of a given word. For instance, in the short corpus of *vaisseaux*, five meanings occurs, namely the anatomic sense ('vessel'), ship sense ('ship'), spaceship sense ('spaceship'), botanical sense ('vessel') and architectural sense ('nave').

We will use the term 'short corpus' to refer to both first corpora, that are ones getting with a simple query, and the term 'large corpus' to refer to both second corpora constructed via binary queries. The large corpus of *vaisseau* contains 21018 occurrences of *vaisseau*, and the one of vaisseaux 20363 occurrences of *vaisseaux*.

The word sense discrimination processing, that will be described in the next subsection, is based on the word-form level. Indeed, different forms of the same word could have different meanings. Consequently, according to the French noun taken in illustration, two set of semantic classes will be discovered, one per each form of VAISSEAU ; each set of classes is get from the appropriate large corpus.

3.2 Word Sense Discrimination

In order to discriminate the different senses of the given word, we look at its lexical environment in the corpus. The semantic discrimination process is indeed carried out within a lexical semantic network in which the semantic proximity between collocates of the target word are given. The process consists in three steps. Firstly, a set of collocates that will serve to semantic discrimination will be extracted from the short corpus of the given target word-form. Secondly, the semantic proximity between the collocates of the set of the collocates will be calculated. Thirdly, the data from the preceding step will serve to build the lexical semantic network of the given target word-form, from which the semantic classes will emerge.

3.2.1 Step 1 : building the set of collocates

Starting from the plural form *vaisseaux*, the set of collocates is automatically built from two source corpora. The first one is the short corpus of *vaisseaux*, in which we take all the collocates that occur more than 7 times (these are the collocates that will serve to construct the large corpus of *vaisseaux*). The second one is the large corpus of *vaisseaux*. Within the second

source, the collocates that will be automatically chosen are extracted from the top 15 collocates -in term of frequency- of each collocate selected from the first source corpora. I should be noted that all the collocates in the set are collocates of the target word-form which is under study, since our corpus is in fact a concordance of it.

The division into meaning potentials in a given corpus is thus based exclusively from the lexical environment of the collocates reported in the set. As the semantic classes discovered stand for different meanings of the given target word-form, it was decided to base the process on the semantic similarity between words. Consequently, we have first to calculate the semantic proximity between words (step 2), and then proceed to the gathering of words into classes (step 3).

3.2.2 Step 2 : semantic proximity between words

For this step, the procedure is based on the second-order occurrences. We deal again with the 15 first collocates of a collocate given in the set of collocates. A collocate is defined here in the contextualist framework, since it is extracted on the word-form level from the large corpus of the target word-form under analysis (*vaisseau* or *vaisseaux*). In building the top 15 collocates list, collocates that appear in a stop list are kept out.

In the set of collocates built for the target word-form *vaisseaux*, there are for example the collocates *cellules* and *tissus*. In order to calculate the semantic proximity between these two words, we look at the top 15 collocates of *cellules (cells)* and those of *tissus* (tissues) from the long corpus of *vaisseaux*. In the top lists, words are sorted according to the frequency of the ternary collocations involved, that is for example *vaisseaux* (the target word-form) + *cellules* (a collocate in the set of collocates) + *sanguins (blood)* (a collocate in the top 15 list of *cellules*). The ternary collocations that have less than 3 occurrences are kept out. This means that the top 15 collocates are in fact the first 15 collocates of the binary collocation *vaisseaux* + *cellules*.

Then, we weight the words in the top lists by giving a simple numeric value on a scale of 1 to 15. The weight is assigned according to rank : the word in rank 1 has the value 15, the one in rank 2 has the value 14, etc. The weighted top lists of *cellules* and *tissus* are given below in Table 1.The weighting will serve to calculate the semantic proximity between two words.

vaisseaux		
cellules	***tissus***	**Weight**
sanguins	sanguins	**15**
endothéliales	peau	**14**
paroi	sang	**13**
lymphatique	nerfs	**12**
parois	capillaires	**11**
souches	petits	**10**
capillaires	*cellules*	**9**
musculaires	coeur	**8**
tissus	organes	**7**
formation	muscles	**6**
tapissent	nouveaux	**5**
sang	mous	**4**
interne	parois	**3**
nouveaux	artères	**2**

Table 1. Top 15 collocates list of *cellules* and *tissus* extracted from the long corpus of *vaisseaux*

The score of semantic proximity depends on the number of shared words in the top-15 list and on the weight of them. For instance, the semantic proximity score between *cellules* and *tissus* is 181. The principle is to add the weight of the shared words of the two top 15 collocates lists compared. Yet, if among the top 15 collocates list appears either *cellules* or *tissus*, then we change its actual weight in giving the value 20. In this way, we give a stronger weight to the two collocates of the lexical pair. For example, in the collocational top-list of the word *tissus*, the word *cellules* is in it, with the weight 9. So, for calculating the semantic proximity score, we will not take the value 9, but the value 20. As the word *tissus* is in the top 15 collocates list of *cellules*, its actual value of 7 has been changed to 20. For the calculation of the semantic proximity score, we add in both top 15 collocates lists, the word under consideration. So, *cellules* is added to its own list, and *tissus* in its own list, with a weight of 20. In Table 2, the shared word are written in bold.

cellules weighted top-list	
sanguins	**15**
endothéliales	14
paroi	13
lymphatique	**12**
parois	**11**
souches	10
capillaires	**9**
musculaires	8
tissus	***~~7~~ => 20***
formation	6
tapissent	5
sang	**4**
interne	3
nouveaux	**2**
cancéreuses	1
cellules	20

tissus weighted top-list	
sanguins	**15**
peau	14
sang	**13**
nerfs	12
capillaires	**11**
petits	10
cellules	***~~9~~ => 20***
coeur	8
organes	7
muscles	6
nouveaux	**5**
mous	4
parois	**3**
artères	2
lymphatiques	**1**
tissus	**20**

Table 2. Illustration of the calculation of the semantic proximity score

At the end of step 2, each lexical pair, generated with the collocates of the set of collocates, has a semantic proximity score.

The theoretical maximal semantic proximity score is 318, and the minimal one is 0. If a lexical pair have a score 0, that means that the two words of the pair doesn't share any word in their top-15 collocates list. In the contrary, if a lexical pair has the maximal score, that means that there is a total overlap between the two top-15 collocates lists involved. We consider that a score of 68 is sufficient to judge a semantic proximity, even if this value is very far from the maximal possible score. The threshold was chosen according to two reasons. The first reason, is that a lexical pair that shares 8 words in their top-15 collocates lists is judged semantically similar, and with at least 8 shared words, the minimal semantic proximity score is 72 (given the 8 words in the tail of the list, that is words in rank 8 to 15), and the maximal is 184 (given the 8 words in the head of the list, that is words in rank 1 to 8). The second reason matches a lexical pair that shares less than 8 words. With a threshold of 68, we want to judge pertinent lexical pairs that share very little words, that is to say two or three words but which are in the head of the top-15 collocates list. For example, if the words shared are the three words in the head of the two lists compared, so the minimal possible semantic score is 84. In the situation where only two words are shared, we want to judge pertinent the lexical pair if this overlap imply words in the first three ranks (rank 1, 2 and 3). But if this is only the two head words that are shared, and these words are not the two ones imply in the lexical pair, then the minimal possible

semantic score is 58. Indeed, in the situation of only two words shared, we want to judge pertinent only lexical pairs in which either the two words in the lexical pairs (so the semantic score is 80, independently to the ranks of words) or only one of these two are among the two words shared in the lists compared. If only one of these two words in the lexical pair are among the two words shared in the lists compared, then there is restriction : given the lexical pair A – B, in the top-15 collocates lists of A, the words must be in ranks 1, 2, or 3. Given lexical pair *A* – *B*, in which the two words *A* and *C* are shared in the top-lists. To reach the threshold 68, as the change of weight to 20 is realized only for the A in the top-list of *B*, the word *C* must be at rank 2 in the top 15 collocates list of *A* : 20 (weight of the word *A* in the list of *A*) + 14 (weight of the rank 2). In the top-15 collocates list of B, the word C must be in rank 2 and the word A can be in any rank : 14 (weight of the rank 2) + 20 (changed weight of the word A).

The threshold 68 is consequently established considering in fact the sum 34 in this borderline situation described above ; the sum of 34 getting from the list A and the sum 34 getting from the list B. This explains so that the threshold 68 is quite far from the maximal possible semantic proximity score.

Thus, in the following step, we will not consider that the discovery of semantic class for lexical pairs that have a score below 68. The word sense discrimination is realized in the semantic lexical network building with the collocates that are in the set of collocates and showing the semantic proximity between them. The semantic classes will be discovered within the semantic lexical network by clustering the words. The processing constitute the next step.

3.2.3 Step 3 : clustering

For step 3, we elaborate a process inspired by the *HyperLex* algorithm, created by Jean Véronis (2003, 2004). First, the semantic network is built in which nodes are the words in the set of collocates, and two nodes are linked if they have between them a semantic proximity score superior or equal to 68. After building the network, the nodes that have no links at all are eliminated ; in the case of the building network of *vaisseaux*, no node were eliminated. We give in Table 3 the data of the network of *vaisseaux* and those of *vaisseau*.

Target word-form	***Nb nodes***
vaisseaux	142
vaisseau	161

Table 3

The semantic classes are discovered inside the lexical network. Each node will be analysed by itself. So, we have to decide in which order the nodes will be analysed. After comparing the

results obtained from four distinct orders (Millon, 2011), we identified the following one as the best : the order of the analysis of the nodes is according to the number of appearances in the top 15 collocates list of all the collocates in the set of collocates. We add then their numeric value (the weight in a scale on 1 to 15) that have been given in the top-15 collocates lists during the step 2. According to the sorting, in the case of the semantic discrimination of the word-form *vaisseaux*, the first node to be analysed is *sanguins*.

If the local network of the node *sanguins* fills the set 1 of properties or the set 2 of properties, that we describe below, then this node will be considered as a "founding node" of a new semantic class :

Set 1 of properties of a founding node :

- Proportion of its direct neighbours that have been removed from the network : < 0,7
- Number of direct neighbours present in the actual state of the network : >= 6.

Set 2 of properties of a founding node:

- Proportion of its direct neighbours that are eliminated from the lexical network : < 0,7
- Number of direct neighbours present in the actual state of the network : < 6.
- Agglomeration coefficient between its direct neighbours that are present in

The calculation of the proportion of the direct neighbours of a given node that have been removed from the network is the following :

number of its direct neighbours that are no longer present in the actual state of the network / number of its original direct neighbours

Agglomeration coefficient (one of the properties of the set 2) measures the connection strength between a specific set of nodes in the network. In our case, the set of nodes is all the direct neighbours of the potential founding node under analysis, present in the current state of the network. Agglomeration coefficient is calculated in the following way : given the potential founding node A, and a set of 25 neighbours nodes of it. First, the 'maximal possible number of links of each neighbour' is calculated : *number of neighbours of A – 1*, that is to say 24 in our example. Second, the 'maximal possible number of links between all the neighbours of A' is calculated : *(number of neighbours of A * maximal possible number of links of each neighbour) / 2*. Hence, in our illustration, the result is 300, because (25*24)/2 = 300. Third, we get the real number of links between all the neighbours of A. Given 236 the number of real links. Last, the agglomeration coefficient is calculated in the way : *number of real links / maximal possible*

number of links between all the neighbours of A. In our illustration, the agglomeration coefficient is 0,79 (236/300).

The agglomeration coefficient varies between 0 and 1. A coefficient of 1 means that all nodes are linked between them. In the contrary, a coefficient of 0 means that no node is linked to any other in the considered set of neighbours. This calculation assures a semantic coherence of the semantic class will be formed. Indeed, the future semantic class will have a certain semantic proximity with the founding node, but words that are in the class will not have a semantic proximity between them. If the potential new semantic class has a very low semantic coherence in the case for example where the agglomeration coefficient is 0, the potential founding node is finally not a good candidate for creating its own semantic class. Agglomeration coefficient gives a supplementary surety to the semantic coherence of the class. Agglomeration coefficient is a tool to theoretically avoid the formation of an heterogeneous semantic class. However, we decide to calculate the agglomeration coefficient only for potential founding node that have less than 6 direct neighbours in the current stat of the network.

If the potential founding node matches set 1 or set 2 of properties established for identifying a founding node, then this node is a founding node. Hence, a new semantic class is formed, filled with the founding node and all its direct neighbours present in the current state of the network. All the nodes integrated in the semantic class are then removed from the network. We use the term 'founding node' for the node that leads to the creation of a new semantic class.

If the potential founding node doesn't match either set 1, or set 2, no new class is created. Then, if this rejected node because of the proportion >= 0,7 of its direct neighbours that were removed from the lexical network, it will potentially integrate one of the semantic classes already discovered. The assumption is that the node rejected could be semantically close to one the semantic classes. Indeed, even if it has a semantic proximity score lower than 68 with the founding node of a given semantic classes already discovered, it could enter in it, because having semantic proximity superior or equal to 68 with other nodes in the class. Consequently, some semantic classes could develop with some nodes that are indirect neighbours of the founding node. We speak of an 'indirect integration' if a word appears in a semantic class, but is not a direct neighbour of the founding node.

An 'indirect integration' processing is realized either when a potential founding node does not match set 1 and set 2 of properties of a founding node, or as soon as a new semantic class is formed –the process will then analyse, according to the order of the sorted list, each remaining node in the network. For the indirect integration in a semantic class already discovered, we look at the links of the node. We establish to different sets of properties according to the nature of

the node : either it is a potential founding node that have been not identify as a founding node, or it is a remaining node.

The set of properties for an indirect integration of a potential founding node is given below :

> Indirect integration (potential founding node rejected as founding node)
> Proportion of its direct neighbours that are eliminated from the lexical network : >= 0,7
> Number of its original direct neighbours >= 4
> Among its direct neighbours that have already been removed from the network and that have been placed in semantic classes, only and only one semantic class must be involved.

The set of properties for an indirect integration of a remaining node in the network:

> Indirect integration (remaining node in the network)
> Proportion of its direct original neighbours that have been already placed in a given semantic class discovered : >= 0,7
> Number of its direct original neighbours : >= 4

So, after each analysis of a given node as a potential founding node the state of the network change. If the potential founding node under analysis is identified as a founding node, this node and all its direct neighbours in the actual state of the network are removed. If the potential founding node is not identified as such, then this node is also removed from the network ; the node is then either integrated in an indirect way in a semantic class already discovered, or is simply kept out. Concerning the analysis of the remaining nodes, this takes place directly after the discovery of a new semantic class, if a given remaining node is indirectly integrated to a semantic class, then it is removed from the network.

As the node *sanguins* match the criteria of set 1 of properties, it becomes a founding node. The node *sanguins* has actually 89 direct neighbours (see below, alphabetic sort). So, the semantic class is first formed with these 89 nodes plus the node *sanguins*.

> aorte ; artères ; artérioles ; atteinte ; calibre ; capillaires ; cardiaques ; cellules ; cerveau ; chaleur ; cholestérol ; circulation ; circulatoire ; circule ; coeur ; combat ; constitué ; contient ; contraction ; coronaires ; corps ; couperose ; croissance ; cutanés ; dilatation ; dilate ; dilatés ; due ; développement ; endothéliales ; fibres ; fins ; foie ; formation ; forment ; ganglions ; grand ; grands ; gros ; inflammation ; interne ; long ; lymphatique ; lymphatiques ; lymphe ; lésions ; maladies ; mous ; muscles ; musculaires ; nerfs ; nerveux ; niveau ; nom ; nouveaux ; organes ; paroi ; parois ; peau ; permet ; permettent ; petit ; petite ; petits ; poumons ; provoque ; reins ; rouge ; rougeurs ; réseau ; résistance ; sang ; sanguine ; santé ; superficiels ; surface ; système ; taille ; tendons ; tissu ; tissus ; travers ; tumeur ; types ; tête ; valves ; vasculaire ; veines ; yeux

After the formation of a new semantic class, the indirect integration processing of the remaining nodes in the network is launched. In our example, the remaining node *institut* has been integrated in the semantic class formed by the founding node *sanguins*, because it presents the following properties :

- Proportion of its direct original neighbours that have been already placed in the semantic class 'sanguins' : 1 (>= 0,7)
- Number of its direct original neighbours : 4 (>= 4) : *coeur nom sang santé*.

Among the remaining nodes, *spatiaux* is the next node to be analysed as a potential founding node. As its properties also match the set 1, it is a founding node, leading to the formation of a new semantic class. Among its 35 original direct neighbours, 5 are not present in the current state of the network, namely *petite*, *grand*, *nom*, *combat* and *système*, because they were integrated in the semantic class founded by *sanguins*. Consequently, this second new semantic class is first composed of 31 nodes : *spatiaux* plus its 30 remaining nodes in the network. After the indirect integration processing, none remained to be integrated into the semantic class 'sanguins', but three nodes, that are *races*, *images* and *histoire*, were in the class 'spatiaux'. For naming the semantic classes, we use the founding node. So, just 17 nodes are now present in the network, and within this actual state, the node *frégates* is the next potential founding node to be analysed. As the properties of *frégates* match the set 1 (see below), it is a founding node, leading to the creation of a third semantic class. Its properties are the following :

- Proportion of its direct neighbours that are eliminated from the lexical network : 0,48 (11/23)
- Number of direct neighbours present in the actual state of the network : 12

The 11 nodes that were no longer present in the current state of the network were integrated in the semantic class '*sanguins*' (*types*, *combat*, *petite*, *grands*) and in that founded by the founding node *spatiaux* (*armada*, *guerre*, *canons*, *détruire*, *construire*, *type*, *flotte*). The semantic class founded by *frégates* is first composed of its 12 remaining neighbours, that are then eliminated from the network, namely *armés, bâtiments, commerce, croiseurs, français, ligne, marine, navires, port, pirates, roi* and *transport.* The indirect integration processing of the remaining nodes lead to no integration within any of the three semantic classes discovered. Now, the current state of the network contains only 4 nodes : *fantômes*, *mer*, *soyouz* and *brûlés*. Among these four nodes, the next one to be analysed as a potential founding node is *fantômes*. It is not a founding node, because 0,8 of these original direct neighbours are no longer present in the network. It is then not indirectly integrated into one of the semantic classes 'sanguins', 'spatiaux' and 'frégates' because its direct neighbours that appear in semantic classes

discovered are spread in the classes ‘spatiaux’ and ‘frégates’. The node *fantômes* is then simply removed from the network. For the same reasons, the nodes *mer*, *soyouz* and *brûlés* not lead to the creation of new classes, and were not integrated into one of the semantic classes.

Nevertheless, the nodes *soyouz* and *brûlés* formed two new semantic classes. Indeed, once the whole content of the network has been analyzed, a last processing is applied, aiming to potentially form new semantic class(es) with the nodes that have been kept out, that is to say nodes that were rejected as a founding node and had not been indirectly integrated to one of the semantic classes. In the case of *vaisseaux*, the nodes *fantômes*, *mer*, *soyouz* and *brûlés* have been kept out. Conditions for the creation of such supplementary semantic class are the following :the node must have only one original direct neighbor, and its original neighbour has to appear in one the semantic classes discovered. If the two conditions are filled, the node is a founding node leading to the formation of a new semantic class composed by it and its original neighbour. In doing this, we hope to detect meanings that have very low representation in the corpus.

These semantic classes will serve to the semantic classification processing of the ‘syntactic collocations’. Indeed, the collocations will be associated to one or several of the discovered semantic classes. In the following subsection, we describe the syntactic collocation extraction.

3.3 Collocations extraction

Collocations are searched here in the large corpus gathering contexts of *vaisseaux* and *vaisseau* in accordance to grammatical relations, in a part-of-speech tagged version of corpus. The grammatical relations are pre-defined. The set of grammatical relations are the following : NOM + ADJ (attributive and predicative), NOM + PREP(DE) + NOM, NOM + PREP (≠DE) + NOM, NOM + VERBE, VERBE + NOM, NOM + VERBE_PARTICIPEPASSÉ, VERBE + PREP(DE) + NOM, VERBE + PREP(≠DE) + NOM, VERBE + PREP(DE) + VERBE_INFINITIF, VERBE + PREP(≠DE) + VERBE_INFINITIF, VERBE + ADJ. The grammatical relations between words are identified within grammatical schemes encoded in Perl scripts. Syntactic collocations extracted can involve, or not, the noun VAISSEAU. Syntactic collocations are extracted using the word-form level, but there are finally gather according the lemma level. The word-form level allows to save the morpho-syntactic contexts of each collocation : distribution of articles of the noun, absence of article, and the modes of verbs (infinitive, gerund, etc.).

Table 4 below shows the ten first collocations in term of frequency for the following grammatical patterns : NOM + PREP(DE) + NOM, NOM + PREP (≠DE) + NOM, NOM + VERBE, VERBE + NOM.

VERBE - NOM	NOM - VERBE	NOM - PREP - NOM	NOM - PREPde - NOM
construire & vaisseau	vaisseau & être	sang & dans vaisseau	capitaine & de vaisseau
avoir & vaisseau	vaisseau & avoir	vaisseau & en orbite	lieutenant & de vaisseau
détruire & vaisseau	vaisseau & transporter	vaisseau & dans espace	vaisseau & de guerre
voir & vaisseau	vaisseau & arriver	vaisseau & en forme	paroi & de vaisseau
piloter & vaisseau	vaisseau & étayer|être	vaisseau & à niveau	dilatation & de vaisseau
entrer & vaisseau	vaisseau & calibrer	vaisseau & sur planète	enseigne & de vaisseau
lancer & vaisseau	vaisseau & entrer	espace & à_bord_de vaisseau	vaisseau & de ligne
envoyer & vaisseau	vaisseau & voler	orbite & autour_de terre	flotte & de vaisseau
dilater & vaisseau	vaisseau & décoller	vaisseau & à travers	vaisseau & de flotte
diriger & vaisseau	vaisseau & aller	terre & à_bord_de vaisseau	intérieur & de vaisseau

Table 4. Ten first collocations (according to the raw frequency) within a grammatical classification – collocations extracted from the both large corpus vaisseau and vaisseaux.

The ten first collocations classified following grammatical patterns reveal one major meaning, that is the 'spaceship' sense of the French noun VAISSEAU. As for the word frequency distribution in a corpus that follows the Zip's law, meanings of words have as well a skewed distribution. But the fact that meanings are very little represented in the corpus does not mean that they are not pertinent for speakers. The semantic classification of collocations that Millon (2011) proposes makes it possible to reduce the difficulty related to the dispersion of meanings, which is particularly useful for example in a context of a lexicographical work.

The next subsection describes the automatic semantic categorization processing according to the set of semantic classes that have been discovered during the word meaning discrimination processing.

3.4 Automatic semantic categorization of the collocations

The aim of the automatic semantic categorization collocations processing is to attach a given collocation to at least one of the semantic classes discovered. As each semantic class is composed by a set of words, the present processing depends on the degree of overlap between the collocates of the given collocation and the content of the semantic classes. A collocate is defined here in the contextualist framework, since it is extracted from all the contexts of *vaisseau* and *vaisseaux*, according to the word-form level. The analysis corpus is then the large one gathered from the two word-forms of the noun VAISSEAU. While no statistical dimension is considered, a lexical one is applied, keeping out so the collocates that are in the stop list –the same used for the word meaning discrimination.

As two distinct sets of semantic classes are obtained, one for each word-form of the French noun VAISSEAU, a given syntactic collocation will be analysed twice, one according to the set

of semantic classes discovered for the plural form *vaisseaux*, and one according to the other set of semantic classes.

For each semantic class discovered, two proportions are calculated in order to determine if the given collocation will or will not attach to it. Both proportions are described below :

- **Proportion 1** : *number of the collocates of the given collocation that appear in the given semantic class / number of the collocates of the given collocation that appear within all the semantic classes*.
- **Proportion 2** : *number of the collocates of the given collocation that appear in the given semantic class / number of the words of this semantic class*.

Given the two proportions, if **Proportion 1 >= 0,5** and **Proportion 2 >= 0,1**, then the given collocation is classified in the semantic class involved. To illustrate the process, we take the collocation *DILATATION & DE VAISSEAU*, that occurs 367 times in the corpus. According to its lexical distribution, 54 collocates appear in the semantic classes which have been discovered for the plural form *vaisseaux*. This partial overlapping follows the following distribution : one collocate appears in the semantic 'frégates' class, one in the 'spatiaux' class and 52 in the 'sanguins' class. Proportion 1 and Proportion 2 of each of these three semantic classes are showed in Table 5 below. The collocation *DILATATION & DE VAISSEAU* is then attached to the 'sanguins' class.

Class name	Size of the class	P1	P2	Collocats impliqués de la collocation
frégates	13	0.02	0.08	transport ;
spatiaux	34	0.02	0.03	jeu ;
sanguins	91	0.96	0.57	permettent ; surface ; superficiels ; cellules ; artérioles ; cerveau ; paroi ; musculaires ; tissus ; dilate ; capillaires ; due ; peau ; réseau ; sanguine ; coronaires ; coeur ; veines ; cholestérol ; chaleur ; muscles ; petits ; rouge ; sang ; couperose ; circulation ; contraction ; nom ; maladies ; artères ; nerveux ; parois ; niveau ; permet ; cutanés ; inflammation ; rougeurs ; provoque ; sanguins ; tête ; calibre ; corps ; cardiaques ; grands ; yeux ; tissu ; lymphatiques ; forment ; dilatation ; atteinte ; système ; fibres ;

Table 5. P1 and P2 scores for the collocation *DILATATION & DE VAISSEAU* according the semantic classes of *vaisseaux*.

Nevertheless, the collocation *DILATATION & DE VAISSEAU* is not classified at all within the semantic classes discovered for the singular form *vaisseau*. According to the content of this latter set of semantic classes, 16 collocates of the given collocation appear : 4 in the semantic class 'sanguin' and 12 in the class 'spatial'. Proportions P1 and P2 calculated for these two SG classes are given in Table 6 below.

Class	Size if the class	P1	P2	Collocats impliqués de la collocation
sanguin	6	0.25	0.67	niveau ; coeur ; sanguin ; sang ;
spatial	138	0.75	0.09	nom ; forme ; prend ; fin ; pose ; retour ; temps ; intérieur ; jeu ; grande ; contrôle ; système ;

Table 6. P1 and P2 scores for the collocation *DILATATION & DE VAISSEAU* according the semantic classes of *vaisseau*.

In the following subsection, we give the results of both the semantic discrimination processing and the semantic syntactic collocations categorization processing.

3.5 Results

3.5.1 Results of the word sense discrimination

The results are shown in Tables 7 and 8. Table 7 gives the set of the semantic classes discovered for the plural form *vaisseaux*, and Table 8 gives the result for the singular form *vaisseau*.

Semantic classes of the word-form vaisseaux	
Founding node	**Content of the class**
sanguins (91)	aorte ; artères ; artérioles ; atteinte ; calibre ; capillaires ; cardiaques ; cellules ; cerveau ; chaleur ; cholestérol ; circulation ; circulatoire ; circule ; coeur ; combat ; constitué ; contient ; contraction ; coronaires ; corps ; couperose ; croissance ; cutanés ; dilatation ; dilate ; dilatés ; due ; développement ; endothéliales ; fibres ; fins ; foie ; formation ; forment ; ganglions ; grand ; grands ; gros ; inflammation ; institut ; interne ; long ; lymphatique ; lymphatiques ; lymphe ; lésions ; maladies ; mous ; muscles ; musculaires ; nerfs ; nerveux ; niveau ; nom ; nouveaux ; organes ; paroi ; parois ; peau ; permet ; permettent ; petit ; petite ; petits ; poumons ; provoque ; reins ; rouge ; rougeurs ; réseau ; résistance ; sang ; sanguine ; sanguins ; santé ; superficiels ; surface ; système ; taille ; tendons ; tissu ; tissus ; travers ; tumeur ; types ; tête ; valves ; vasculaire ; veines ; yeux
spatiaux (34)	armada ; armes ; canons ; combats ; commandes ; construction ; construire ; détruire ; ennemis ; espace ; fiches ; flotte ; guerre ; histoire ; images ; jeu ; jeux ; laser ; nouvelles ; orbite ; personnages ; planète ; planètes ; races ; site ; spatiaux ; star ; terre ; trek ; type ; univers ; véhicules ; wars ; épisodes
frégates (13)	armés ; bâtiments ; commerce ; croiseurs ; français ; frégates ; ligne ; marine ; navires ; pirates ; port ; roi ; transport
soyouz (2)	espace ; soyouz
brûlés (2)	brûlés, petite

Table 7. Semantic classes discovered for vaisseaux

Semantic classes of the word-form vaisseau	
Founding node	**Content of the class**
spatial (138)	alien ; amiral ; américain ; angoisse ; années ; apollo ; appelé ; arrimage ; arrimé ; astronautes ; base ; baïkonour ; bord ; canons ; capable ; capitaine ; cargo ; cassini ; chine ; chinois ; cinéma ; combat ; commandant ; commande ; commandes ; compagnie ; construire ; construit ; contrôle ; cosmodrome ; célèbre ; deuxième ; divin ; détruire ; enterprise ; envoyé ; espace ; extra-terrestre ; extraterrestre ; fantôme ; film ; fin ; forme ; français ; fusée ; galactica ; grand ; grande ; gros ; guerre ; habité ; heures ; histoire ; homme ; hommes ; horizon ; icarus ; image ; immense ; internationale ; intérieur ; iss ; jeu ; jours ; kirk ; lancement ; lancer ; lancé ; lieu ; ligne ; lune ; mer ; mir ; mis ; mise ; mission ; monde ; mère ; navire ; nom ; nommé ; nouvelle ; orbite ; paris ; part ; petit ; petite ; place ; placé ; planète ; porte ; pose ; première ; prend ; prendre ; pris ; progress ; ravitaillement ; rejoindre ; retour ; russe ; réussi ; serenity ; seul ; shenzhou ; site ; siècle ; soviétique ; soyouz ; spatial ; spatiale ; star ; station ; succès ; système ; série ; temps ; terre ; terrestre ; titan ; touriste ; transportant ; type ; uss ; version ; vie ; vient ; ville ; vitesse ; vol ; volant ; vostok ; voyage ; voyager ; écrase ; énorme ; équipage ; évent
lieutenant (14)	chef ; classe ; commandement ; enseigne ; flotte ; française ; frégate ; grade ; lieutenant ; marine ; marins ; navale ; officier ; pierre
sanguin (6)	artère ; coeur ; **caillot** ; niveau ; sang ; sanguin
lego (2)	lego ; star

Table 8. Semantic classes discovered for *vaisseau*

Five semantic classes are automatically discovered for *vaisseaux*, and four for the singular form *vaisseau*. These two distinct sets of semantic classes are obtained from two distinct corpora : the first one from the large corpus of *vaisseaux*, and the second one from the large corpus of *vaisseau*. In the present and next sections of the paper, we will refer to the semantic classes discovered specifically for the plural form *vaisseaux* and the ones for the singular form *vaisseau*, respectively 'PL semantic classes' and 'SG semantic classes'.

If we look at the content of these 9 semantic classes, the following senses can be identified : the PL class 'sanguins' and the SG class 'sanguin' concerned the anatomic sense of the French noun VAISSEAU, the PL classes 'spatiaux' and soyouz' and the SG class 'spatial' concerned the spaceship sense, and finally the PL class 'frégates' and the SG class 'lieutenant' concerned the ship sense. The architectural and botanic senses founded in the lexical entry of the dictionary TLFi given in the appendix, are nevertheless present in the short corpus of *vaisseaux*, but as only both senses occur in only three contexts of *vaisseaux*, no semantically pertinent collocates of these two meanings have been extracted for building the large corpus, and none appear either in the set of collocates building during the step 1. In the short corpus of the singular form *vaisseau* only the botanic sense occurs, but in only one context.

Even if the anatomic, spaceship and ship senses concerned both sets of semantic classes, sizes of the SG and PL classes concerning each of these senses are quite different, and there is almost no lexical overlapping. Comparing PL and SG semantic classes between them, around 29 words are in both, that is to say 10% of the total words founded in the two sets of semantic classes. The 12 following words are found for example in the classes 'spatiaux' and 'spatial' : *canons, commandes, construire, détruire, guerre, histoire, jeu, orbite, planète, site, terre, type*. Three, namely *coeur, niveau* and *sang* are both in the PL class 'sanguins' and in the SG class 'sanguin'. The examples given above concern couples of PL and SG classes that involved the same sense of VAISSEAU, but the overlapping also entails couple of PL and SG classes of a completely distinct meaning. For instance, the six following words *combat, grand, gros, nom, petit* and *système* are found in the SG class 'spatial' that concerns the spaceship sense and in the PL class 'sanguins' concerning the anatomic sense. We can notice that many of these of words are very general ones, so that overlap is not surprising, but we have to see if this have an impact on the results of the semantic categorization of the syntactic collocation.

3.5.2 Results of the categorization of collocations of the French noun VAISSEAU

For our processing, we consider only syntactic collocations that occur at least 10 times in the corpus (given the lemma level). Tables 9 and 10 give the ten first nom + PREPde + nom

collocations in which appears the lemma VAISSEAU according to the frequency, respectively within the SG semantic classes and the PL ones.

NOM + PREPde + NOM				
Class name	**Nb total coll**	**Nb total Avec MC**	**Nb coll in the pattern**	**First ten collocations**
lieutenant	28	13	7	vaisseau & de royale (16) ; fait & de vaisseau (16) ; vaisseau & de réserve (15) ; vaisseau & de commandement (12) ; marin & de vaisseau (10) ; vaisseau & de roy (10) ; vaisseau & de mautort (10)
sanguin	209	87	23	formation & de vaisseau (109) ; obstruction & de vaisseau (55) ; vaisseau & de peau (54) ; cellule & de vaisseau (50) ; inflammation & de vaisseau (49) ; lésion & de vaisseau (36) ; contraction & de vaisseau (31) ; perméabilité & de vaisseau (29) ; vasoconstriction & de vaisseau (28) ; éclatement & de vaisseau (25)
spatial	577	312	64	capitaine & de vaisseau (796) ; lieutenant & de vaisseau (589) ; paroi & de vaisseau (522) ; vaisseau & de guerre (520) ; enseigne & de vaisseau (240) ; vaisseau & de ligne (216) ; flotte & de vaisseau (179) ; intérieur & de vaisseau (156) ; vaisseau & de flotte (154) ; vaisseau & de classe (145)

Table 9. Ten first NOM + PREPde+ NOM collocations with VAISSEAU categorized in SG semantic class

NOM + PREPde + NOM				
Class name	**Nb total coll**	**Nb total Avec MC**	**Nb coll in the pattern**	**Ten first collocations**
frégates	60	32	11	vaisseau & de compagnie (38) ; vaisseau & de haut (21) ; escadre & de vaisseau (21) ; vaisseau & de marine (18) ; vaisseau & de rang (18) ; vaisseau & de royale (16) ; tête & de vaisseau (16) ; vaisseau & de bord (16) ; genre & de vaisseau (11) ; vaisseau & de escadre (10)
spatiaux	466	227	65	flotte & de vaisseau (179) ; lancement & de vaisseau (137) ; vaisseau & de combat (123) ; commande & de vaisseau (121) ; commandant & de vaisseau (108) ; vaisseau & de transport (104) ; vaisseau & de angoisse (73) ; vaisseau & de exploration (65) ; vaisseau & de star (63) ; pilote & de vaisseau (48)
sanguins	498	258	71	paroi & de vaisseau (522) ; dilatation & de vaisseau (357) ; intérieur & de vaisseau (156) ; type & de vaisseau (130) ; réseau & de vaisseau (122) ; formation & de vaisseau (109) ; nom & de vaisseau (81) ; rupture & de vaisseau (79) ; nombre & de vaisseau (78) ; niveau & de vaisseau (67)
soyouz	76	26	5	vaisseau & de secours (23) ; arrimage & de vaisseau (19) ; version & de vaisseau (13) ; vaisseau & de shenzhou (11) ; décollage & de vaisseau (9)

Table 10. Ten first NOM + PREPde+ NOM collocations with VAISSEAU categorized in PL semantic class

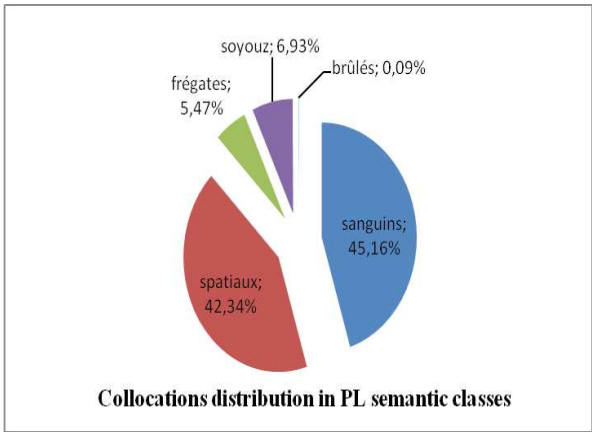
Not all the collocations will be attached to a semantic class. Firstly, only those that have at least one collocate that appears in the lexical set of the semantic classes will be considered for

the semantic categorization. Secondly, those that do have at least one collocate that appear in one of the semantic classes but that don't match P1 and P2 will not be classified.

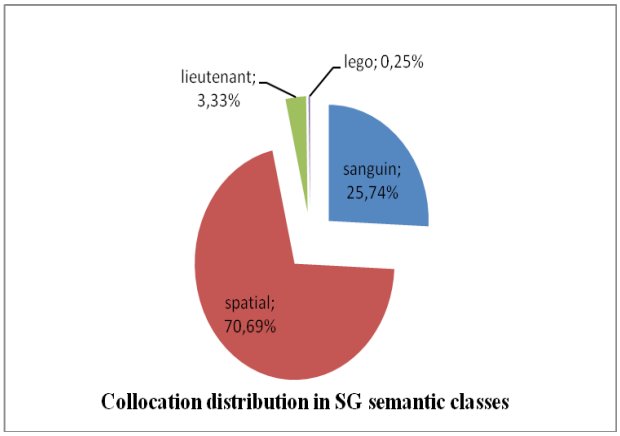
Among the collocations that occur at least 10 times in the large corpus gathering contexts of *vaisseaux* and *vaisseau*, 32 were not given in input to the semantic categorization collocations processing according the set of PL semantic classes, because there is no overlap between their lexical distribution and the content of the classes. And 39 were not given in input to the semantic categorization collocations processing according the set of SG semantic classes for the same reason. The set of PL semantic classes has in total 140 words, and the SG set one has 159 words. The involved collocations here are not very frequent, but while the 32 collocations are not semantically pertinent, the majority of the 39 collocations that have not been given in input are pertinent to the anatomic sense of VAISSEAU. This can be explained by the small size of the semantic class 'sanguin'. As regards the 32 collocations kept out because of the absence of overlapping with the PL semantic classes, most concern a person, which is a commander of a vessel, and one, *vaisseau & de nef*, concerns the architectural sense of VAISSEAU. In fact, in the whole large corpus, that is to say with the contexts of *vaisseaux* and *vaisseau*, 88 contexts of both word-forms concern these last sense, but, firstly, the syntactic collocations extracted were not given in the input since they occur less than 10 times, and, secondly, these collocations would not have been categorized in one of the semantic classes discovered because of the absence of a semantic class that represents the 'architectural' sense of VAISSEAU, even if this sense is associated with specific terms like *vouté*, *travée*, *choeur*, *abside, transept, charpente*. Finding a way to automatically detect such very infrequent meanings in the corpus is an aspect that remains to be explored.

In total, 2128 syntactic collocations were given in input for the set of semantic classes detected for the singular form *vaisseau*, and 2134 were given according to the one of the plural form *vaisseaux*. In view of the automatic categorization collocations processing described earlier, a collocation could be classified, according to the same set of semantic classes, in several semantic classes. Nevertheless, this does not happen with the data obtained for the French noun VAISSEAU.

Some 1096 collocations have been classified within the semantic classes of the plural form *vaisseaux* (51% of the total collocations given in input), and 812 within the semantic classes discovered for the singular form *vaisseau* (38% of the total collocations given in input). Graphs 1 and 2 below show the detailed distribution (expressed in percentages) of the collocations according respectively to the PL semantic classes and to the SG semantic classes.



Graph 1. Collocation distribution in PL semantic classes

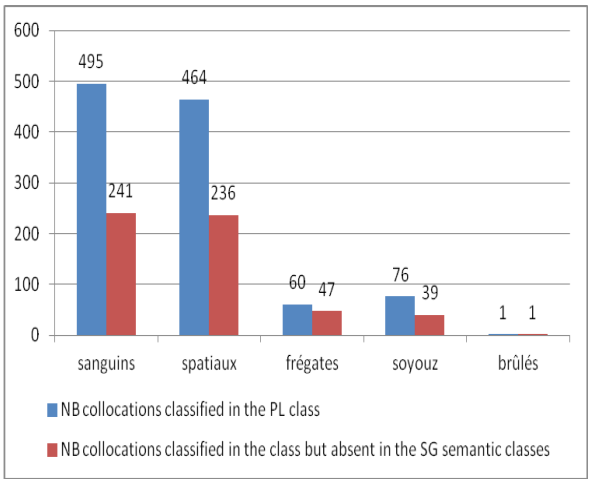


Graph 2. Collocation distribution in SG semantic classes

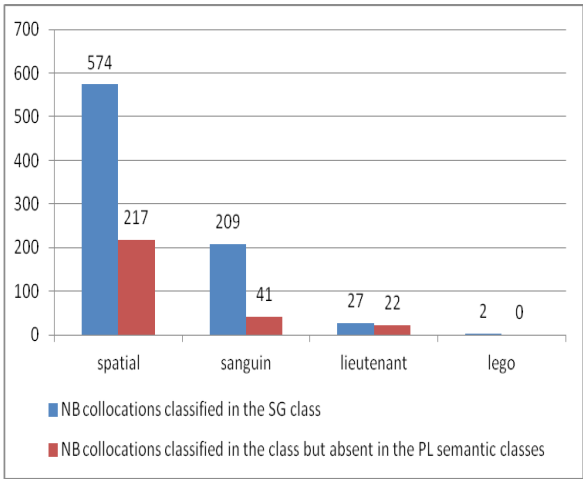
Regarding the distribution of the PL semantic classes (see Graph 1), the classes 'spatiaux' and 'sanguins' have almost the same distribution. This is also the case for the classes 'frégates' and 'soyouz'. But, while the classes 'spatiaux' and 'sanguins' constitute the major classes of the PL semantic classes in term of the number of syntactic collocations classified in, the two others are minor classes. If we now look at the distribution of the SG semantic classes (see Graphic 2), these are highly skewed in favour to the class 'spatial'. The class 'sanguin' is the second important class, but there is a big difference in number of collocations between the classes 'spatial' and 'sanguin'.

Comparing the results between the SG and PL semantic classes, three situations appear. The first situation arises when a given collocation is classified only in one of the semantic classes of only one word-form, as it is the case for the collocation *DILATATION & DE VAISSEAU* attached to the PL class 'sanguins' but any SG semantic classes. The second situation occurs when a given collocation is attached to a PL semantic class and to a SG class as well, and when these two classes are concerned with the same sense of VAISSEAU. The last situation shows that, similarly to situation 2, when a given collocation is attached to a PL semantic class and to a SG class, these two classes are not concerned the same sense of VAISSEAU.

Concerning situation 1, of the 1096 collocations classified within the semantic classes of the plural form *vaisseaux*, 564 have not been classified in any semantic classes discovered for the singular form *vaisseau*. Amongst the 812 collocations that have been classified within the semantic classes of the singular form *vaisseau*, 280 have not been classified at all in the semantic classes discovered for the plural form *vaisseaux*. Graphs 3 and 4 below give the statistical information.



Graph 3. Statistics data on the collocations classification within PL semantic classes



Graph 4. Statistics data on the collocations classification within SG semantic classes

Concerning the nature of the collocations classified only in one set of semantic classes, these concern semantically pertinent collocations, but more importantly they match the sense concerned by the semantic class involved. Table 11 below gives the ten first collocations (in terms of raw frequency), independently to the grammatical pattern.

Classified in the PL class 'sanguins', but absents whithin all SG classes	Classified in the SG class 'sanguin', but absents whithin all PL classes	Classified in the PL class 'spatiaux', but absents whithin all SG classes	Classified in the SG class 'spatial', but absents whithin all PL classes
capillaire & vaisseau	augmenter & résistance	impérial & vaisseau	spatial & vaisseau
dilatation & de vaisseau	rétiniens & vaisseau	vaisseau & détruire	à_bord_de & vaisseau
nerf & vaisseau	caillot & dans vaisseau	embarquer & dans vaisseau	capitaine & de vaisseau
lymphatique & sanguin	renforcer & de vaisseau	partie & de vaisseau	fantôme & vaisseau
réseau & de vaisseau	sur & coeur	jeu & de vaisseau	amiral & vaisseau
coeur & vaisseau	rétrécissement & de vaisseau	attaque & de vaisseau	lieutenant & de vaisseau
nouveau & sanguin	muscle & vaisseau	pas & de vaisseau	vaisseau & de guerre
dilater & vaisseau	lésion & traumatique	vaisseau & utiliser	DIGIT & vaisseau
circuler & dans vaisseau	dans & cerveau	plus & rapide	ennemi & vaisseau
rupture & de vaisseau	par & caillot	géant & vaisseau	grand & vaisseau

Table 11. Ten first syntactic collocations classified in only one of the two set of semantic classes.

Between the two classes that concern the ship sense of VAISSEAU, that is to say the classes ‘frégates’ and ‘lieutenant’, there is little overlapping between the collocations classified : 78% of the collocations classified in the PL semantic class ‘frégates’ are not classified in the set of SG semantic classes. This is almost the same statistics apply to the collocations attached to the SG class ‘lieutenant’. This can be explained by the lexical content of these two semantic classes, in that they shared only the word-form *marine*.

Concerning situations 2 and 3, 532 collocations are classified in both. Table 12 gives their distribution.

Nb collocations		SG Classes		
		spatial	*sanguin*	*lieutenant*
PL classes	*sanguins*	86	**168**	0
	spatiaux	**225**	0	1
	frégates	9	0	**4**
	soyouz	37	0	0
	brûlés	0	0	0

Table 12. Distribution of the collocations that are classified in the two set of semantic classes of VAISSEAU

Given situation 2, in total 434 collocations are concerned : 225 are indeed classified in the PL class 'spatiaux' and SG class 'spatial', 168 are classified in PL class 'sanguins' and SG class 'sanguin', 4 are in PL class 'frégates' and SG class 'lieutenant', and 37 are in the PL class 'soyouz' and in the SG class 'spatial'. Looking at the set of PL semantic classes, unlike the set of SG semantic classes, two classes denoted the 'spaceship' sense of VAISSEAU (the classes 'soyouz' and 'spatiaux'), but the classes do not overlap, since none of the collocations are classified in more than one class in the same set. Comparing the size of these classes, the class 'soyouz', with only two words (*soyouz* and *espace*), is very small compared to the class 'spatiaux'. Nevertheless, thanks to this small size and importantly to the presence of the specific term *soyouz*, several semantically pertinent collocations could have been classified. Of course, it is a very specific semantic class that covers one particular spaceship, with syntactic collocations such as *tma-9 & vaisseau*, *DIGIT & shenzhou*, *progress & ravitaillement*, *DIGIT & soyouz*, *spatial & vi*, *lancement & habiter*, *vaisseau & de secours*, *vaisseau & à station*. Naturally, the word *espace* is a general term in the spaceship sense, while the word *soyouz* is very specific in that it is a series of 'real' spacecraft – indeed, there are a lot of contexts of *vaisseau* and *vaisseaux* that occur in science-fiction books, movies or TV series, specifically with Star Trek or Star Wars. Consequently, the class 'soyouz' shows that a very small semantic class in term of size could supply several pertinent collocations, if at least one word is very specific. Naturally, some collocations involving the word *soyouz* are categorized in the class 'spatiaux', such as *soyouz & vaisseau*, *à_bord_de & soyouz* and *soyouz & spatial*. Table 12 above shows that 37 collocations attached to the PL class 'soyouz' (49%) are classified in the SG class 'spatial', and 51% of the collocations categorized in the PL class 'soyouz' are not categorized at all in any of the SG classes. This state confirms first that this class is strongly linked to the 'spaceship' sense of VAISSEAU, and second that this class is semantically specific. Collocations categorized in PL class 'soyouz' and absent in the categorization for the

PL semantic classes are semantically pertinent, because of the very specific sense of the word *soyouz*.

Given situation 3, 86 are concerned, since they are classified in two classes that imply two radical distinct senses, namely the anatomic sense (the semantic class 'sanguins') and the spaceship sense (with the SG class 'spatial') of the French noun VAISSEAU. But, this is not the case between the respective SG class 'sanguin' and the respective PL class 'spatiaux'. Looking at these 86 collocations, we first notice that the majority of these collocations are semantically vague since they are shared by different senses of VAISSEAU, such as *gros & vaisseau*, *intérieur & de vaisseau, nombreux & vaisseau, petit & plus, autour_de & vaisseau*, *envahir & vaisseau*. Naturally, two different senses of the same word can have a part of their lexical combinations in common, particularly when distinct senses are of the same nature as in the case of the 'anatomic' and 'spaceship' senses of the French noun *vaisseau*, both denoted a concrete object. So, in describing the size of the a vessel or a vehicle, the adjectives *gros* and *petit* could be used. Nevertheless, the adjectives denoting a huge size are only used when describing the spaceship. We have for example the following collocations *immense & vaisseau, énorme & vaisseau, gigantesque & vaisseau*, all three classified in the SG class 'spatial' and in the PL class 'spatiaux'. Secondly, among these 86 collocations, we still see a minority of collocations that are semantically pertinent for only one of these two senses of VAISSEAU. For instance, the collocations *sanguin & vaisseau*, *lymphatique & vaisseau*, *paroi & de vaisseau* and *artère & veine* are strongly linked to the anatomic sense, but are classified in the SG class *spatial*, while correctly classified in the PL class *sanguins*. To explain this incorrect categorization, we have to look at the collocates of the syntactic collocations, for example with the collocation *sanguin & vaisseau*. This collocation occurs 4103 in the corpus and has in total 4620 collocates, of which 77 appear within the content of the SG semantic classes : 70 in the class 'spatial', 6 in the one 'sanguin' and one in the class 'lieutenant'. P1 and P2 for each of these three classes are the following : 'spatial' (P1 = 0,91 ; P2 = 0,51), 'sanguin' (P1 = 0,08 ; P2 = 1), and 'lieutenant' (P1 = 0,01 ; P2 = 0,07).

This example shows the negative aspect of the current automatic semantic collocation process in raising two problems. First, as the ratio of P1 is the total number of collocates of the collocation involved in any semantic classes of the given set, P1 of *sanguin & vaisseau* calculated for the class 'sanguin' is very low while P2, which the ratio is the size of the given semantic class, is the maximum. This explains the categorization within the class 'spatial'. Second, in order to precisely understand the categorization of *sanguin & vaisseau* in the class 'spatial', we have to look at the nature and frequency of the 70 collocates of *sanguin &*

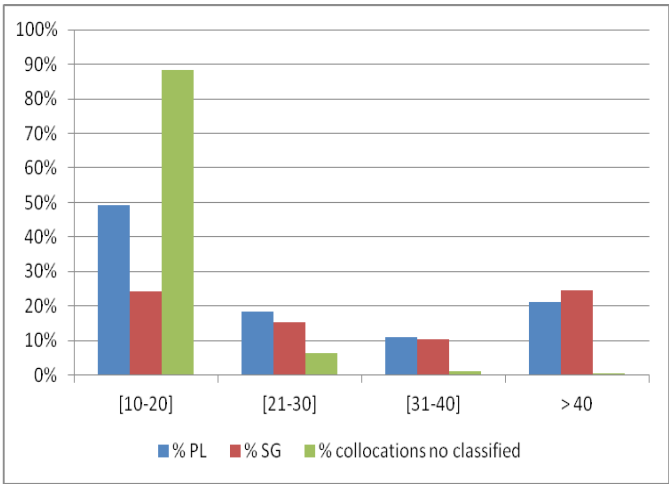
vaisseau that appear in this class. These collocates are the following (sorted according their frequency in the context of the given collocation) :

> système (66) ; gros (55) ; intérieur (38) ; petit (25) ; petite (20) ; forme (19) ; fin (15) ; type (15) ; grande (12) ; base (10) ; grand (10) ; nom (9) ; retour (8) ; seul (8) ; temps (7) ; image (6) ; capable (6) ; réussi (6) ; images (6) ; place (6) ; prendre (5) ; espace (5) ; première (5) ; porte (5) ; lieu (5) ; combat (4) ; ligne (4) ; vitesse (4) ; site (4) ; détruire (4) ; pose (3) ; mis (3) ; planète (3) ; années (3) ; part (3) ; mère (3) ; contrôle (3) ; vie (3) ; succès (3) ; transportant (3) ; vient (3) ; construit (3) ; construire (3) ; jours (3) ; prend (2) ; pris (2) ; placé (2) ; ville (2) ; mise (2) ; appelé (2) ; monde (2) ; heures (2) ; homme (2) ; spatial (2) ; série (2) ; nouvelle (2) ; deuxième (1) ; terre (1) ; français (1) ; siècle (1) ; histoire (1) ; lancement (1) ; mission (1) ; lune (1) ; ravitaillement (1) ; guerre (1) ; bord (1) ; fusée (1) ; rejoindre (1) ; équipage (1) ; écrase (1)

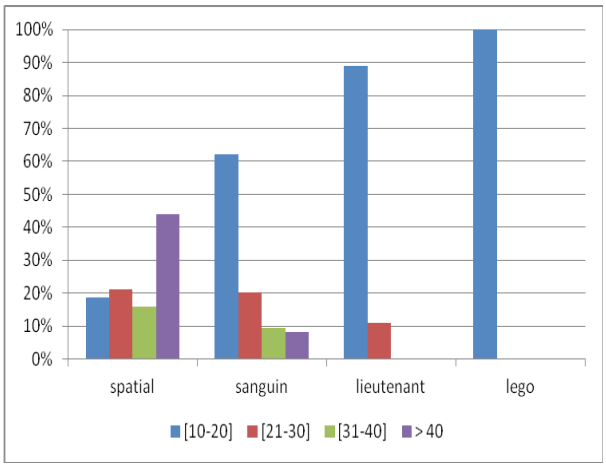
Most of these collocates have a low frequency, but more importantly most are not semantically pertinent to the anatomic sense of the collocation *sanguin & vaisseau*. On the contrary, the six collocates of *sanguin & vaisseau* that appear in the class ‘sanguin’, namely *artère*, *caillot*, *coeur*, *niveau, sang* and *sanguin*, are semantically pertinent and, as well, occur frequently with the collocation : *sanguin* (596 occurrences), *sang* (196 occurrences), *coeur* (163 occurrences), *niveau* (69 occurrences), *artère* (37 occurrences) and *caillot* (36 occurrences).

The *sanguin & vaisseau* example raises a degree of pertinence between the words that fill the semantic classes. The automatic semantic categorization of syntactic collocations processing can be improved by calculating a degree of pertinence for each word, according to the semantic proximity score between first a given word and other words within its own semantic class and second the given word and other words within other semantic classes.

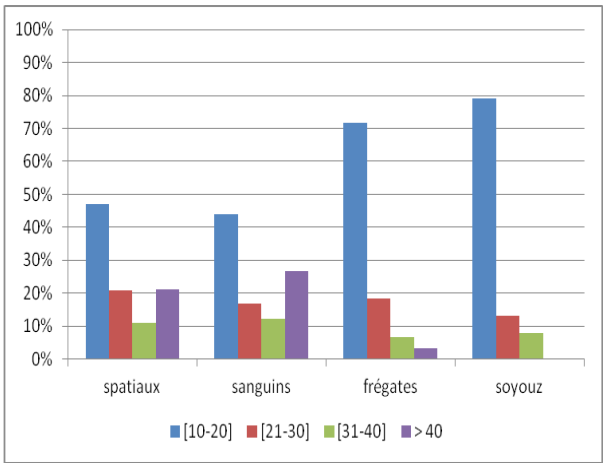
Lastly, we must look at the collocations that have not been classified either within the PL classes, or within the SG classes. Graph 5 given below shows the distribution of collocations not classified following the four intervals of frequency, namely [10-20], [21-30], [31-40] and > 40. Graph 5 also shows the distribution in the four intervals of frequency of the collocations that have been classified in the PL semantic classes and in the SG semantic classes. Graphs 6 and 7 show the detailed distribution of collocations that have been classified in the PL semantic classes (Graph 7) and in the SG semantic classes (Graph 6) in this four intervals of frequency according each semantic class.



Graph 5. Collocations data according to intervals of frequency



Graph 6. Intervals of frequency among SG semantic classes



Graph 7. Intervals of frequency among PL semantic classes

First, graph 5 shows that the majority of collocations no classified at all are of lower frequency ([10-20 interval of frequency). Graphs 6 and 7 show that the size of the semantic classes plays an important role. Indeed, the smaller semantic class, the more collocations of lower frequency will be classified. Conversely, the syntactic collocations that frequently occur (more than 40 times), have more chance of being classified in large (in term of size) semantic classes.

4 Conclusion

In an endeavour to isolate phraseological units within a given language, the use of a corpus makes possible the identification of collocations in language use, in this case as found on internet. Judging the acceptability of a given lexical combination is indeed particularly difficult, since this decision is generally based on the intuitive phraseological knowledge of the analyst. Corpus data can not only be used to compare collocations identified from the linguist's

introspective knowledge to regular lexical co-occurrences arising from by many language users attested in a corpus, but also the identification of collocations that would have simply been forgotten, or not thought through introspection alone.

Many methods of automatic extraction of collocations candidates have been proposed and advanced concordancing tools such as *Word Sketch Engine* or *Xaira* have emerged. Such tools provide great help in the identification of collocations, as conventional expressions in a given language. For lexicographical works, lists of potential collocations of a given word are automatically built, using, in the case of Work Sketch Engine tool statistics and syntactic dimensions, so the analyst has only identify the meaning of the collocations in order to treat each meaning of the target word. However, in the case of the compilation of a dictionary of collocations, if the collocations of a word are simply listed regardless of its meanings, the dictionary will be of limited use, especially for a non-native speaker of the language. As emphasized Blumenthal (2008), the collocational profile has to be a semantic collocational profile of a given word, because if not "these data only reflect the polysemous richness of the target word, issue irrelevant when one examines the conceptualization of a certain reality by the combinatorial possibilities of language."[1]. This is where differences lie with the early work of Palmer, and much general language phraseology, as these list phraseological units essential by linguistic properties, in a Firthian approach, it is the wider context that is necessary to demonstrate meaning potentials and context of use. Sinclair's idiom principle requires that we go beyond the so-called bound collocations of so-called general language to explore more precisely contextualised usage.

Thus, the tools mentioned above apply statistical filters to the collocation extraction process. As a result, a number of relevant collocations to meaning(s) that are poorly represented in the corpus will be either excluded from the final list or located at the tail of the list and, hence, probably will not be analyzed by the analyst. The focus is on the significant collocations of a given target word, and not on those significant for a particular meaning of that given target word. Consequently, some relevant collocations will be probably unnoticed by the lexicographer who uses such tools. On the subject of methods of extracting collocations applying statistical measures, Sinclair (2004 [1970]) declares that “we can develop measures that are closer to our intuitions, of our thought. Some problems remain. One problem is the homography. It can distort any statistical analysis.”. Since then much more precise measures have been found, but the basic problem of relying too heavily on statistics remains.

[1] Translate from the French following Blumenthal’s citation : « ces données ne refléteraient que la richesse polysémique du mot de base, problématique peu pertinente quand on s’interroge sur la conceptualisation d’une certaine réalité par les possibilités combinatoires d’une langue ».

The main objective of the doctoral thesis of Millon (2011) was to establish a methodology for building, in a automatic way, the combinatorial profile of a given word according the its different meanings. This methodology would combine the Palmerian perpsective of locating phraseological units with a Firthian one of using their lexicographical environment for their categorisation. This drawing together of phraseological and contextualist threads becomes possible through the harnessing of the internet as a source of large amounts of data, whilst finding means to place that data within more precise contextual areas of use.

The methodology was tested on several French nouns that belong to the semantic field of 'vehicles' (air, shipping, land) , of which the noun VAISSEAU taken for illustration in this paper. To discriminate the meanings in corpus of this polysemous and homonymous word ('ship', 'spaceship', 'anatomic vessel', 'botanical vessel', 'nave') and to semantically classify collocations according to meaning potentials, an exclusively corpus-driven methodology is applied, which means that no linguistic sources, such as dictionaries, are exploited. Moreover, the senses of the word are discovered in corpus using the textual data of the corpus given the word-form level, because different word-forms of the same word could have different meanings. The result of the word sense discrimination is two sets of semantic classes, one per word-form of VAISSEAU, composed by words that have semantic proximity. Each semantic class stand for a particular meaning of VAISSEAU. Collocations extracted from corpora according the phraseological approach, that is to say in searching collocates linked between them by grammatical relations, are then attached to one or several semantic classes automatically discovered.

In this case, collocations are seen as having strong lexical syntagmatic links between words that can be identified via different viewpoints regarding the phraseological and contextualist approaches of the notion of collocation. Millon (2011) fixed her work in both of these approaches. Such semantic collocations database can be used for example to assist lexicographic tasks, such as the lexicographic project of Williams and Millon (2010). In currently ongoing work, the automatic meaning discovery processing will be adapted to the evaluation of the content of huge specialized web-based comparative corpora that have been built to represent a given particular domain, namely that of heritage, within the METRICC research programme.

5 References

Béjoint, H. (2010). The lexicography of English. Oxford: OUP.

Benson M., Benson E., Ilson R. (1986) (2ème edition) The BBI Dictionary of English Word Combinations. Amsterdam : John Benjamins.

Blumenthal P. (2008). "Histoires de mots : affinités (s)électives", in: J. Durand/B. Habert/B. Laks (éds.): *Congrès Mondial de Linguistique Française - CMLF'08*, Paris: Institut de Linguistique Française.

Firth J.R. (1957). Modes of Meaning. Papers in Linguistics 1934-1951. Oxford : OUP

Haussman F-J. (1985b). Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels, dans Bergenholtz H., Mugdan J., (eds) Lexikographie und Grammatik. Tübingen, Niemeyer.1985: 118--129 {= Lexicographica. Series Maior 3}

Hoey, M. (2005). Lexical Priming: A New Theory of Words and Language. London: Routledge.

Hunston S., Francis G. (2000). Pattern Grammar: A corpus-driven approach to the Lexical Grammar of English. Amsterdam et Philadelphie : John Benjamins.

Louw, B. (1993). 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in Baker, M. 1993. Text and Technology. Amsterdam: John Benjamins. Pp. 157-76.

Millon, C. (2011). *Acquisition automatique de relations lexicales désambiguïsées à partir du Web*. Thèse de doctorat, Université de Bretagne-Sud.

Moon, R. 1998. *Fixed idioms and expressions in English*. Clarendon Press: Oxford, U.K

Palmer, H. (1933). Second Interim Report on English Collocations. Tokyo: Kaitakusha.

Sinclair, J. McH., Jones, S., Daley, Robert. (1970|2004). English Lexical Studies: The OSTI Report. London and New York: Contiuum

Sinclair, J. McH. (1966). 'Beginning the study of lexis'. In Bazell C. E, Catford J. C., Halliday M. A. K., Robins R. H. (éds). In Memory of JR FIRTH. Londres : Longman, pp. 410-430.

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

Williams G. (1998). 'Collocational Networks : Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles'. International Journal of Corpus Linguistics. Vol 3/1 : 151-171.

Williams G and Millon C. (2010). Going organic: Building an experimental bottom-up dictionary of verbs in science. Proceedings of the 13th EURALEX International Congress, Leeuwarden, Pays Bas. Pp 1251-1257.

Appendices

Définition de *vaisseau* dans le TLFI (Trésor de la Langue Française Informatisé):

Sens A.1 : *Vx* ou *littér.* **Vase, récipient**.
Vaisseau d'argent, d'argile, de bois, de cuivre, d'or, de verre; vaisseau d'eau bénite. C'est à quoi doivent se borner les présens qu'on peut faire à des peuples qui, n'ayant point de vaisseaux propres à la cuisson des alimens dans les fluides, ne feraient aucun usage des légumes et des fruits qui exigent cette préparation pour être mangés (*Voy. La Pérouse*, t. 1, 1797, p. 206). *Il reprit le vaisseau des saintes huiles, en ôta le couvercle, vint se placer devant le lit* (Zola, *Rêve*, 1888, p. 190).

Sens A.2. *Vieilli.* **Récipient d'une forme ou d'une matière particulière destiné à un usage technique**. *Vaisseau vinaire. On le prépare* [*l'oxyde de mercure*] *(...) par l'évaporation jusqu'à siccité d'une dissolution nitrique de mercure, et la calcination du résidu en poudre dans un vaisseau de verre ou de grès non verni* (Kapeler, Caventou, *Manuel pharm. et drog.*, t. 2, 1821, p. 581). *Pour pouvoir suivre des yeux le mécanisme de la congélation, ils placent l'eau à geler dans un vaisseau en verre transparent, posé lui-même au milieu d'un mélange réfrigérant de glace et de sel* (Metzger, *Genèse sc. cristaux*, 1918, p. 132).
Sens A.3. *P. anal., RELIG. Vaisseau d'élection.* „Créature choisie pour sa pureté et sa sainteté`` (Littré). *Vaisseau d'iniquité.* „Méchant, pécheur obstiné`` (*Lar. 19e*).

Sens **B.1.** *ANAT.* **Organe tubulaire, canal par lequel circule un liquide organique**.
Vaisseau excréteur; vaisseau biliaire, galactophore, spermatique. Lorsque l'urine, limpide, aqueuse et sans couleur, annonce l'irritation et le spasme de ses vaisseaux sécrétoires, il faut détendre, relâcher et adoucir (Geoffroy, *Méd. prat.*, 1800, p. 374). *Vaisseau chylifère*. Vaisseau lymphatique**.
− *En partic. Vaisseau sanguin* et, p. ell., *vaisseau.* Artère, veine ou capillaire. *Vaisseau dilaté, oblitéré; rupture, oblitération d'un vaisseau* ; *vaisseaux anastomoses; vaisseau artériel* (vieilli). *Un lobule pulmonaire (...) est une masse pyramidale essentiellement constituée par des ramifications des bronches et des vaisseaux sanguins, le tout emballé par du tissu conjonctif* (Camefort, Gama, *Sc. nat.*, 1960, p. 140). V. *mydriase* ex.
♦ *Vaisseau veineux** (vieilli). *Vaisseau capillaire.* V. *capillaire*[1] B 1.
Sens B.2. *BOTANIQUE*
a) Petit canal par lequel circule la sève des végétaux.
Plante à vaisseaux (synon. *plante vasculaire*); *vaisseaux ligneux. Le bois se compose de vaisseaux spiralés et annelés appelés* faisceaux ligneux *vers le centre de la tige, et faisceaux* libériens *à la partie externe* (*Arts et litt.*, 1935, p. 22-1). *L'ascension de la sève brute se fait par la lumière des vaisseaux, elle relève de processus purement physiques dont les causes ne sont que partiellement connues* (*Hist. gén. sc.*, t. 3, vol. 1, 1961, p. 460).
b) *Vaisseau lactifère.* Canal dans lequel circule le lait des plantes à latex. V. *lactifère* ex. de H. de Graffigny.
Sens B.3. *P. anal.*, *ACUP.* Réunion de points ayant une action commune sur une zone particulière du corps.
C'est alors que les huit vaisseaux ont cette propriété d'accaparer cette énergie maléfique, de la retenir dans leurs « étangs », dans leurs « lacs », afin de l'éliminer peu à peu (R. de La Fuÿe, *L'Acup.*, 1956, p. 75).

Sens C.1. *MAR.* Synon. *bateau*[1], *bâtiment.*

a) Vieilli ou *littér.* **Navire de dimensions importantes servant au combat, au transport de passagers ou de marchandises**.
La plage entière est bordée de carcasses de vaisseaux naufragés, à demi ensevelis dans le sable; quelques-unes montrent encore leur haute proue fracassée (Lamart., *Voy. Orient*, t. 2, 1835, p. 57):

b) Mod. **Grand navire de combat muni d'une artillerie**.
Vaisseau de charge, de combat, d'escorte; lieutenant de vaisseau. Le 18 juin suivant les vaisseaux français rencontrèrent les vaisseaux russes *avec les escadres allemandes dans les eaux de Kiel* (Maurras, *Kiel et Tanger*, 1914, p. 20). *Nous disposons en Afrique française libre des forces de terre suffisantes et en Méditerranée des vaisseaux de guerre nécessaires pour l'escorte* (De Gaulle, *Mém. guerre*, 1954, p. 599).

c) **[P. allus. littér. ou myth.]**
♦ *Vaisseau fantôme**.
♦ *Vaisseau des Argonautes.* Nef sur laquelle Jason s'embarqua pour la quête de la Toison d'Or. *Par vous l'heureux vaisseau des premiers Argonautes Flotte encor dans l'azur des airs* (Chénier, *Ïambes*, 1794, p. 261). [Souvent avec majuscule] *Le Vaisseau des Argonautes*, p. ell., *le Vaisseau*. Constellation de l'hémisphère austral. *Nous trouvons le soir au bord oriental, le vaisseau céleste, appelé vaisseau des Argonautes par tous les anciens* (Dupuis, *Orig. cultes*, 1796, p. 237).

d) *P. anal.*
α) *Vaisseau du désert.* Chameau ou dromadaire utilisé comme monture. Synon. *méhari. Savez-vous que lorsqu'on a créé les premières troupes méharistes, en 1890, on a songé à faire appel à des marins, sous prétexte que chevaucher les « vaisseaux du désert » provoquait des nausées, comme le mal de mer?* (L. Gardel, *Fort Saganne*, 1980, p. 62).
β) Avion. *La réverbération est intense et notre vaisseau métallique* [*un avion*] *est lancé comme un volant par de puissants souffles chauds* (Morand, *Air indien*, 1932, p. 105).
γ) Véhicule de grandes dimensions. *La cour de Saint-Lazare était vide. Le flot humain des heures de pointe s'en était comme retiré. C'était un bassin vide, sans même à l'ancre ces rapides vaisseaux qui sont les autobus et qui mènent par leur réseau mille fois tissé au cours du jour, à travers la ville* (Vialar, *Bête de chasse*, 1952, p. 86).

e) *P. anal.* ou *au fig.* ou *p. métaph.* Chose concrète ou abstraite considérée dans la manière dont elle est dirigée, conduite ou soumise aux événements et phénomènes importants.
α) La Terre, la durée de vie terrestre. *Rappelle-toi-le bien; nous sommes sur ce vaisseau démâté pour souffrir. C'est un mérite, pour l'homme, que Dieu l'ait jugé capable de vaincre ses souffrances les plus graves* (Lautréam., *Chants Maldoror*, 1869, p. 153). *Songez maintenant à cette traversée sans fin, et toujours périlleuse, que nous faisons tous sur le grand vaisseau; songez qu'il n'y a point de port. C'est tempête de monnaie, tempête de travail, tempête de guerre toujours!* (Alain, *Propos*, 1927, p. 719).
β) Institution, ensemble de valeurs intellectuelles ou morales. *Le vaisseau de l'Église, de la religion. Le vaisseau de la République vogue, comme j'ai dit, entre deux écueils, le modérantisme et l'exagération* (Desmoulins ds *Vx Cordelier*, 1793-94, p. 130). *Il déprimoit sans relâche les causes finales, qu'il*

appeloit des rémoras *attachés au vaisseau des sciences* (J. de Maistre, *Soirées St-Pétersb.*, t. 1, 1821, p. 388).

γ) Pays, État. *C'est vous qui, sages et heureux pilotes, avez conduit le vaisseau public au port du bonheur, c'est-à-dire de la liberté* (*Le Moniteur*, t. 2, 1789, p. 511). *Quel allègement pour le vaisseau de l'État, en 1836, si tout ce qui a plus de cinquante ans passait tout d'un coup* ad patres *!* (Stendhal, *H. Brulard*, t. 1, 1836, p. 320).

f) *HÉRALD.* **Représentation d'un bâtiment de mer dont on ne peut préciser l'espèce**. *D'azur, à un vaisseau de guerre équipé d'or, ayant au pavillon de poupe les armes de Rohan* (Grandm. 1852).

Sens 2. a) ARCHIT. Espace que forme à l'intérieur d'un édifice une voûte allongée ou ovoïde de vastes proportions. Inégalement semées à travers la forêt de piliers et d'arcades qui soutient les trois nefs de la cathédrale, ces masses de lumière éclairaient à peine l'immense vaisseau (Balzac, M^e^ Cornélius, 1831, p. 200). La première chorale des Cheminots (...) s'est fait entendre dans l'immense vaisseau du cinéma Gaumont (Enseign. mus., 2, 1950, p. 10).

− P. ext. Édifice aux dimensions importantes. Vers Reims, la vaisseau de la cathédrale dominait la mer des maisons (Hamp, Champagne, 1909, p. 178). On apercevait un océan de têtes moutonnant, grondant, s'écrasant contre les flancs du lourd vaisseau de pierre de l'hôtel de ville (Van der Meersch, Invas. 14, 1935, p. 411).

b) *AÉRON., ASTRON.*

α) *Vieilli. Vaisseau aérien.* **Dirigeable, aéronef**. *Röthe a proclamé à Berlin, au commencement de la guerre, que, dans cette dure période, le* Faust *de Goethe, la* Symphonie héroïque *de Beethoven (...) luttaient autant contre les adversaires de l'Allemagne que les canons de Krupp et les vaisseaux aériens de Zeppelin* (Barrès, *Cahiers*, t. 14, 1922, p. 81).

β) Engin spatial de grandes dimensions souvent conçu pour être habité. Synon. *astronef* (rem. 2 *s.v. astro-* I A), *cosmonef* (rem. *s.v. astro-* I A), *sonde* spatiale, spationef* (rem. 2 *s.v. spatial*). *Vaisseau spatial, interplanétaire, habité; vaisseau de l'espace. Les équipages des vaisseaux lunaires que Russes, puis Américains comptent envoyer d'ici là subiront ces effets* [*des radiations*] (*Le Figaro littér.*, 22 juin 1963 ds Guilb. *Astronaut.* 1967).