What are you afraid of?
The construction of meaning
in *X-(o)phobie* lexemes

Marine Lasserre
CLLE-ERSS & University of Toulouse
marine.lasserre@univ-tlse2.fr

Abstract: French neoclassical compounds *X-(o)phobie* display two distinct meanings : 'fear' and 'hostility'. In order to determine which meaning is involved in a given complex word, *X-(o)phobie* lexemes are studied according to the contexts they are used in. In fact, in a non-compositional analysis, complex words cannot be studied in isolation since they construct their meaning whenever they are employed. The distributional hypothesis offers a new way to examine the semantics of complex words. Two methods make it possible to investigate the variation of meanings in *X-(o)phobie* lexemes. Moreover, the use of Google data enables an analysis of the construction of neologisms meaning.

Keywords: French; neoclassical compounds; distributional analysis; semantics; morphology

> *I hate the word* homophobia. *It's not a phobia. You are not scared. You are an asshole.*
> (Morgan Freeman on Twitter)

1. Introduction

This paper deals with neoclassical compounds constructed by means of the element *-(o)phobie*. Two readings can be provided for *X-(o)phobie* lexemes: 'fear of *X*' and 'hostility against *X*'. Taking into account the co-text may help the hearer determine which of these two meanings is intended in a given sentence, and allow the speaker to attribute one of the meanings to a coined lexeme. The paper aims at setting up a protocol that could then be automatically implemented.

Neoclassical compounds were massively introduced into many European languages from the 17th century onwards, initially to meet the needs of emerging scientific disciplines faced with the task of denoting new fields, tools, and concepts. These complex lexemes, which are built on Ancient Greek or Latin constituents, are now widely used in everyday language, especially in French. They are non-native complex lexemes, which means first that the order of constituents is reversed as compared to native constructs

(neoclassical compounds are right-headed whereas native compounds are left-headed) and second, that their constituents are not available in the French lexicon, in other words, their constituents are not autonomous French lexemes.[1] The origin of these constituents raises many issues for the analysis of modern languages, among which the construction of the meaning of such lexemes by speakers who have no etymological knowledge, i.e., who lack an understanding of the semantics of classical constituents. Somewhat surprisingly, speakers who are unacquainted with classical languages are nonetheless able to construct and interpret lexemes involving Latin or Ancient Greek constituents. We may have learned in school that *-cracy* means 'power': *democracy* is 'the power of the people'. However, *aristocracy* does not mean synchronically 'the power of the best people' but 'a group with a self-identified high value'. Likewise, we all learned that *-phobia* means 'fear' but as Morgan Freeman in the quotation above shows, this is not always the case. The interpretation of such complex lexemes does not depend on the transparency of the complex lexeme or on understanding the classical constituent, as a segmental analysis might suggest. Rather, the assumption defended here is that speakers do not need to access the meaning of each element in order to construct the meaning of a complex lexeme; instead, they take lexemes which are already present in their lexicon as models. The notion of *leader word* (Rainer 2003; Roché 2011) is interesting in this respect in order to apprehend the lexical dynamics. A leader word is a complex lexeme which constitutes the origin of a model, the starting point of a serial effect. It initiated a pattern of formation (or contributed to establishing it) and, synchronically, still remains salient and frequent. In the structure of the mental lexicon, a leader word acts as a model on which new lexemes can analogically be constructed. It exerts a strong attraction and functions as a magnet: when a speaker builds a lexeme according to a pattern of formation, he has to place it at a certain semantic distance from leader words in order to give it a semantic interpretation. The position of the new construct with respect to each leader word determines the meaning it acquires. To go back to *-(o)cratie*, in French two leader words emerged: *démocratie* with the meaning 'power of X' (X represents the first constituent) and *aristocratie* with the meaning 'group with a self-identified high value'. A lexeme such as *cathocratie* (< *catho*, the clipped form of *catholique* 'Catholic') should not be considered as built by adding the base *catho-* to an element *-cratie* meaning 'power'. Rather,

---

[1] In fact, constituents may be autonomous lexemes in French, as in *théâtrothérapie* (< *théâtre* 'theater' + *thérapie* 'therapy'). Nevertheless, the order of the constituents shows that the complex lexeme is non-native.

it is constructed by taking *aristocratie* as a model via a process of analogy and means 'a group with a self-identified high value composed of Catholics' (for a fuller overview of analogy in the construction of the meaning of neoclassical compounds, see Lasserre & Montermini 2014a). But how can a lexeme be placed at a greater or lesser distance from a leader word in order to give it a certain semantic interpretation? The claim here is that in a semantic analysis of complex lexemes, it is essential to take the context into account since a lexeme is never built in isolation: the distribution of a new lexeme is crucial in a semasiological approach, since a hearer has to find clues concerning its semantic properties in order to interpret it correctly.

These preliminary considerations raise the issue of the status of the constructs under discussion. Neoclassical compounds have been extensively studied in various languages (see Bauer 1998; Fradin 2000; Lüdeling et al. 2002; Iacobini 2004; Amiot & Dal 2007; Villoing 2012 among others). These lexemes are traditionally defined by means of several properties, even though no single property is absolutely necessary. First, they form technical or scientific vocabularies; we will see however in this paper that neoclassical formations are commonly used in everyday language. Second, they include a linking vowel (/i/ for Latin constituents and /ɔ/ for Ancient Greek ones) at the junction of their constituents. In a non-segmental analysis, this linking vowel cannot be considered as an interfix,[2] nor can it be linked to either of the two constituents. It can be handled, however, using the notion of constraint, put forward by Optimality theory (Prince & Smolensky 1993); the presence of this vowel is a constraint that weighs on the output of the construction. This property is then non obligatory but very frequent (another vowel or even a consonant could show up). Its presence is not only due to morphophonological conditions but also to serial effects. For example, Lasserre (2013) argues that the origin of the first constituent does not affect the presence or absence of the linking vowel: lexemes constructed with the elements *-logie* and *-phone* (to name speakers) display the vowel /ɔ/ in the same proportions, even though almost all those with *-phone* have non classical first constituents, i.e., names of languages. More than a morphophonological condition, a serial effect

---

[2] An anonymous reviewer suggested that the linking vowel could be interpreted as an interfix. Nevertheless, on the one hand, an interfix should have the form of a suffix, be added to a monosyllabic base and have the structure VC (Roché 2003), which is not the case for /ɔ/ in *X-(o)phobie* lexemes. On the other, it is different to consider the presence of the vowel as a constraint on the ouput or as a segment, even if this segment is meaningless. The serial effect that prevails in the construction of *X-(o)phobie* lexemes dictates the presence of the vowel.

requires the vowel: the complex lexemes have to look like other neoclassical compounds. The vowel is then included in the spelling of neoclassical elements, between parentheses: it does not belong to the constituent *-phobie* but is highly likely to occur in the whole compound. Lastly, neoclassical compounds are formed by means of neoclassical elements, that is to say elements which were either autonomous lexemes in classical languages but lost their autonomy in modern languages, or non autonomous classical elements which were already used in Ancient Greek or Latin as construction elements (as we will see with *-phobie*). However, this umbrella term of neoclassical element refers in fact to a heterogeneous class. If these elements are viewed from a purely synchronic point of view, we observe that some of them can be considered as suppletive stems of lexemes (cf. Amiot & Dal 2007), while others form long series of lexemes and display a behavior which is closer to that of canonical affixes. Lasserre and Montermini (2014b) attempted to establish a typology of the final elements of complex lexemes and considered the elements which are particularly available for new constructions *(-(o)logie, -(o)cratie, -(i)cide, …)* as the exponents of constructions in the sense of Booij (2010): they are the formal manifestation of semantic and categorical operations applied to lexemes, just like affixes. The present study focuses on the French element *-(o)phobie* which is considered as belonging to the latter class. In section 2, I will present this constituent from a diachronic and a synchronic point of view as well as the corpus which was collected. To answer the main questions raised by this paper, I will propose in section 3 a distributional analysis, in order to take the context of the lexemes under discussion into account for the study of their meaning. Finally, as a conclusion, I will focus on the problem of homophony versus polysemy.

2. Lexemes involving the *-(o)phobie* element in French

In Ancient Greek, φόβος was an autonomous noun denoting a panic fear; combined with an autonomous lexeme, it could also form adjectival compounds with the meaning 'who is afraid of something', for example ἀεροφόβος, 'who is afraid of air'. The form φοβία never appeared as an autonomous lexeme, and was in fact fairly infrequent. Perseus,[3] which includes several dictionaries such as the Liddell-Scott-Jones Greek-English Lexicon, gives only two occurrences: ἀφοβία and ὑδροφοβία. Surprisingly, as can be seen in Table 1, in French and in other modern languages, the

---

[3] http://www.perseus.tufts.edu

equivalent of *φόβος* is not autonomous whereas the equivalent of -*φοβία* is. As already pointed out in the introduction, in French and in other modern languages (for instance in English) a second meaning emerged. According to Cottez (1980), the meaning 'hostility' was first employed in French by Joseph de Maistre in 1821 in the lexeme *théophobie* (< *théo-* 'God'). This meaning was no doubt initially a figurative extension of the 'fear' meaning. Nevertheless, its current widespread use, devoid of a 'fear' dimension, argues for a separate meaning. This hypothesis will be justified in section 3, as each meaning implies different distributions and belongs therefore to a separate class of distribution.

Table 1: Comparison of Ancient Greek and Modern French

Ancient Greek				French			
Lexeme	Category of lexeme	Gloss	Example	Lexeme	Category of lexeme	Gloss	Example
φόβος	N	'panic, fear'		*phobe			
X-φόβος	A	'afraid of X'	*ἀεροφόβος* 'afraid of air'	X-phobe	A/N	'(person) afraid of X'	*arachnophobe*
						'(person) opposed to X'	*homophobe*
*φοβία				phobie	N	'panic fear'	
X-φοβία	N	'horror of X'	*ὑδροφοβία* 'horror of water'	X-phobie	N	'fear of X'	*arachnophobie*
						'hostility towards X'	*homophobie*

For the needs of the present analysis, I collected a database based on two dictionaries of French (*Le Trésor de la langue française informatisé* (*TLFi*) and *Le Grand Robert* (online version)) and on Google. Since the main focus here is the construction of the meaning of complex lexemes, the Web is the only existing resource that gives us access to a large number of newly coined lexemes, even if this resource is often criticized for its volatility or unreliability. Hathout et al. (2009) listed some of the precautions that need to be taken when using data from the Web for morphological analyses. For the present study, a lexeme was accepted in the database even if it had a single hit on Google, on condition that the context of the occurrence was uncontroversial, for example another lexeme from the same derivational family or from the same derivational set.[4] This condition makes it possible to avoid typing errors. Table 2 shows that Google provided a large number of lexemes which were not recorded in dictionaries:

[4] According to Roché (2011) or Hathout (2011), two lexemes belong to the same derivational family if they are constructed on the same base, and are part of the same derivational set if they are constructed using the same process.

Table 2: Occurrences of *-(o)phobie* lexemes in the database from various sources

	Dictionaries	Google	Total
-(o)phobie	73	456[5]	529
	13.80%	86.20%	

As we have already seen, at least two meanings emerged in French for the bound element *-(o)phobie*, namely 'fear' and 'hostility'. For the first one, leader words such as those listed in (1a) were identified. All the lexemes that can receive this interpretation denote medical disorders. For the second meaning, the leader words in (1b) were identified. This interpretation leads to nouns referring to attitudes that need to be fought against. A third meaning emerged in the technical terminology of chemistry, 'chemical repulsion', exemplified in (1c). However, as the main focus of this paper is lexemes coined in everyday language, this meaning was not taken into account.

(1)   a.   claustrophobie 'claustrophobia'
          agoraphobie 'agoraphobia'

      b.   xénophobie 'xenophobia'
          homophobie 'homophobia'

      c.   hydrophobie 'hydrophobia'
          oléophobie 'oleophobia'

Unlike what happened with the complex lexemes *X-(o)cratie*, where leader words were the most frequent and the first lexemes which were coined according to a pattern of formation (see (1)), the determination of leader words for *X-(o)phobie* lexemes is much more difficult. As the present study takes a synchronic perspective, leader words were identified here principally on a frequency basis (based on the number of Google hits). Moreover, the presence of a lexeme in dictionaries reinforces the likelihood of its being a leader word. For example, while *christianophobie* has the same proportion of Google hits as *claustrophobie* (respectively 197,397 and 154,640 on July 14, 2013), the fact that it does not feature in dictionaries makes it unlikely to have the status of leader word, as it is probably shared by fewer speakers.

The leader words identified in (1) will serve as poles of attraction for the creation of new lexemes; newly coined lexemes will construct their meaning by analogy with the meaning of the leader words. Examples which

[5] These occurrences were collected between March 04, 2013 and July 08, 2013.

are best interpreted as bearing the 'fear' meaning are listed in (2a) and examples carrying the 'hostility' meaning are listed in (2b). These interpretations are the main ones found in Google but, as this paper claims, the lexemes can also have other meanings.

(2)   a.   nyctophobie ← nyct- 'night'[6]
             pisciphobie ← pisc(i)- 'fish'
             anginophobie ← angine 'throat infection'

       b.   artistophobie ← artiste 'artist'
             footballophobie ← football 'football, soccer'
             fromageophobie ← fromage 'cheese'

While the first constituent can contribute to interpreting an *X-(o)phobie* lexeme with either meaning, i.e., to placing *X-(o)phobie* in proximity to one of the two leader words, this is not sufficient. Intuitively, it can be assumed that if the first constituent *X* refers to pests, insects, etc., then the *X-(o)phobie* lexeme will probably have the meaning 'fear'. If *X* refers to a class of humans, *X-(o)phobie* will probably mean 'hostility'. Nevertheless, as shown in Table 3, the first constituent can also belong to another semantic category: how are we to interpret *X-(o)phobie* in that case? *Vélophobie* could refer either to the 'fear of bikes' (3a) or to 'hostility toward bikes and bikers' (3b):

(3)   a.   Moi j'ai la **vélophobie** :snif: quand je voie ou entend un vélo je suis au bord de l'arrêt cardiaque c'est handicapant surtout que je n'est pas de permis.
             'I have bike-ophobia :sniff: when I see or hear a bike I am on the verge of a heart attack it is incapacitating especially because I haven't got a driving licence.'

       b.   Face à un tel déluge de « **vélophobie** », les propagandistes de la petite reine ont pris la plume pour vanter les bienfaits de la vélocipédie et faire accepter ces merveilleux fous roulant sur leurs drôles de machines.
             'Faced with such a deluge of "bike-ophobia", defenders of the "little queen" [= bike] took up their pens to extol the benefits of biking and to gain acceptance for those magnificent men on their strange machines.'

Moreover, even if *X* belongs to one of the semantic categories listed above (i.e., 'animal', 'human'), the complex lexeme may be used with an unexpected meaning. This is the case of *clownophobie* which denotes the fear of clowns.

---

[6] The final vowel of the classic root citation form is deliberately omitted, since the vowel depends on the constraint which is weightier in the construction of the output. In *nyctophobie*, the output aligns on the paradigm of *X-(o)phobie* lexemes and has the vowel /ɔ/ whatever the classical root vowel may be, whereas in the following

Table 3: Complex lexeme meaning(s) according to the semantic type of the first constituent

X denotatum	Example	X-(o)phobie: 'fear'	X-(o)phobie: 'hostility'
pests, insects	*arachnophobie*	✓	
human being			✓
other	*vélophobie* 'bike-ophobia'	?	?
human being	*clownophobie* 'clown-ophobia' 'clown-ophobia'	✓	

If the first constituent is not sufficient, we have to turn to other means in order to construct and/or interpret an *X-(o)phobie* lexeme. An apparently unique lexeme can be used with different meanings. In (4), *Microsoftophobie* denotes either a disorder (albeit an imaginary one (4a)) or hostility towards the software giant (4b).

(4) a. Souffrant de **Microsoftophobie** aiguë, j'essaie de ne pas utiliser les programmes mis "gracieusement" à disposition par Bill Gates.[7]
        'Suffering from acute Microsoft-ophobia, I try to avoid using the software "graciously" offered by Bill Gates.'

    b. Je ne veux pas défendre Microsoft pour sa position monopolistique, mais c'est un peu lassant cette **Microsoftophobie** primaire et systématique.
        'I do not want to defend Microsoft for its monopolistic position, but this rabid and systematic Microsoft-ophobia is a bit boring.'

The simple observation of *X* is not helpful here; from a semasiological point of view, the only way to access the meaning of these lexemes is to observe their distribution. That is why we opted for a distributional analysis in order to analyze the meaning(s) of *X-(o)phobie*.

---

example, *pisciphobie*, the ouput prefers to remain faithful to the first constituent that is largely found as *pisci* in French.

[7] http://tinyurl.com/np99wrc

3. A distributional analysis

Distributional analysis is a relevant way to integrate the context in the study of the meaning of complex lexemes. This approach, based on the distributional method advocated by Harris (1954; 1968), has been mainly used in semantics and computational linguistics (for an overview, see Lenci 2008; Sahlgren 2008) but rarely applied to morphology (Lazaridou et al. 2013). The distributional hypothesis (DH) is a usage-based perspective and states that there is a correlation between the similarity of meaning and the distributional similarity of two lexemes: "the degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear" (Lenci 2008, 3). In other words, linguistic items sharing a similar distribution may have a similar meaning. Studying the linguistic contexts of given lexemes should therefore provide evidence for characterizing them semantically. Corpora are essential tools in this task because, as Lenci (*op.cit.*, 9) says, "as repositories of linguistic usages, they represent the primary source of information to identify the word distributional properties". Lasserre and Montermini (2012) applied this method in a cross-linguistic comparison between the French element *-cide* and the Italian elements *-cida* and *-cidio*. The present analysis of *-(o)phobie* adopts the same protocol.

3.1. Top-down method

The method employed here investigates the weak version of DH presented by Lenci (2008). This version only assumes a correlation between linguistic distribution and semantic content and "exploits such correlation to get better understanding of the semantic behavior of lexical items" (*op.cit.*, 14). Unlike the strong version of DH, which investigates the causal role that linguistic distribution may play in the formation of the semantic representation of a given lexeme, the weak version disregards the causal role.

For the top-down method, the first step was to select the ten most frequent lexemes in the database. Frequency is based here on the number of pages indexed by Google containing each lexeme.[8] By chance, five of these lexemes have the meaning 'fear' and five have the meaning 'hostility', as Table 4 shows.

---

[8] These searches were conducted on July 2014. While the mutability of Google might have produced different results on a different date, the important point here is not the number of occurrences in themselves but rather the comparative number of occurrences.

Table 4: The ten most frequent lexemes in the corpus

Lexeme	Gloss	Hits	'fear'	'hostility'
homophobie	'homophobia'	5,516,500		✓
xénophobie	'xenophobia'	3,321,100		✓
islamophobie	'islamophobia'	1,344,890		✓
dysmorphophobie	'fear caused by a physical defect' or 'fear of having a physical defect'	602,170	✓	
agoraphobie	'agoraphobia'	241,380	✓	
christianophobie	'hostility towards Christians'	197,397		✓
transphobie	'hostility towards transsexuals'	162,400		✓
claustrophobie	'claustrophobia'	154,640	✓	
photophobie	'photophobia'	124,520	✓	
aquaphobie	'fear of water'	78,695	✓	

The second step was to investigate the distribution of these ten lexemes in different resources. First, a sample of Web pages was collected. In order to obtain a list of the contexts in which each lexeme was used, I took the first two pages of Google hits for each,[9] in the singular and plural forms, and identified their distribution: qualifying adjectives, predicates, etc. Then I used Frantext, a corpus mainly composed of French literary texts, in order to obtain a broader range of contexts. Finally, I used two resources developed at Toulouse University, Les Voisins du Monde[10] and Les Voisins de Wikipédia,[11] which are based, respectively, on ten years of the French newspaper *Le Monde* and on the French Wikipedia. These two resources are tagged for syntactic context and are presented via an interface which makes it possible to search for the syntactic relations a lexeme enters into. The measure of mutual information which can be obtained via these two resources can be taken as a clue of a privileged link between two lexemes. Thanks to these different means, a list of contexts for each of the lexemes considered was obtained. The third step consisted in selecting the contexts which are the most relevant and frequent for each interpretation. The final step consisted in conducting a search on each lexeme with each selected

---

[9] Investigating the whole Internet is impossible. However, as the Internet is, at least today, the best resource to obtain a real account of how speakers use a language and coin new lexemes, this sample gives us an overview.

[10] http://redac.univ-tlse2.fr/voisinsdelemonde/

[11] http://redac.univ-tlse2.fr/voisinsdewikipedia/

distribution on Google and noticing any potential hits. The selected contexts[12] are listed in Table 5; each was searched on Google with several inflected forms, based on those most frequently encountered on Google.

Table 5: Implemented co-texts for the top-down method

	Selected co-texts	Google search[13]
'fear'	X souffrir DE _____ 'X suffer from _____'	"souffrir de (d') Xphobie" "souffre de (d') Xphobie" "souffres de (d') Xphobie"
	soigner _____ 'to treat _____'	"soigner ∗ Xphobie" "soigne ∗ Xphobie" "soignes ∗ Xphobie"
	crise DE _____ 'attack of _____'	"crise de (d') Xphobie" "crises de (d') Xphobie"
'hostility'	vague DE _____ 'wave of _____'	"vague de (d') Xphobie" "vagues de (d') Xphobie"
	_____ ambiante 'prevailing _____'	"Xphobie ambiante"
	_____ latente 'latent _____'	"Xphobie latente"

Obviously, this list needs to be expanded in order to take more contexts into account. The drawback with this non automatic method is that it does not generate a long list of contexts, but it is almost impossible to apply an automatic method to Google. Furthermore, it is impossible to identify an extended set of contexts which are restricted to only one interpretation. This explains why barely more than half of the lexemes in the corpus were found with at least one required context, as shown in Table 6.

In order to obtain these results, it was considered that, if at least one hit was found with a relevant context, the lexeme was used with the corresponding meaning. In fact, 223 lexemes in the corpus (42.15%) have

---

[12] While the distribution is given in terms of "context" in DH, the use of Google forces one to take into account only the co-text of lexemes, i.e., the immediately preceding or following context.

[13] This column represents the sequence that was input; double inverted commas restrict the search to the requested sequence only and the asterisk functions as a wild card, i.e., the possibility of having any word (in this case, any determiner).

Table 6: Results for the top-down method

Lexemes used only with 'fear' contexts	163	31%
Lexemes used only with 'hostility' contexts	43	8%
Lexemes used with both 'fear' and 'hostility' contexts	85	16%
Lexemes used with none of the implemented contexts	238	45%

less than 1,000 hits on Google, and 485 of them (91.68%) appear less than 10,000 times. It was therefore difficult to establish an arbitrary level of acceptability. The absence of any given sequence can consequently be fortuitous. Furthermore, neologisms, which account for a large proportion of the lexemes in the corpus, are in general less frequent on the Web. The probability of finding them with one of the selected contexts is thus weak.

In addition, the limited number of selected contexts led to a limited number of results. For example, the lexeme *kurdophobie* in (5) could not be captured by this method, even though it is clearly employed with the interpretation 'hostility' in the following example:

(5) La **kurdophobie** des Turcs conduit donc à de nombreuses aberrations de ce type. [...] la Turquie a encore beaucoup de chemin à faire dans le respect des droits du peuple kurde.
'The Kurd-ophobia of Turks leads to several aberrations. [...] Turkey has a long way to go in respecting the rights of Kurdish people....'

3.2. Bottom-up method

Another method had to be found in order to solve some of the problems posed by the top-down approach. This second method takes into account each co-text of each *X-(o)phobie* lexeme contained in the database. This method was called "bottom-up". To understand it, let us look at the following examples:[14]

[14] Glosses for (6) and (7): *claustrophobie* 'claustrophobia'; *homophobie* 'homophobia'; *couplophobie* 'couple-ophobia'; *christianophobie* 'Christian-ophobia'; *microsoftopho-bie* 'Microsoft-ophobia'; *souffrir de* 'suffer from'; *atteint de* 'affected by'; *aigu* 'acute'; *lutte contre* 'fight against'; *manifester contre* 'demonstrate against'; *primaire* 'elementary'; *soigner* 'treat'; *légère* 'superficial'; *observatoire de* 'observatory of'; *lésion cérébrale* 'brain damage'; *affection neurologique* 'neurological disorder'; *raideur* 'rigidity'; *blessure légère* 'superficial injury'; *maladie mortelle* 'fatal illness'; *mal incurable* 'incurable illness'; *maladie de cœur* 'heart disease'; *glaucome* 'glaucoma'; *rhumatisme* 'rheumatism'; *infarctus* 'heart attack'; *terrorisme international* 'international terror-

(6) a.   claustrophobie   souffrir de X; soigner X
     b.   homophobie   lutte contre X; X primaire
     c.   couplophobie   souffrir de X; X aigu; X légère; atteindre de X
     d.   christianophobie   lutte contre X; manifester contre X; observatoire de X
     e.   microsoftophobie X primaire; X aigu

(7) a.   souffrir de X lésion cérébrale; X affection neurologique; X raideur; X blessure légère
     b.   atteint de X maladie mortelle; X mal incurable; X maladie de cœur
     c.   aigu   glaucome X; rhumatisme X; infarctus X
     d.   lutte contre X terrorisme international; X trafic de stupéfiant; X immigration irrégulière
     e.   manifester contre X suppression d'emploi; X insécurité; X racisme
     f.   primaire   anticléricalisme X; antiaméricanisme X

The left-hand column in (6) contains *X-(o)phobie* lexemes selected according to the following criteria: the first two, *claustrophobie* and *homophobie* are those which are considered as the leader words, respectively for the 'fear' and 'hostility' interpretations. The following three are examples of neologisms, that is to say non-dictionarized[15] lexemes, from the corpus. The right-hand column contains some of the contexts found on Google for each of the five lexemes. In (7) some of these contexts are listed on the left; the right-hand column gives some of the most frequent collocations observed in the Les Voisins resources, with the help of the mutual information measure.

The neologism *couplophobie*, in (6c), is used with the contexts *souffrir de, aiguë* and *atteint de.* These three contexts in turn are mainly used with the names of diseases (7a,b,c). Hence, *couplophobie* belongs to the same class of distribution as *infarctus* or *glaucome*, that is to say these different lexemes share the same structures. Like *infarctus* or *glaucome*, *couplophobie* has been coined as a disease. *Claustrophobie* in (6a), which is not a neologism, shares the context *souffrir de* with *couplophobie* and therefore also belongs to this class of distribution: we can deduce that *couplopho-*

---

ism'; *trafic de stupéfiant* 'drug trafficking'; *immigration irrégulière* 'illegal immigration'; *suppression d'emploi* 'job loss'; *insécurité* 'insecurity'; *racisme* 'racism'; *anticléricalisme* 'anticlericalism'; *antiaméricanisme* 'anti-Americanism'.

[15] In order to avoid the lexicalized/non-lexicalized distinction, the term "dictionarized" is used here to indicate whether a lexeme was found in dictionaries or not. If not, it is considered as a neologism, though of course the notion of neologism is more complex than this.

bie was probably built by analogy with *claustrophobie. Christianophobie* in (6d) is used with *manifester contre* 'protest against', *lutter contre* 'fight against', which are not frequently used with the names of diseases, but with nouns containing pejorative connotations (7d,e). It belongs to another class of distribution. Moreover, this lexeme shares contexts with the leader word *homophobie* in (6b). We can then consider that *christianophobie* and *homophobie* have related meanings. *Microsoftophobie* (6e), already discussed in example (4), is found in contexts that are linked to both interpretations. This complex lexeme can thus be considered to be linked to both *homophobie* and *claustrophobie.* The complexity of the global lexical network takes shape both through the contexts shared by different lexemes and through the distance between the meaning of one lexeme and another.

In order to formalize the bottom-up method, the top-down method protocol was extended, by collecting a Web sample (once again comprising the first two pages of Google results for singular and plural forms), but this time for each *X-(o)phobie* lexeme in the database and not only for the ten most frequent ones. These contexts were then indexed according to their semantic features, using the Les Voisins resources. For example, Les Voisins de Wikipédia presents the context *accusation de* 'accusation' as preferentially used with the contexts *abus de pouvoir* 'abuse of power', *corruption* 'corruption' or *racisme* 'racism' which are attitudes that need to be fought against: the context *accusation de* 'accusation' will therefore be used to attribute the interpretation of 'hostility' to an *X-(o)phobie* lexeme. The context *atteindre de* 'be affected by' is only used with diseases such as *daltonisme* 'color blindness', *tuberculose* 'tuberculosis' or *maladie de Parkinson* 'Parkinson's disease'. This implies that if *X-(o)phobie* is used with *atteindre de,* it has to be interpreted with the meaning 'fear'. Some contexts, such as *se plaindre de* 'complain about', are presented in Les Voisins as used both with 'hostility' contexts (*injustice* 'injustice') and with 'fear' contexts (*douleur abdominale* 'abdominal pain'); we can therefore interpret *X-(o)phobie* lexemes used with these contexts either meaning 'fear' or 'hostility'.

In Table 7, some lexemes of the corpus are counted several times: some of them are found exclusively with a context interpreted either as 'fear', 'hostility', both or neither, while yet others are found in contexts belonging to several categories such as *accusation* and *atteindre.* In this case too, 203 lexemes, i.e., 38% of the database, are not covered by the analysis carried out with the bottom-up method. This can be attributed to two reasons: either an-*X (o)phobie* lexeme is only present in humorous lists of diseases (for example: "what are you afraid of when you suffer

Table 7: Results for the bottom-up method

	Exclusively	Non exclusively	Total
'fear'	86	62	148
'hostility'	44	100	144
'fear' or 'hostility'	15	50	65
other	47	101	148
none		203	

from *hippopotomonstrosesquipedaliophobie*?)[16] or, conversely, the lexeme is a frequent and fully lexicalized one and the first two pages of Google hits only contain dictionary entries. However, are these 203 undetected lexemes the same as the 238 lexemes that were not detected with the top-down method? Not necessarily. If we compare the two methods, we see that 215 lexemes are captured by both; 113 are captured only by the bottom-up method, and 87 only by the top-down one. 116 lexemes (22%) are not captured by either method. If the 62 lexemes found only in humorous lists of phobias (and which can be counted with the 'fear' interpretation) are excluded, that leaves only 54 lexemes that cannot be interpreted thanks to the distributional analysis, i.e., 10% of the corpus. These two methods are therefore complementary and show the interest of distributional analysis for morphological data.

3.3. Implications

In addition to providing a method for interpreting the meaning of complex lexemes, the analysis presented has highlighted an etymologic and stylistic difference between *X-(o)phobie* 'fear of *X*' and *X-(o)phobie* 'hostility toward *X*', which is correlated with the origin of the first constituent. Table 8 sums up the number of lexemes which display a majority of contexts compatible with either the 'fear' or the 'hostility' meaning, using the bottom-up method.

   Table 8 reveals a significant difference between *X-(o)phobie* 'fear' and *X-(o)phobie* 'hostility', when the origin of the first constituent is taken into account. While an *X-(o)phobie* lexeme carrying the meaning 'fear' prefers a classical first constituent (61.2%), an *X-(o)phobie* lexeme with the meaning

---

[16] The answer is: you are afraid of too long words! (http://tinyurl.com/q7o5jm2)

Table 8: Origin of the first constituent according to the meaning of the complex lexeme

	C1: Latin	C1: Ancient Greek	C1: French	
'fear'	3 (3.4%)	52 (57.8%)	35 (38.9%)	90 (100%)
'hostility'	3 (3.8%)	6 (7.6%)	70 (88.6%)	79 (100%)

'hostility' displays a preference for a native first constituent (only 11.4% of C1 are classical). This can be explained by the fact that when *X-(o)phobie* means 'fear', it denotes a disease and so the lexeme has to look learned and scientific, whereas when you want to name a particular kind of hostility towards a person, group or object, the target has to be easily understandable by the general public. Most often, the first constituent denotes the name of a social group (on the schema of *homophobie*, lexemes like *lesbophobie* (< *lesbienne* 'lesbian') or *transphobie* (< *trans(sexuel)* 'transsexual') were coined) or a population, an ethnic group (*américanophobie* (< *Américain* 'American'), *kabylophobie* (< *Kabyle* 'Kabyle')). Lasserre (2013) focused on the difference in *X* origins for *X-(o)logie* and *X-(o)phone* lexemes designating speakers of a language and observed a similar phenomenon, reported in Table 9.

Table 9: *X-(o)phone* and *X-(o)logie* according to the origin of *X*

	-(o)phone	-(o)logie
Classical bases	3 (1.32%)	396 (40.04%)
Native bases	224 (98.68%)	593 (59.96%)

An *X-(o)logie* lexeme, like *X-(o)phobie* names of diseases, has to 'sound scientific' and a classical first constituent is used to achieve this. In contrast, *X-(o)phone* speaker names, like *X-(o)phobie* hostility names, not only have to be more transparent, there are also not many classical bases available to designate people or ethnic groups.

The difference in interpretation for *X-(o)phobie* lexemes is therefore perceptible even in the selection of *X* for the construction of the lexemes.

4. Conclusion: homophony or polysemy?

As we have seen throughout this paper, the same sequence, *-(o)phobie*, can give birth to different interpretations. The question therefore arises whether we are dealing with one lexeme construction rule or two. To provide a tentative answer, let us look at the lexemes which yielded ambiguous results in our analysis: some *X-(o)phobie* lexemes, as we have seen for *Microsoftophobie* in (4), can be used with both the 'hostility' and the 'fear' meanings. Two different explanations can be proposed for this. First, the metaphorical use of a construction can be established thanks to its distribution.

(8) Et Bob est à Nankin en 1937 en train de faire <u>une crise de</u> **Japanophobie** <u>aigüe</u>.
     'And Bob is in Nanjing in 1937 having an acute attack of Japan-ophobia.'

In (8), hostility towards the Japanese is presented as a disease. Nevertheless, we are in the presence of the output of a single construction, used metaphorically: *japanophobie* always reflects hostility. This is not the case in (9).

(9) a. Mais je n'ai jamais été à l'aise en auto, je suis comme ça. [...] Je crois sincèrement <u>souffrir d'</u>**autophobie** depuis sa naissance.
         'But I have never felt comfortable in a car, it's just the way I am. [...] I really think that I have suffered from car-ophobia since his birth.'
     b. Si vous vous sentez concernés [...], par l'**autophobie** <u>ambiante</u> savamment entretenue par certains dirigeants et lobbys et par le ras le bol lié à la pression incessante qui règne sur nos vies de conducteur [...]
         'If you feel concerned [...] about the prevailing car-ophobia, adroitly maintained by some leaders and lobbies, and are fed up with the constant pressure on our driving lives [...]'

Here we have two different outputs: a disease which consists in being afraid of cars (9a), hostility towards cars (9b). There is no metaphor. These outputs are, once again, revealed thanks to the distribution of each occurrence. Furthermore, in (9a), as for all the *X-(o)phobie* lexemes meaning 'fear', *autophobie* can be paraphrased as "la phobie des autos", 'car phobia', whereas this is not possible in (9b). As we have seen in Table 1, the autonomous lexeme *phobie* carries only the meaning 'fear'. If (9a) can be linked to an autonomous lexeme while (9b) cannot, two different construction rules should apply.

However, formalizing the difference between a metaphorical use and two different outputs would require further investigation. What the present study shows is that it can hardly be claimed that a meaning is linked to

a lexeme independently of its contexts of use. The meaning of a complex lexeme, especially in the case of neologisms, is established each time it is used in a given context. Morphology constructs objects with a set of potential meanings and their precise meaning can only be specified in real contexts. That is why the meaning of a complex lexeme cannot be studied in isolation.

References

Amiot, Dany and Georgette Dal. 2007. Integrating neoclassical combining forms into a lexeme-based morphology. In G. Booij, L. Ducceschi, B. Fradin, E. Guevara, A. Ralli and S. Scalise (eds.) On-line proceedings of the Fifth Mediterranean Morphology Meeting (MMM5), Fréjus 15–18 September 2005. Bologna: University of Bologna. 323–336.

Bauer, Laurie. 1998. When is a sequence of two nouns a compound in English? English Language and Linguistics 2. 65–86.

Booij, Geert. 2010. Construction morphology. Oxford: Oxford University Press.

Cottez, Henri. 1980. Dictionnaire des structures du vocabulaire savant: éléments et modèles de formation. Paris: Le Robert.

Fradin, Bernard. 2000. Combining forms, blends and related phenomena. In U. Doleschal and A. M. Thornton (eds.) Extragrammatical and marginal morphology. München: Lincom Europa. 11–59.

Harris, Zellig S. 1954. Distributional structure. Word 10. 146–162.

Harris, Zellig S. 1968. Mathematical structures of language. London: John Wiley & Sons.

Hathout, Nabil. 2011. Une approche topologique de la construction des mots. In Roché et al. (2011, 251–317).

Hathout, Nabil, Fiammetta Namer, Marc Plénat and Ludovic Tanguy. 2009. La collecte et l'utilisation des données en morphologie. In B. Fradin, F. Kerleroux and M. Plénat (eds.) Aperçus de morphologie du français. Saint-Denis: Presses Universitaires de Vincennes. 267–287.

Iacobini, C. 2004. Composizione con elementi neoclassici. In M. Grossmann and F. Rainer (eds.) La formazione delle parole in italiano. Tübingen: Niemeyer. 69–95.

Lasserre, Marine. 2013. Neoclassical compounds and language registers. Journées d'études Toulousaines (JETOU), 16–17 mai 2013, Université de Toulouse 2-Le Mirail.

Lasserre, Marine and Fabio Montermini. 2012. Une approche distributionnelle pour l'analyse des composés néoclassiques. Verbum 34. 195–212.

Lasserre, Marine and Fabio Montermini. 2014a. How is the meaning of complex lexemes constructed? A study of neoclassical compounds. Italian Journal of Linguistics 26. 157–181.

Lasserre, Marine and Fabio Montermini. 2014b. Pour une typologie des lexèmes construits: entre composition, composition néoclassique et affixation. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer and S. Prévost (eds.) Actes du Congrès Mondial de Linguistique Française (CMLF) 2014, Berlin, 19–23 juillet 2014. SHS Web of Conferences, 8 (2014). 1797–1812

Lazaridou, Angeliki, Marco Marelli, Roberto Zamparelli and Marco Baroni. 2013. Compositionally derived representations of morphologically complex words in distributional semantics. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. East Stroudsburg, PA: ACL. 1517–1526.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. Italian Journal of Linguistics 20. 1–31.

Lüdeling, Anke, Tanja Schmid and Sawwas Kiokpasoglou. 2002. Neoclassical word formation in German. In G. Booij and J. van Marle (eds.) Yearbook of morphology 2001. Dordrecht: Kluwer. 253–283.

Prince, Alan and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical Report TR-2, Center for Cognitive Science, Rutgers University, New Brunswick, N.J. and Technical Report CU-CS-697-93, Department of Computer Science, University of Colorado, Boulder.

Rainer, Franz. 2003. Semantic fragmentation in word-formation: The case of Spanish *-azo*. In S. Rajendra and S. Starosta (eds.) Explorations in seamless morphology. London: Sage. 197–211.

Roché, Michel. 2003. L'interfixe est-il une unité morphologique? Silexicales 3. 169–178.

Roché, Michel. 2011. Quelle morphologie? In Roché et al. (2011, 15–40).

Roché, Michel, Gilles Boyé, Nabil Hathout, Stéphanie Lignon and Marc Plénat (eds.). 2011. Des unités morphologiques au lexique. Paris: Hermès-Lavoisier.

Sahlgren, Magnus. 2008. The distributional hypothesis. Italian Journal of Linguistics 20. 33–54.

Villoing, Florence. 2012. French compounds. Probus 24. 29–60.