Neoclassical compounds and language registers

Marine Lasserre[1]
(1) CLLE-ERSS & Université de Toulouse
marine.lasserre@univ-tlse2.fr

Résumé Les composés néoclassiques occupent une place à part dans la morphologie du français. Quand ils mettent en jeu des éléments finaux néoclassiques récurrents, ils sont à la frontière de la composition et de la dérivation. S'ils étaient au départ utilisés pour désigner des concepts scientifiques ou techniques, ils sont aujourd'hui pleinement utilisés dans la langue courante. Il s'agit ici de voir comment ces différents registres de langue interfèrent dans la construction non seulement formelle mais aussi sémantique des mots complexes comportant des éléments néoclassiques.

Abstract Neoclassical compounds are special in French morphology. With regard to their final elements, they are at the border of composition and derivation. If originally they were used to designate scientific or technical concepts, today they are commonly used in everyday language. In this paper, we will see how these different language registers come into play in the construction of these words containing neoclassical elements, both from a formal and from a semantic point of view.

Mots-clés : Morphologie, Composition néoclassique, Registres de langue, Corpus, Analogie, Sémantique, Grammaticalisation.

Keywords: Morphology, Neoclassical compounding, Language registers, Corpus, Analogy, Semantics, Grammaticalization.

1 Introduction

The purpose of this presentation is to study some French neoclassical compounds which apparently display the same behavior. These compounds are built with final neoclassical elements and enter in the construction of large series of lexemes. Consequently, we consider that the final neoclassical element is undergoing grammaticalization, and will eventually become a suffix.

In this article, we adopt the framework of lexeme-based morphology ((Boyé *et al.*, 2011) among others). In this framework, the basic unit of morphology is the lexeme, which is seen as a unit having three properties: a phonological form, a semantic content and a grammatical category.

After defining what is a neoclassical compound (1.1) and presenting our corpus (section 2), we will present two different case studies: first, a morphophonological study (section 3) which shows how neoclassical compounds from different registers are formally treated in the same way, and second, a morphosemantic analysis (section 4) showing how non lexicalized words produced by native speakers allow us to refine the semantic models built on the basis of lexicalized compounds.

1.1 What is a neoclassical compound?

Traditionally, native compounds are defined on the basis of a main property: they result from the combination of two (or more) native lexemes. Thus, we have something like [Lexeme + Lexeme]$_{Lexeme}$. In French, we find four productive patterns of native compounds (Villoing, 2012) : (1) [NN]$_N$, (2) [VN]$_{A/N}$, (3) [AN]$_A$ and (4) [AA]$_A$.

(1)	poisson-chat	'fish-cat'	*catfish*
(2)	ouvre-boîte	'open-can'	*can-opener*
(3)	rouge-brique	'red-brick'	*red like a brick*
(4)	aigre-doux	'sour-sweet'	*sweet-and-sour*

The particularity of Romance languages is that the head of the compound is on the left.

Neoclassical compounds are different (cf. (Bauer, 1998; Iacobini, 2004) among others). First, they imply at least one Greek or Latin constituent which is no longer autonomous in modern languages, here in French. Moreover, the order of the constituents is reversed as compared to Romance native compounds. (Villoing, 2012) found seven regular patterns of neoclassical compounding in French: (5) [NN]$_A$, (6) [NN]$_N$, (7) [AN]$_A$, (8) [AN]$_N$, (9) [NA]$_A$, (10) [AA]$_A$ and (11) [NV]$_{A/N}$.

(5)	androgyne	*androgynous*
(6)	vélodrome	*velodrome*
(7)	macrocéphale	*macrocephalic*
(8)	microorganisme	*microorganism*

(9) photosensible *light-sensitive*

(10) médico-social *medico-social*

(11) xénophobe *xenophobic*

In addition, neoclassical compounds differ from the native ones in the presence of a linking vowel between the two constituents, generally /o/ with Greek constituents and /i/ with Latin ones.

1.2 Problems in defining the status of neoclassical constituent

The neoclassical constituent can be defined through four properties (Amiot & Dal, 2007). First, these elements were lexemes in the source languages (Greek or Latin) but second, they have no syntactic realization in modern languages, since they can only appear as bound constituents. Third, they serve to construct the learned vocabulary of the language. Finally, they display a linking vowel between the two constituents in a particular phonological context. So, neoclassical constituents inherit from Greek or Latin lexemes a semantic content just like autonomous French lexemes. Yet, their non-autonomous nature makes them look like affixes. We argue here that neoclassical constituents do not constitute a homogenous set since the status of each of these elements may vary. A constituent like *xéno-* seems to be closer to modern lexemes since it has a stable lexical meaning ('foreigner') whereas a constituent like *-logue* tends to be closer to affixes, and can even be considered as an affix: it lost its original semantic content as it became very productive, as we will see below.

1.3 Different language registers

The third property stated by (Amiot & Dal, 2007) cannot be applied to all neoclassical constituents. If originally these constituents were used in specialized lexicons (medical, learned, technical vocabularies), this is no longer the case. In fact, the same constituent can be employed in learned language, e.g. *-logue* is primary used in the medical vocabulary (*endocrinologue*, 'endocrinologist'; *cardiologue*, 'cardiologist'), but the current language adopted it either to imitate the medical vocabulary (*chevillologue*, 'specialist of ankles') or to form the names of other kinds of specialists (*chaussettologue*, 'specialist of socks'). Most of the words constructed on the model of "pure" neoclassical words are used with an ironical sense.

2 The Corpus

In order to account for the differences between these language registers, we chose six French neoclassical elements which seem to have the same behavior (Table 1).

All these elements are inherited from Greek: *-phage* comes from a verb and *-crate*, *-cratie*, *-phone*, *-logue* and *-logie* from nouns.

Our corpus is built on two different resources. First, we looked at the neoclassical compounds contained in two dictionaries : *Le Grand Robert* (online version) and *Le Trésor de la Langue*

French neoclassical element	*-crate*	*-cratie*	*-phone*	*-logie*	*-logue*	*-phage*
English gloss	'-crat'	'-cracy'	'speaker of a language'[1]	'-logy'	'-logist'	'-phagous'
Original lexeme	κράτος		φωνή	λόγος		φαγειν

Table 1: Studied elements and their origins

Française informatisé (TLFi). Then, we looked for non lexicalized compounds in a corpus built on the Google Ngrams resource. Finally, the corpus was completed by occasional findings. Table 2 gives the total number of lexemes for each element.

	-crate	*-cratie*	*-phone*	*-logie*	*-logue*	*-phage*
Dictionaries	18	30	15	535	226	78
Google	146	230	212	455	562	92

Table 2: Number of entries in the corpus for each element

3 Morphophonological analysis

At first, we studied the compounds of our corpus from a formal point of view. In particular, three main constraints were observed that determine the output form of these complex lexemes (3.1): Fidelity to the base, Fidelity to the output form, which has to be, preferentially, /ɔkʁat/, /ɔkʁasi/, /ɔfɔn/, /ɔfaʒ/ and so on, and finally a prosodic constraint on the size of the base. The variation observed in the attested forms is the result of the interaction of these three constraints (3). What we can conclude about morphophonology is that the differences in register or the origin of the bases do not play any role in the formal construction of the neoclassical compounds of our corpus (3.3).

3.1 The constraints involved

The first constraint involved is a constraint of Fidelity. It stipulates that the first element of the compound has to be maximally intact. Here, we refer to the first element as the base, as we consider that the final elements involved are quasi suffixes. If we consider that the Fidelity-to-the-base constraint is respected when there is no truncation, no epenthesis, no haplology and no learned allomorphy, this constraint is not violated in 65% to 85% of the cases, depending on the different elements.

The second constraint relates to the output form, which has end in /ɔkʁat/, /ɔkʁasi/, /ɔfɔn/, /ɔlɔg/, /ɔlɔʒi/ or /ɔfaʒ/, according to the element involved. Obviously, this constraint has a diachronic justification. The presence of the vowel -o itself is part of the definitional properties of neoclassical compounds. In ancient Greek (and still in Modern Greek), the vowel -o marked the compositional character of a word. Hence, the output form of a neoclassical compound, whether built with modern or ancient material, necessary display this vowel. We see in fact that /o/ is present in more than 80% of the words in our corpus at the boundary between the first and

the second element. In the remaining cases, neoclassical compounds mainly display a different vowel (mostly /i/ or /a/) and in a large minority of cases a consonant (mostly liquids /l/ or /ʁ/, which are close to vowels in the universal scale of sonority).

The third constraint that comes into play concerns the size of the complex lexeme. In the light of what we have studied in our corpus, we can infer that words constructed with the neoclassical elements *-crate*, *-cratie* and *-phage* are preferentially preceded by two syllables and *-phone*, *-logie* and *-logue* by three syllables.

3.2 The interaction of constraints

In the OT framework (cf. (Prince & Smolensky, 2008); for an application to morphology cf., among others, (Plenat, 1996)), constraints are hierarchically ordered and can be violated. Here we do not order them in a very precise hierarchy, but we can nevertheless observe their interaction. Thus, pédagocratie ('pedagogue' + *-cratie*) does not respect either the Fidelity to the base constraint or the size constraint. However, the truncation of the base permits respecting Fidelity to the output constraint, which states that the output has to end in /ɔkʁasi/.

On the other hand, a lexeme like *médiacratie* ('media' + *-cratie*) respects the Fidelity to the base constraint and violates the Output constraint, maybe to avoid homophony with *médiocratie* (*medio-* for 'middle class' + *-cratie*).

Table 3 shows how the same base may give different outputs words depending on the constraints which are respected or violated[2].

corporation 'ship' + *-cratie*	(i) Fidelity to the base	(ii) Fidelity to the output /ɔkʁasi/	(iii) Size of the base
corporatocratie (10355)	+	+	– –
corporacratie (1830)	–	–	–
corpocratie (1308)	–	+	+

Table 3: Different outputs words for a same base

Let us now look at the behavior of bases containing /o/, once again based on examples in *-cratie*. If the vowel /o/ occurs in the first syllable a base is never truncated after it (e.g. *politocratie* ← 'politician' + *-cratie*). If the /o/ occurs in the second or in the third syllable, in a large majority of cases the base is truncated after it (e.g. *consocratie* ← *consommateur* 'consumer' + *-cratie*; *tortiocratie* ← *tortionnaire* 'torturer' + *-cratie*; *pédagocratie* ← *pédagogue* 'pedagogue' + *-cratie*).

If the base ends in a consonant, the vowel /o/ allows a resyllabification and hence, respect for the Size constraint by one-syllable bases (e.g. *fricocratie* ← *fric* 'dough' + *-cratie*).

[2]In Table 3, + indicates that a constraint is respected and – that it is violated. Concerning constraint (iii), – indicates each syllable which violates the two-syllables preference. In brackets we indicate the number of occurrences found on Google (research of 1st March 2012)

3.3 Different registers, same constraints

The Fidelity to the output constraint is automatically respected with bases of Greek origin, which systematically end in /ɔ/. However, when a suppletive Greek stem is not available, French native stems are subject to the same constraint. For instance, in our corpus very few classical bases are available for words in *-phone*, since their bases are language names, often containing sequences which are uncommon in French. We can compare them with the words in *-logie*, for which large number of bases of classic origin[3] are available. Table 4 shows the difference between the two subcorpora:

	-phone	*-logie*
Classical bases	3 (1.32%)	396 (40.04%)
French native bases	224 (98.68%)	593 (59.96%)

Table 4: Words in *-phone* and *-logie* with classical and French native bases

In spite of these differences, words constructed with the two elements display /ɔ/ at the boundary between the first and the second element in a comparable proportion[4]:

	-phone	*-logie*
Output containing /ɔ/ between the two elements	81.94%	94.40%
Of which, with /ɔ/ not present in the base	90.86%	62.30%

Table 5: Percentage of words in *-phone* and *-logie* containing /ɔ/ between the two elements

So, even if the origin of the bases varies, the outputs of neoclassical compounds respect the same formal constraints. We observe the same phenomenon when we look at the size of the bases, whether they are of classic origin or not.

In what follows, we will see that the homogeneity observed in morphophonology is not found when semantics is concerned.

4 Morphosemantic analysis

4.1 Lexical constraints

The neoclassical elements considered here contribute to lexical enrichment. Each new word is built in connection with the other lexemes in the lexicon. This connection may take different ways. For instance, some lexemes may acquire the status of "leader words" (Roché, 2011). In general, a "leader word" corresponds to the first (or one of the first) derivate which is attested, which becomes the source of a word formation pattern by virtue of its frequency or saliency. Once the pattern is established in the language, new words which are coined can be positioned

[3]We choose to consider bases as French ones when they actually exist in French, even if they originally come from Greek or Latin (e.g. "théâtre" in *théâtrologie*).

[4]We consider here that the segment -o- is present in the base when the base actually contains this segment (e.g. “bobo” in “bobologie”) or when the base is a Greek one (e.g. “psycho-” in “psychologie”).

with respect to the original model. For instance, we may consider that, as far as meaning is concerned, leader words function as semantic catalysts.

Concerning *-crate* and *-cratie*, two leader words seem obvious : *démocratie* (‘democracy’), first documented in 1370, may have generated the meaning "power of X", while *aristocratie* (‘aristocracy’), also first documented in 1370, the meaning "self-valorizing closed social group"; this meaning has been documented since the end of the eighteenth century. On the other hand, for words in *-phage*, it is more difficult to find a clear leader word (or a set of leader words) on a diachronic or on a logical basis. However, we can also observe the emergence of different semantic patterns. The main semantic patterns of *-phage* words are given below:

(12) ‘who eats X’ e.g. *anthropophage* (‘anthropophagous’)

(13) ‘who/which consumes X’ e.g. *chronophage* (‘time wasting’)

(14) ‘who loves X in a pathological manner’ e.g. *téléphage* (‘TV addict’)

(15) ‘who hates X / who annihilates X’ e.g. *turcophage* (‘who hates Turks’)

If we look at the words constructed with *-logue* or *-logie*, no clear leader word seems to emerge. In this case, lexical organization is semantically "distributed": the lexemes in our corpus display a range of different but connected meanings without any clear attracting pole emerging. In all these cases, new lexemes are constructed by analogy with the already existing lexicon.

What neoclassical compounds tell us about lexical construction is that although lexicalized compounds give us a global vision of the semantic models involved, words productively coined by speakers allow us to refine this models.

4.2 How non lexicalized compounds contribute to semantic modelization

If we look at compounds containing the element *-cratie*, we see that *démocratie*, as we said before, functions as a leader word for the semantic reading ‘power of’. The description of this model can be refined looking at non lexicalized words. In fact, speakers coined such words as *corruptocratie* (‘power exercised by means of corruption’), *fricocratie* (‘power exercised in order to get money’) or *ruecratie* (‘power exercised in the streets’). Along with this original model, another semantic pole emerged, with *aristocratie* as its leader word. The original notion of power is absent from the semantic reading of words in this class, which refer to ‘self-valorizing closed social groups’. Several new words were coined in every-day language according to this model, such as *cathocratie* (‘Catholics as a self-valorizing social group’). On their turn, these two main poles generated some sub-models. Thus, we found such words as *Sarkocratie* (‘a state in which Sarkozy has the power’), modeled on the *démocratie* type, or *footocratie* (‘omnipresence of football’) modeled on *aristocratie* type.

As far as words in *-phage* in our corpus are concerned, the original meaning is ‘who eats X’, as illustrated by words like *anthropophage*. Two antithetical models were generated by metaphor from this original sense, mainly in common language. On the one hand, we have the reading ‘who hates X / who annihilates X’ exemplified by *turcophage* (see (15)). On the other hand, we have the reading ‘who/which consumes (a lot of) X’ illustrated by chronophage (see (13)). This model, in turn, generated a fourth reading ‘who loves X in a pathological manner’. Such

lexemes as *cinéphage* (‘movie addict’) or *téléphage* (‘TV addict’) exemplify this pattern. From a semantic point of view, words in *-phone* in our corpus are not so interesting, as we limited our observation to words indicating the speaker of a language, ignoring, for the moment, such models as ‘music instrument using something’ (*xylophone* ‘xylophone’) or ‘measuring device’ (*tensiophone* ‘blood pressure meter’). However, common language also tends to refine the first model, to a lesser extent. Thus, a lexeme like *belgophone* can refer to someone who speaks Belgian (the variety of French spoken in Belgium) or, in some contexts, a resident of Belgium.

We saw in 4.1 that words in *-logie* cannot be considered to be constructed by analogy with any clear leader word. Rather, their semantic organization has to be considered "distributed". This kind of semantic organization shows the same type of phenomena that we observed with words having a clear leader word. In this case, it is not the range of semantic readings that is modified in newly constructed words but the number of possible bases. The ironical dimension plays a crucial role here. Thus, the ‘study of a disease’ like *cancérologie* ‘research on cancer’, *paludologie* ‘research on paludism’ becomes the ‘study of a false or a minor disease’: *ampoulologie* ‘the study of blisters’. The ‘study of a part of the body’ (*dermatologie*, *cardiologie*) is turned into the ‘study of a part of the body which is normally not studied’: *moustachologie* ‘the study of moustaches’. In addition, a lexicalized compound may be the source of a new sub-paradigm. We can observe it with *œnologie* ‘oenology’, which generated *vinologie* (same meaning but base of different origin), *biérologie* ‘the study of beer’ or *cidrologie* ‘the study of cider’. Another example is *démonologie* ‘demonology’ which gave birth to *elficologie* ‘the study of elves’ or *vampirologie* ‘the study of vampires’.

All these examples show that the observation of register variation allows a more detailed analysis of semantic variation. The lexicalized occurrences in our corpus gave us a global vision of the semantic behavior of each neoclassical element: we saw for instance that *-logue* and *-logie* can receive different analyses since we cannot sort out clear leader word or attracting poles, but only identify a general range of connected meanings. However, regardless of the behavior of each individual element, nonlexicalized compounds contribute to the lexical dynamics of the language, contributing to refine the original semantic models. Today neoclassical elements are not confined to the scientific, technical or learned registers of the language; their use has extended to common language. Below we will see to what extent these refined semantic models are productive.

4.3 Semantic models and productivity

Let us look at the semantic models attested for words containing the element *-phage*. We distinguished four different semantic types, exemplified, respectively, by *anthropophage*, *chronophage*, *téléphage* and *turcophage* (see above). Table 6 (see end of the paper) gives the number of occurrences belonging to each type and the distribution of lexicalized / non lexicalized compounds in our corpus. Note that some lexemes have been classified into different types. For instance, *bibliophage* can refer to a paper-eating insect or to someone who reads a lot. Very few words cannot be easily classified in any of the four types: *mélophage* (‘parasite living in the fleece’), *mallophage* (‘living on mammals and biting without stinging’) and *métrophage* (‘who regularly travels by subway’). It is clear from Table 6 that some semantic types are more productive than others. Thus, the *anthropophage* type is very predominant among the lexicalized compounds of our corpus (84.81%) whereas it represents less than a half (47.92%) of nonlexicalized compounds. The *chronophage* and *téléphage* types show the contrary trend. Most of the occurrences

of these types are not found in dictionaries (only 13.33% et 6.90%, respectively, of the occurrences of these types are present in the TLFi and/or in Le Grand Robert). If we consider that nonlexicalized compounds tend to be the most recent, we can suppose that these two semantic types are more recent than the first one. This corroborates the fact that it is nonlexicalized words that refine semantic models and generate new sub-models.

In addition, we can observe that some semantic patterns are not productive anymore. It is the case for the 'discourse characterized by' interpretation of *-logie* (e.g. *dyslogie* 'language disorder', *tachylogie* 'speech disorder characterized by the accelerating pace of emission of words') which is not found in new lexemes while the interpretation 'set / collection of' seems to occur in coined words (*critérologie* 'set of criteria', *discologie* 'set of disks', *bouletologie* 'set of drags').

5 Conclusion

What does language register variation tell us in the study of neoclassical compounds? First, formally, even compounds that have been coined recently with native material (not only from Greek or Latin bases) and are not lexicalized respect the same constraints than compounds using Greek or Latin bases, which have been in the lexicon for a long time, used in scientific, scholar domains. Second, from a semantic point of view, nonlexicalized compounds allow us to refine the global analysis elaborated on the basis of lexicalized compounds: they provide shades which may be then analyzed in sub-models. Some semantic patterns lost their productivity, while some others got a new life in the common language. In this respect, these elements are representative of the lexical dynamics of the language.

Acknowledgements

I would like to thank Fabio Montermini for reviewing this article and Nabil Hathout for helping me in collecting data with Google Ngram.

References

AMIOT D. & DAL G. (2007). Integrating neoclassical combining forms into a lexeme-based morphology. *G. Booij et al*, p. 323–336.

BAUER L. (1998). Is there a class of neoclassical compounds, and if so is it productive? *Linguistics*, **36**(3), 403–422.

BOYÉ G., HATHOUT N., LIGNON S., PLÉNAT M. & ROCHÉ M. (2011). *Des unités morphologiques au lexique*. Hermes-Lavoisier.

IACOBINI C. (2004). Composizione con elementi neoclassici. *M. Grossmann/F. Rainer (a c. di) La formazione delle parole in italiano, Tübingen, Niemeyer*.

PLENAT M. (1996). De l'interaction des contraintes: une étude de cas. *Current Trends in Phonology: Models and Methods, ESRI, University of Salford, Salford*, p. 585–615.

PRINCE A. & SMOLENSKY P. (2008). *Optimality Theory: Constraint Interaction in Generative Grammar*. Wiley.

ROCHÉ M. (2011). Quel traitement unifié pour les dérivations en-isme et en-iste. *M. Roché, G. Boyé, N. Hathout, S. Lignon, et M. Plénat, Des unités morphologiques au lexique. Hermes-Lavoisier*.

VILLOING F. (2012). French compounds. *Probus, International Journal of Latin and Romance Linguistics. Mouton de Gruyter*, p. 26–60.

	TLFi/Grand Robert			Google		
	Number of occurrences	Percentage with respect of the total number of occurrences of the resource	Percentage with respect of the semantic type	Number of occurrences	Percentage with respect of the total number of occurrences of the resource	Percentage with respect of the semantic type
Type 1 ”anthropophage”	67	84.81%	59.29%	46	47.92%	40.71%
Type 2 ”chronophage”	2	2.53%	13.33%	13	13.54%	86.67%
Type 3 ”téléphage”	2	2.53%	6.90%	27	28.13%	93.10%
Type 4 ”turcophage”	6	7.59%	40%	9	9.38%	60%
Others	2	2.53%	66.67%	1	1.04%	33.33%

Table 6: Distribution of the four semantic types of -phage derivates in dictionaries and Google