Neo-Classical Word Formation in WM Electronic Dictionaries

Evanthia Petropoulou* & Pius ten Hacken×

*Englisches Seminar
Universität Basel
Nadelberg 6
CH-4051 Basel
evanthia.petropoulou@unibas.ch

×Abt. Geisteswiss. Informatik
Universität Basel
Petersgraben 51
CH-4051 Basel
pius.tenhacken@unibas.ch

Abstract

The lexicon of several modern languages such as English, Italian, French, and German contains words which are composed of components corresponding to Ancient Greek and Latin words, but have not been borrowed from these ancient languages. The coverage of this part of the lexicon involves the recognition of the basic units, the description of the word formation processes, and the analysis of existing words in terms of these units and processes. The Word Manager system for reusable morphological dictionaries provides a formalism in which neo-classical word formation can be described in a very natural way. It is assumed that neo-classical formatives do not have a normal syntactic category, so that they have to go through word formation processes which turn them into lexemes of a regular class. By using the same principles and guidelines in the coverage of English and Italian, similarities and differences between these languages stand out.

1 Introduction

In the vocabulary of many European languages there are words of Greek (i.e. Ancient Greek) and Latin origin. Often these words occur in several modern languages with an almost identical form and meaning. Thus, corresponding to Ancient Greek θεολογια, we find English *theology*, German *Theologie*, Italian *teologia*, etc. Given the common cultural heritage of the people speaking them, it is no surprise to find that these languages share a stock of words taken from a language which is, or at some point was, consideredthe language representing important aspects of this culture. This phenomenon is not only attested for Greek but also for French and more recently for English. An originally French word such as *garage* is attested in English, Italian, German, etc. with the same meaning and form. Similarly, English *feedback* is found in Italian, German, French, etc.

An interesting property of such words with Greek origin is their morphological complexity. The components of *theology* are also found in *theosophy* and *morphology*, both of which do not occur in Ancient Greek. They were coined in the 19th century, with corresponding words appearing in German, French, Italian, etc. As the form and meaning of these words are largely predictable on the basis of their parts, the formation process can be described in terms of word formation rules. This approach to neo-classical word formation requires that the constituent parts are described in the lexicon.

In this paper we describe an approach to the coverage of both the elements involved in neo-classical word formation and the rules for combining them. As a starting point we take the Word Manager system for reusable morphological dictionaries.

2 Word Manager

Word Manager (WM) is a system for the specification, use, and maintenance of electronic dictionaries, described by Ten Hacken & Domenig [1996]. A WM lexicon is built up around a description of the morphological system of a language, including both inflectional and word formation rules. Lexemes are entered by assigning them to an inflection class and, where applicable, to the word formation process which formed them. The resulting dictionaries are intended to be used as a basis for lexical tools in a large variety of NLP-contexts. In the context of the project "Word Formation as a Structuring Device of the English and Italian Lexicons: A Large-Scale Exploration", WM lexicons for English and Italian are developed conforming to a common set of lexicographic specifications.

For the coverage of neo-classical word formation this means that first the underlying rule system has to be described in terms of the WM formalism. As an initial hypothesis we assumed that neo-classical word formation is a separate phenomenon, with restricted interaction with other word formation processes.In a second stage the lexicon entries are classified in terms of the available rules. The coverage of the internal structure and the formation of new words require the representation of the neo-classical formatives as entities.

3 Neo-Classical Formatives

The basis for the description of neo-classical word formation is the assumption that there is a class of neo-classical formatives (NCFs). NCFs are formatives available for the formation of new lexemes but they are not lexemes themselves, which implies that they are bound morphemes. Examples are *morpho*, *anthropo*, *hydro* in English and *morfo*, *antropo*, *idro* in Italian. Their bound status is explained by the absence of a regular syntactic categorywhich is encoded in practice by means of a special syntactic category `(Cat NCF)`. There are no syntactic rules that refer to this category, but only morphological rules. In a system such that described by LŸdeling et al. [2001] the bound status of these formatives has to be stipulated.

Following Bauer [1998] we assume that the basic form of NCFs includes the final *-o*. According to the OED, in Ancient Greek combinations (and their adaptations and imitations in Latin), the combining stem usually ended in *-o* , as thematic vowel or its representative, or as an addition to a consonant stem. In modern Latin and English, it has come to be the usual connecting vowel in scientific terms in general and it is affixed not only to terms of Greek origin, but also to those derived from Latin. Typically, NCFs such as *morpho* appear as first part of a compound with the final *-o*, e.g. *morphology*, but in final position the *-o* is lost before the suffix, e.g. *anthropomorphic*. The loss of the final vowel before the initial vowel of a suffix is a process widely attested in English and Italian, whereas the insertion of an *-o* would be unique. For modelling such morphological processes, WM has Spelling Rules (SRules), which derive the form without the final-*o* before a suffix.

It is a question for the lexicographer to determine for a particular element whether it is a formative or not and whether it is a lexeme or an NCF. The following properties can be taken as general indications for NCF-hood:

- NCFs have a meaning and form of their own. The meaning and form of an NCF is based on the meaning of a formally similar word in Ancient Greek and/or Latin.
- NCFs can be used in word formation appearing as left- or right-hand elements in a compound.
- NCFs select for specific suffixes, such as *-y*, *-ic*, *-ous*, *-ist*, *-ism*, *-itis*, *-ia*, etc. in English.
- NCFs do not normally combine with native lexemes.

The idea is that after the borrowing of a number of Greek words involving a certain set of basic components, a reanalysis of these items has occurred in the lexicon of English and Italian. This historical analysis is supported by the data given in the OED. Until the 18th century, *morpho* is attested in English only as part of words that were borrowed from Greek as entire words, e.g. *metamorphosis*, *anthropomorphous*. In the 19th century, new words appear, e.g. *morphology*, which do not have a parallel in Ancient Greek. The appearance of *morpho* in words such as *morphology* shows that it has become a formative in English. Conversely, for Greek words that have no role in English word formation, there is no reason to consider them as formatives in English.

By way of example, the criteria apply to Italian *antropomorfico* in the following way, showing that it is a complex entry based on the NCFs *antropo* and *morfo*:

- *antropo* and *morfo* can be recognized as forms associated with a constant meaning. Their meanings are related to the Greek words ανθρωπος and μορφη respectively.
- *antropo* and *morfo* can be used in the creation of new lexemes, appearing in the left- or right-hand side of compounds and selecting for specific suffixes, e.g. *filantropismo*, *antropologico*, *polimorfia*, *morfologico*.
- *antropo* and *morfo* do not readily combine with native lexemes.

4 Word Formation Rules

As mentioned above, NCFs do not have a real syntactic category, but only a feature `(Cat NCF)`, which is not referred to in syntax. In order to have a syntactic distribution at all, they have to undergo morphological processes that assign a lexical category to their result. (The analytically simplest process is suffixation. Thus, the English suffix *-ic* may attach to NCFs and assigns the syntactic category of adjective to the result, as in *anthropic*.)

In most cases, neo-classical lexemes are not combinations of a simple NCF with a suffix, but rather consist of two NCFs followed by a suffix. In the case of such compound lexemes that contain NCFs, it is first necessary to specify the order of the processes of compounding and suffixation. Thus for the lexeme *anthropology* there are three possible analyses, shown in (1).

(1) a. [[$_X$ anthropo logo] -y]
b. [anthropo [$_Y$ logo -y]]
c. [anthropo logo -y]

Both X in (1a) and Y in (1b) are constituents which do not correspond to attested words. Nevertheless these analyses are theoretically more attractive than the ternary branching (1c). Comparing the nature of X and Y, we observe that the hypothetical *-logy* in (1b) would be an entity of a new kind. Being the combination of an NCF with a suffix, it is complex, fictional, and able to assign a syntactic category. It is difficult to explain why *-y* would have the latter property but *-logy* is not a noun. No such problems arise in structure (1a). Here *anthropologo* has all the properties we know of simple NCFs except that it is complex. If simple NCFs have no syntactic category, we expect that compound NCFs do not have one either. This expectation is borne out in the sense that any of a range of suffixes can turn *anthropologo* into a word, cf. *anthropology,anthropological, anthropologist.* The semantic relationships among these words also correspond to what can be expected if *anthropologo* is first formed as a complex NCF with a specific meaning.

Apart from compound NCFs, we also find complex NCFs involving a prefix. An example is *polytheism.* In principle, we could imagine the structures as in (2), which closely correspond to the ones in (1).

(2) a. [[$_X$ poly theo] -ism]
b. [poly [$_Y$ theo -ism]]
c. [poly theo -ism]

The arguments for the preference of (2a) follow the same pattern as in the discussion of (1). There is one typical difference, however, in the sense that Y in (2b) has the same form as an existing word, *theism.* We reject the idea that this constitutes an argument in favour of (2b), because the most prominent meaning of *theism* shows a specialization which does not fit in with the meaning of *polytheism.* In fact, *theism* refers to a quite specific view of God, in contrast to *deism* as introduced by Voltaire. The component *theo* as it is used in *polytheism,* however, refers to the general meaning of 'god', which can be modified by *poly* without contradiction.

The distinction between prefixes and NCF-stems is important in the context of WM, because prefixes are encoded as elements of the morphological rule system whereas stems areentered by the lexicographer. There are a number of criteria that can be brought to bear on this distinction. When a formative can be either the first or the last element in a compound NCF, it must be an NCF itself [Scalise 1984]. This applies to *filo* in Italian *filosofico* and *anglofilo.* If nothing else can be found, a classification along the lines of Greek syntactic categories is used as a heuristic criterion, Greek nouns corresponding to NCFs, other categories in Greek corresponding to prefixes. There are only few cases where this generalization results in unsatisfactory classification. Interestingly, these cases are relatively harmless, because the very reason why they are felt as stems, their specific meaning, at the same time has the typical side effect of a relatively small range of use.

The other process for turning NCFs into lexemes is conversion, i.e. derivation without an affix. As long as the base for conversion is a compound NCF, there is no other analysis. For example, English *photograph* is formed from the NCF *photographo,* a compound of *photo* and *grapho,* by deleting the final *-o.* In Italian, the situation is slightly more complicated because of the inflectional system. In the Italian noun *antropologo* ('anthropologist') the final *-o* is the singular ending of the noun, not the thematic vowel of the NCF *logo.*

Therefore, it is derived from the NCF *antropologo* by deleting the thematic character and assigning it to the *o/i* noun class.

In the case of simple NCFs, conversion competes with borrowing. For instance, the lexeme *crypt* ('an underground room'), which is a borrowing from Latin through French, can be analysed asa conversion of the NCF *crypto* ('hidden, concealed, secret'), since it is both morphologically and semantically related to it, carrying only a special use of its meaning. The adjective *cryptic* ('hidden, mysterious')with its much more general meaning cannot be based on the noun *crypt* but should rather be related to the NCF *crypto* by suffixation.

The strength of the phenomenon of neo-classical compounding is shown by the analogous formation of words such as *filmography* in which, words of clearly non-neo-classical origin combine with neo-classical elements. The case of *filmography* is analogous to that of *bibliography*, which can be analysed into the complex NCF *bibliographo*, consisting of the NCFs *biblio* and *grapho*, and the suffix *-y*. For the treatment of *filmography* in a parallel way, a formative *filmo* has to be hypothesized, which can attach to *grapho* and create the NCF *filmographo*. A conversion rule turns the regular noun *film* into an NCF, by adding the *-o*.

A special problem arises in those cases where both a borrowed word, e.g. *larynx*, and the corresponding NCF, in this case *laryngo* as in *laryngoscopy*, occur in the language. They correspond to two formsof the same paradigm in Greek, the nominative λαρυγξ and the combining form λαρυγγο, and their semantic relation is much tighter than in the case of *crypt* and *crypto*. This relationship is expressed by taking *larynx* as the base form and applying the conversion rule for nouns to NCFs to produce *laryngo*.

5 Implementation

Neo-classical word formation is implemented as a set of word formation rules in WM, organized in a parallel fashion for English and Italian. There are three parts: NCF-formation, suffixation and conversion. NCF-formation includes one rule each for compounding, prefixation, and conversion from lexemes to NCFs. The output of each of these processes is a new NCF. Suffixation and conversion, on the other hand, as it has been described above, produce lexemes. In suffixation, it is specified for each suffix how the resulting lexeme is inflected by assigning it to a particular inflection class. Therefore, separate rules had to be written for each syntactic category and inflection class. For English there is one rule each for nouns, adjectives, and verbs. In Italian, different rules are necessary for the different inflection classes of nouns and adjectives, resulting in five word formation rules. For conversion, a similar specification is required.

At the time of writing, the lexicon databases are not large enough to draw many general conclusions about the difference between English and Italian in the domain of neo-classical word formation. The problem is that in order to describe a complex lexeme in WM, the simple lexeme(s) on which it is based have to be entered first. Therefore, in a database of approximately 40,000 entries, there are disproportionately many simple lexemes. Nevertheless a number of interesting similarities and differences have emerged so far. First, in approximately 2000 entries in each lexicon resulting from neo-classical word formation

processes, NCF formation covers $^1/_3$ of all entries, while the restare lexemes resulting from suffixation and conversion. An interesting difference between the two languages is that conversion of NCFs to lexemes is much more frequent in Italian, consisting of $^1/_4$ of the resulting lexemes, compared to only $^1/_{10}$ in English. A reason is that Italian has also conversion to adjectives, which English lacks, and more than twice as many converted nouns. No doubt this is due to the formal similarity of NCFs with their final *-o* to the large class of singular masculine nouns and adjectives in *-o*. Second, in the case of suffixation, numbers are comparable for both languages, except for adjectives, where English has twice as many as Italian. This compensates for the lack of conversion to adjectives in English. Third, Italian seems to have a generally higher number of NCFs than English. Although it is not clear yet how significant this difference is, it can be expected as a result of the different relationship to Latin, which often served as a mediator between Ancient Greek and modern languages. As opposed to English, for Italian Latin is not only a learned language taken for borrowing, but also constitutes the earlier stage of the language.

6 Conclusion

The impact of neo-classical word formation on the lexicons of English and Italian is significant. This impact is not reduced by the fact that these words are formed by a conscious, creative process, in particular for scientific terminology. The formalism available in Word Manager makes a systematic coverage of the formatives and processes involved possible. These processes involve compounding and prefixation producing bound stems of `(Cat NCF)` and suffixation and conversion transforming NCFs into lexemes. Starting from the assumption that neo-classical word formation is a separate phenomenon in word formation turned out to be remarkably unproblematic.

By using the same lexicographic specifications for the recognition and analysis of neo-classical word formation in English and Italian, similarities and differences between these languages stand out. The most salient difference is the frequency of conversion in Italian, in part compensated for by more frequent suffixation in English. More precise conclusions in this area can be drawn when the lexicon database is larger.

Acknowledgement

The work described here was funded by the Swiss National Science Foundation under grant nr. 1214-058936.99.

Bibliography

[Bauer 1998] Bauer, Laurie, 1998. Is there a class of neoclassical compounds and if so is it productive?, in: *Linguistics, 36*, Mouton de Gruyter, Berlin.

[Ten Hacken & Domenig 1996] ten Hacken, Pius & M. Domenig, 1996. Reusable Dictionaries for NLP: The Word Manager Approach, in: *Lexicology, 2*, De Gruyter, Berlin.

[LŸdeling et al. 2001] LŸdeling, Anke, T. Schmid & S. Kiokpasoglou, 2001. Neoclassical word formation in German, in: *The Yearbook of Morphology 2001*, Kluwer Academic, Dordrecht.

[Scalise 1984] Scalise, Sergio, 1984. *Generative Morphology*. Foris Publications, Dordrecht.