Neoclassical Compound Alignments from Comparable Corpora

Rima Harastani, Béatrice Daille, Emmanuel Morin

Neoclassical Compound Alignments from Comparable Corpora

Rima Harastani, Béatrice Daille, and Emmanuel Morin

University of Nantes
LINA, 2 Rue de la Houssinière
BP 92208, 44322 Nantes, France
{Rima.Harastani, Beatrice.Daille, Emmanuel.Morin}@univ-nantes.fr

Abstract. The paper deals with the automatic compilation of bilingual dictionary from specialized comparable corpora. We concentrate on a method to automatically extract and to align neoclassical compounds in two languages from comparable corpora. In order to do this, we assume that neoclassical compounds translate compositionally to neoclassical compounds from one language to another. The method covers the two main forms of neoclassical compounds and is split into three steps: extraction, generation, and selection. Our program takes as input a list of aligned neoclassical elements and a bilingual dictionary in two languages. We also align neoclassical compounds by a pivot language approach depending on the hypothesis that the neoclassical element remains stable in meaning across languages. We experiment with four languages: English, French, German, and Spanish using corpora in the domain of renewable energy; we obtain a precision of 96%.

1 Introduction

Describing new concepts usually requires creating new terms. Neoclassical word-formation is one process used by Romance and Germanic languages (among others) in order to produce domain-specific terms. It combines some elements borrowed from Greek or Latin, called neoclassical elements, to create neoclassical compounds. For example, combining the neoclassical elements *hydro* and *logy* leads to the neoclassical compound *hydrology*. New neoclassical elements could be borrowed when needed to form new terms. The productivity of neoclassical compounds, especially in scientific domains such as medicine, makes their translation difficult since many of them are unlikely to be found in bilingual dictionaries.

Many neoclassical compounds possess a compositional property (the meaning of the whole can be restored from the meaning of the parts) [1]. This property is as well valid for many complex terms (terms that consist of more than one component). For example, the complex term *washing machine* is in fact a machine designed to wash. Thus, some approaches have been proposed to translate complex terms depending on this compositional property [2] [3] [4] [5]. They translate a complex term by translating each of its components individually

using a bilingual dictionary. Then, they combine these individual translations according to some predefined templates to produce the final translation of the complex term. For example, a complex term that is of the form: [Adjective Noun] in English (e.g. comparable corpora), could be translated by a term that is of the form: [Noun Adjective] in French (e.g. corpus comparables). On the other hand, unlike components that compose complex terms, equivalents of neoclassical elements are not expected to be found in monolingual or bilingual dictionaries. For that reason, previous works depend on other resources than bilingual dictionaries to deal with neoclassical elements. For example, [6] use a pivot language (e.g. Japanese) in order to automatically acquire the semantical meaning of neoclassical elements. They suppose that neoclassical compounds are translated in Japanese to terms that consist of simple words. In this way, they align each neoclassical element with its equivalent simple word in Japanese. [7] build multilingual lists of neoclassical elements with their meanings and the relations between them. They define this list for each language in order to morphosyntactically analyze neoclassical compounds. [8] focuses on translating Italian constructed neologisms by prefixation into French. He relies on lexical resources and a set of bilingual lexeme formation rules in order to detect constructed neologisms and to generate their translations. Some of the differences between our work and his, is that [8] does not differentiate between native prefixes and neoclassical elements, he deals only wih prefixation while we treat other forms of neoligisms and experiment with four languages.

We propose a method to automatically extract and align neoclassical compounds from bilingual comparable corpora. Indeed, comparable corpora (collection of multilingual texts that belong to the same domain) have been successfully used for terminology alignment by many approaches [2] [3] for their advantages over parallel corpora (collection of multilingual texts that are translations of each other), such as their availability [9]. Identifying translations of neoclassical compounds can help in enriching bilingual dictionaries used by several applications such as Machine Translation tools (MT tools) and Computer-Aided Translation tools (CAT tools). We suppose that most neoclassical compounds in a source language translate compositionally to neoclassical compounds in a target language. We use a predefined aligned list of neoclassical elements between the two languages, and we define the template that will combine the translations of the individual parts as being the original Greco-Latin template in forming terms [10].

The paper is organized as follows. In section 2, we give a breif introduction to neoclassical compounds. Then, we explain the alignment method in section 3. We present an evaluation of this method in section 4. Finally, we conclude in section 5.

2 Neoclassical compounds

We define neoclassical compounds as single-word terms consisting of at least one neoclassical element. Neoclassical elements or *combining forms* are elements

that are borrowed from Greek and Latin languages (e.g. *patho*, *bio*, *logy*, etc.). These elements are not considered as lexical units as they cannot play the role of independent words in a language syntax, i.e., they are always seen in the combined form with other elements (e.g. *biology*) [10] [11]. Each language may assimilate its borrowed neoclassical elements phonologically (but not totally) [12], in other words, a Greek or Latin word goes under a minimal adaptation before being adopted by a host language. For example, both FR[1] *pathie* and EN *pathy* were borrowed from the Greek word *pathos*. In addition, each element can have different allomorphs, which means that a borrowed Greek or Latin element can be assimilated to different forms in one language. For example, the English neoclassical element *neuro* can have two forms in French: *neuro* like in FR *neurologie* and *névro* like in FR *névrodermite*.

Neoclassical elements can appear at different positions in neoclassical compounds: (1) initial position in a neoclassical compound, like *homo-* in *homomorphic*, (2) final position such as *-cide* in *genocide*. According to [13], we distinguish between Initial Combining Forms (ICFs) and Final Combining Forms (FCFs). ICFs include forms of neoclassical elements that appear at initial positions (e.g. *bio-*, *cardio-*, *patho-*, etc.), while FCFs include forms of neoclassical elements that appear at final positions (e.g. *-logy*, *-cide*, *-pathy*, etc.). Moreover, more than one ICF may appear sequentially in a neoclassical compound (e.g. *histo-* and *patho-* in *histopathology*).

A neoclassical element (borrowed from the same Greco-Latin word) can be seen both as ICF and FCF, for instance, *patho-* in *pathology* and *-pathie* in *cardiopathie*, both elements being adapted from *pathos*. This property enables to distinguish between neoclassical elements and affixes. The latter appear at fixed positions: either at initial positions (prefixes, e.g. *pre-* like in *premature*) or at final positions (suffixes, e.g. *-ist* like in *chemist*). Furthermore, neoclassical elements have been introduced later than prefixes and suffixes to many european languages, around the 19th century to English [14].

3 Neoclassical compound alignment

In this section, we give the assumptions we make to align neoclassical compounds, the forms of neoclassical compounds that can be aligned, and finally we explain the main steps of our alignment approach.

3.1 Assumptions

The method is based on the following assumptions:

Compositional property
Neoclassical compounds are translated compositionally to neoclassical compounds. Each element in a neoclassical compound is translated individually and the final translation is the combination of the translated elements; as

[1] FR, EN, DE, and ES denote to French, English, German, and Spanish respectively.

the meaning of neoclassical compounds can be in many cases obtained compositionally [1]. For example, the translation of EN *hydrology* in French is *hydrologie*, which can be interpreted by the combination of the translation of the composing elements, *hydro* (water): FR *hydro* and *logy* (study): FR *logie*.

Translating a neoclassical compound compositionally to a neoclassical compound can give accurate results even in cases where a neoclassical compound is not fully compositional. From [15], we take EN *leukopathia* (a disease involving loss of melanin pigmentation of the skin) as an example of a neoclassical compound that is indeterminent by its elements; as its definition does not contian an explicit reference to *white* which is the meaning of its first composing element *leuko*. However, the equivalents of *leukopathia* in other languages are: FR *leucopathie*, DE *leukopathie*, ES *leucopatia*, this means that neoclassical compounds can still be translated compositionally even when their exact meaning cannot be restored compositionally.

Preserving the order of elements

The order of the elements of a source neoclassical compound is preserved in the equivalent target neoclassical compound. Taking *hydrology* as an example, the equivalent of *hydro* will appear before the equivalent of *logy* when combining the equivalents of the constituent elements to form the final translation. This assumption is based on the fact that neoclassical word-formation in different languages follows the model of Greek and Latin languages in forming terms [10]. The Greco-Latin template for a term XY that consists of two elements X and Y is: [determinant determinatum]. According to this template, *cardiology* consists of *logy* (study) being the determinatum (identifies the class of which the neoclassical compound is a kind) and *cardio* (heart) being the determinant (gives the differentiating feature).

According to the second assumption, the order of elements should be respected when translating a neoclassical compound. Therefore, each neoclassical constituent element is translated with a neoclassical element of the same type (for instance an ICF by an ICF, an FCF by an FCF).

Hereafter, we assume that neoclassical compounds are either adjectives or nouns since this is true for most of the cases. In spite of the fact that some verbs can contain neoclassical elements, e.g. *hydrogenate*.

3.2 Handled forms

Neoclassical compounds can take different forms. They contain at least one neoclassical element but also can contain native words and/or prefixes combined in different orders. Our method can only align neoclassical compounds that belong to one of the following forms:

- **ICF+ FCF**

  The first form includes neoclassical compounds that consist only of neoclassical elements. One or more ICFs can appear sequentially along with one

FCF. This form is equivalent to the combination of the first and the last forms presented in [11]. Examples of neoclassical compounds are given in (1).

(1) FR histopathologie ($histo_{ICF}$ / $patho_{ICF}$ / $logie_{FCF}$), EN radiology ($radio_{ICF}$ / $logy_{FCF}$), DE radiometrie ($radio_{ICF}$ / $metrie_{FCF}$), ES geomorfologia (geo_{ICF} / $morfo_{ICF}$ / $logia_{FCF}$)

– **ICF+ Word**
This form includes one or more ICFs combined with a native word (lexical item that is a single word). This form is equivalent to combining the third, forth, and the last forms defined in [11]. Examples of neoclassical compounds are illustrated in (2).

(2) FR cardiovasculaire ($cardio_{ICF}$ / $vasculaire_{Word}$), EN photobioreactor ($photo_{ICF}$ / bio_{ICF} / $reactor_{Word}$), DE ferroelektrisch ($ferro_{ICF}$ / $elektrisch_{Word}$), ES multidisciplinario ($multi_{ICF}$ / $disciplinario_{Word}$)

3.3 Approach

The method that we propose firstly extracts neoclassical compounds for source and target languages from comparable corpora using two lists of neoclassical elements, the first for the source language (NE_{l_s}) and the second for the target language (NE_{l_t}); this results in lists of source and target neoclassical compound candidates: NC_{l_s} and NC_{l_t}. Then, each neoclassical compound in NC_{l_s} is aligned with its equivalent(s) in NC_{l_t}, with the help of a bilingual dictionary Dic_{Bi} and an aligned list of neoclassical elements NE_A. The method follows the three main steps of the compositional methods for aligning complex terms [2] [3]: i) the extraction of candidates, ii) the generation of translation candidates, and iii) the selection of correct translations.

1. **Extraction of neoclassical compound candidates**
Source and target neoclassical compound candidate lists (NC_{l_s} and NC_{l_t}) are obtained by projecting NE_{l_s} on the corpus of the source language l_s, and NE_{l_t} on the corpus of the target language l_t. The adjectives or nouns that have at least one neoclassical element (ICF or FCF) are considered as neoclassical compound candidates. An ICF can appear in the beginning or anywhere in the middle of a neoclassical compound, e.g. ICFs *bio-*, *geo-* and *morpho-* appear in *biogeomorphological*. FCFs are found at the end of neoclassical compounds such as *-pathy* in *neuropathy* and *-logie* in *biotechnologie*.
2. **Generation of translation candidates**
The projection made in the extraction phase results in decomposing each extracted neoclassical candidate into two or more elements, in which at least one of these elements is a potential neoclassical element. The form of a neoclassical candidate is checked, and in case it is identified as one of the two handled forms presented in section 3.2, the method will try to generate its

translation candidates while respecting the assumptions explained in section 3.1. Equivalents of identified ICFs and FCFs are found using NE_A whereas the translations of native words are obtained from Dic_{Bi}. The translation candidates are all the possible combinations (following the Greco-Latin template) of the translations of each element of the neoclassical compound candidate (NC_s). The generation succeeds only if all elements of NC_s are identified. For example, suppose that we identify the two elements (*neuro-* and *-logy*) as neoclassical elements in the neoclassical compound EN *neurology*. To generate its French translation candidates, we search for the ICF equivalent(s) of EN *neuro-* in NE_A, which would be FR *neuro-* and FR *névro-*, as well as the FCF equivalent(s) of EN *-logy*, which would be FR *-logie*. Accordingly, two translation candidates will be generated by combining the translations of elements: *neurologie* and *névrologie*. Taking another example: we want to generate English translation candidates for FR *bioscience*, which matches the second form of neoclassical compounds. We can identify FR *bio* as neoclassical element and FR *science* as a word in the dictionary. The equivalent ICF of FR *bio* is EN *bio* that we could obtain from NE_A, while the translations of FR *science* in Dic_{Bi} could be *art*, *science*, *information*, *knowledge* and *learning*. Consequently, five translation candidates will be generated: *bioart*, *bioscience*, *bioinfomation*, *bioknowledge* and *biolearning*.

3. **Selection of correct translations**
   We look up each translation candidate (obtained in the generation phase) in the target neoclassical compound list NC_{l_t}. In case the candidate is found, it will be considered as a correct translation for its respective source neoclassical compound NC_s. For example, if two French translation candidates were generated for EN *neurology*: *neurologie* and *névrologie*, they will be searched in the target neoclassical list NC_{l_t}. The candidate *névrologie* would not be found as it is not the correct translation, but there is a probability that *neurologie* would be found in NC_{l_t}, and therefore considered to be a valid translation.

The main steps of the method are summarized in 1.

Algorithm : Neoclassical compound alignment

```
NCls[] = ExtractNeoclassicalCompoundCandidates(Cs)
NClt[] = ExtractNeoClassicalCompoundCandidates(Ct)
for each NC in NCls
Candidates[] = GenerateTranslationCandidates(NC)
for each Candidate in Candidates
if (Candidate exists in NClt)
       Select Candidate as translation for NC
```

Fig. 1. Algorithm for aligning neoclassical compounds. Cs = source corpus. Ct = target corpus.

4 Evaluation

We present in section 4.1 the resources that we used to do the experiments that led us to the results that we present in section 4.2.

4.1 Resources

We carried out the experiments using comparable corpora built with the BABOUK crawler [16]. The corpora are related to the renewable energy domain in four languages. We pre-processed each corpus by running a word tokenizer, a POS tagger, and a lemmatizer. Table 1 lists the languages with the corresponding number of unique nouns and adjectives in the corpus.

Table 1. Corpora statistics

Language	**No. of unique adjectives and nouns**
English	24,250
French	13,625
German	51,624
Spanish	15,785

As for neoclassical elements, we have taken 113 French neoclassical elements from [17]. Then, we have manually aligned 83 of these neoclassical elements with their English equivalents. We then aligned 61 English neoclassical elements with their German equivalents as well as 58 English neoclassical elements with their Spanish equivalents. This led us to obtaining three lists of aligned neoclassical elements for the pairs of languages (EN-FR, EN-DE, and EN-ES), and four lists of monolingual neoclassical elements (see Table 2). We have also used three bilingual dictionaries for EN-FR, EN-DE, and EN-ES that contain 145,542, 69,876, and 61,587 single-word entries respectively[2].

Table 2. Sizes of monolingual neoclassical element lists

	EN	**FR**	**DE**	**ES**
NE size	83	113	61	58

[2] Dictionaries were obtained from EURADIC French-English dictionary http://catalog.elra.info/product_info.php?products_id=666, http://www.dict.cc/, and http://www.phrozensmoke.com/projects/pythonol/.

4.2 Results and Discussion

Tables 3 and 4 present the results of the experiments carried out on the method for the three pairs of languages (EN-FR, EN-ES, and EN-DE) in both directions. For example, Table 1 shows that we were able to extract 1215 English neoclassical compound candidates using the English neoclassical element list. Although many of these are false neoclassical compounds (e.g. *decision*, *communication*). For the pair of languages EN-FR, French translation candidates were automatically generated for 264 of the English extracted neoclassical compound candidates when using 83 FR-EN neoclassical aligned elements and the FR-EN bilingual dictionary (presented in section 4.1). The correct translation(s) among the generated candidates were found in the target neoclassical compound list for 100 of the 264 candidates. A generated translation candidate that is not found does not necessarily mean that it is a wrong translation; it just could possibly be a correct translation that is missing from the target corpus. The precision obtained from the alignment was about 98%, and the recall was about 37%. The recall was calculated by dividing the number of true positives (correct translations) on the sum of true positives and false negatives (identified neoclassical compounds with no suggested translation).

For all languages, false positive alignments were mainly the translations that were obtained from false extracted neoclassical compounds (noise, non-neoclassical compounds), e.g. the French word *histoire* was extracted from English corpus since *histo* was identified as neoclassical element and *ire* was identified as English word. Thus, *histoire* was aligned with FR *histoire* (*ire* is a French translation of EN *ire*). Erroneous translations were also obtained because of the fact that neoclassical compounds are not always translated to neoclassical compounds from one language to another, e.g. FR *télécommande* was translated to EN *telecontrol*, while the correct translation is EN *remote control*.

Table 3. Alignment of neoclassical compounds for (EN-FR, EN-DE, and EN-ES)

Languages	**Aligned neoclassical elements**	**Neoclassical compound candidates**	**Generated translations**	**Found translations**	**Precision**	**Recall**
EN-FR	83	1215	264	100	98%	37%
EN-DE	61	1215	266	100	96%	36%
EN-ES	58	1215	219	68	97%	30%

A neoclassical compound candidate NC_s can be extracted (since it contains a possible neoclassical element) but still cannot be identified as one of the neoclassical forms our method handles, the generation of its translation candidates will fail. This can be due to several reasons:

– **False neoclassical element**: a candidate like EN *decision* will be decomposed into two elements; the first of which is *deci* will be considered as neoclassical element (false neoclassical element). The second is *sion* which

Table 4. Alignment of neoclassical compounds for (FR-EN, DE-EN, and ES-EN)

Languages	Aligned neoclassical elements	Neoclassical compound candidates	Generated translations	Found translations	Precision	Recall
FR-EN	83	1068	263	94	97%	35%
DE-EN	61	3538	437	105	96%	23%
ES-EN	58	2126	363	69	97%	18%

is neither a neoclassical element nor a known word in the bilingual dictionary.

- **Missing neoclassical element from** NE_A: if a candidate like FR *métronome* is extracted and only the equivalent of *métro* is found in NE_A, the generation will fail because the equivalent(s) of *nome* (a real neoclassical element) are not found in NE_A.

- **Untreated neoclassical form**: a true neoclassical candidate could be extracted but it belongs to a form that we do not handle. There exist other forms of neoclassical elements, for example, EN *antibiogram* (*anti*: prefix, *bio*: ICF, *gram*: FCF) is a form the method does not cover.

Bilingual alignment using a pivot language

We can obtain bilingual lists of neoclassical compounds using a pivot language. Generally speaking, source-to-target translation of a word through a pivot language occurs in two steps: (1) the word is translated to the pivot language using a bilingual source-to-pivot dictionary, (2) the obtained translation from the previous step is translated to the target language using a bilingual pivot-to-target dictionary. Pivot language approach is known to be highly noisy because of polysemy in languages and intransitivity of lexicons. However, bilingual alignments of neoclassical compounds through a pivot language should be as precise as bilingual neoclassical compound alignments obtained by our method (presented in section 3) because neoclassical elements remain stable in meaning across languages.
We chose EN as pivot language to obtain a list of FR-DE alignments as well as ES-FR and ES-DE alignments using the aligned lists of neoclassical compounds for (EN-FR, EN-DE, and EN-ES). Accordingly, we obtained the results shown in Table 5. We conclude that using a pivot language to align neoclassical elements is a confident approach as languages follow the Greco-Latin template when creating neoclassical compounds.

5 Conclusion

In this paper, we presented a compositional-like method to align neoclassical compounds in two languages (source-target). The method handles two forms

Table 5. Alignment of neoclassical compounds using English as pivot language

Languages	Alignments	Precision
FR-DE	61	98%
ES-FR	50	100%
ES-DE	44	97%

of neoclassical compounds. For this task, it uses predefined monolingual neoclassical elements to extract neoclassical compound candidates, and a list of aligned neoclassical elements in addition to a bilingual dictionary to align the extracted candidates. The results showed high precision (more than 96%) in aligning neoclassical compounds of the two handled structures. Moreover, we showed that a pivot language approach gives high-precision neoclassical compound alignments too. We aim at expanding the method so that it covers other possible forms of neoclassical compounds. We also aim to investigate the possibility of automatically extracting neoclassical elements for one language and aligning them with their equivalents in another language.

References

1. Estopa, R., Vivaldi, J., Cabre, M.T.: Use of greek and latin forms for term detection. In: The 2nd international conference on language resources and evaluation. Volume 78. (2000) 885–859
2. Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., Utsuro, T.: Compiling french-japanese terminologies from the web. In: EACL. (2006) 225–232
3. Baldwin, T., Tanaka, T.: Translation by machine of complex nominals: Getting it right. In: ACL Workshop on Multiword Expressions: Integrating Processing. (2004) 24–31
4. Vintar, S.: Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. Terminology **16** (2010) 141–158
5. Grefenstette, G.: The world wide web as a resource for example-based machine translation tasks. In: Translating and the Computer 21, London, ASLIB (1999)
6. Vincent, C., Ewa, K.: Analyse morphologique en terminologie biomédicale par alignement et apprentissage non-supervisé. In: Conférence Traitement automatique des langues naturelles TALN, Montréal, Québec, Canada (2010)
7. Namer, F., Baud, R.H.: Defining and relating biomedical terms: Towards a cross-language morphosemantics-based system. I. J. Medical Informatics (2007) 226–233
8. Cartoni, B.: Lexical morphology in machine translation: A feasibility study. In: EACL. (2009) 130–138
9. Bowker, L., Pearson, J.: Working with specialized language: a practical guide to using corpora. London, Routledge (2002)
10. Amiot, D., Dal, G.: La composition néoclassique en français et l'ordre des constituants. La composition dans les langues, Artois Presses Université (2008) 89–113
11. Namer, F.: Morphologie, lexique et traitement automatique des langues. Lavoisier (2009)
12. Lüdeling, A.: Neoclassical word-formation. In: Keith Brown (ed) Encyclopedia of Language and Linguistics, 2nd Edition, Oxford, Elsevier (2006)

13. Bauer, L.: English word-formation. Cambridge university press (1983)
14. Baeskow, H.: Lexical properties of selected non-native morphemes of English. Gunter Narr Verlag (2004)
15. McCray, A., Browne, A., Moore, D.: The semantic structure of neo-classical compounds. In: The Annual Symposium on Computer Application in Medical Care. (1988) 165168
16. Groc, C.D.: Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. In: The IEEEWICACM International Conferences on Web Intelligence, Lyon, France (2011) 497–498
17. Béchade, H.D.: Phonétique et morphologie du français moderne et contemporain. Presses Universitaires de France (1992)