INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 78S (2009) S48–S55

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Morphosemantic parsing of medical compound words: Transferring a French analyzer to English

Louise Deléger[a,b,c,*], *Fiammetta Namer*[d], *Pierre Zweigenbaum*[e,f]

[a] *INSERM U872, Eq. 20, 15 rue de l'Ecole de Médecine, Paris F-75006, France*
[b] *UPMC, Paris F-75006, France*
[c] *Paris-Descartes University, Paris F-75006, France*
[d] *UMR 7118, ATILF, Université Nancy 2, CLSH, Nancy F-54015, France*
[e] *CNRS UPR3251, LIMSI, Orsay F-91403, France*
[f] *INALCO, CRIM, Paris F-75007, France*

ARTICLE INFO

Article history:
Received 20 February 2008
Received in revised form
24 June 2008
Accepted 30 July 2008

Keywords:
Natural language processing
Morphosemantic analysis
Word definition
Neoclassical compounds
English
French

ABSTRACT

Purpose: Medical language, as many technical languages, is rich with morphologically complex words, many of which take their roots in Greek and Latin—in which case they are called neoclassical compounds. Morphosemantic analysis can help generate definitions of such words. The similarity of structure of those compounds in several European languages has also been observed, which seems to indicate that a same linguistic analysis could be applied to neo-classical compounds from different languages with minor modifications.
Methods: This paper reports work on the adaptation of a morphosemantic analyzer dedicated to French (DériF) to analyze English medical neo-classical compounds. It presents the principles of this transposition and its current performance.
Results: The analyzer was tested on a set of 1299 compounds extracted from the WHO-ART terminology. 859 could be decomposed and defined, 675 of which successfully.
Conclusion: An advantage of this process is that complex linguistic analyses designed for French could be successfully transposed to the analysis of English medical neoclassical compounds, which confirmed our hypothesis of transferability. The fact that the method was successfully applied to a Germanic language such as English suggests that performances would be at least as high if experimenting with Romance languages such as Spanish. Finally, the resulting system can produce more complete analyses of English medical compounds than existing systems, including a hierarchical decomposition and semantic gloss of each word.

1. Introduction

Medical language, as many technical languages, is rich with morphologically complex words, many of which take their roots in Greek and Latin. These so-called neoclassical compounds [1] are present in many areas of the medical vocabulary, including anatomy (gastrointestinal), diseases (encephalitis, cardiomyopathy), and procedures (gastrectomy). Segmenting morphologically complex words into their base components is the task of morphological analysis. When the analysis is complemented by semantic interpretation, the process is called morphosemantic analysis. This type

* *Corresponding author*. Tel.: +33 153109213.
E-mail address: louise.deleger@spim.jussieu.fr (L. Deléger).
1386-5056/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2008.07.016

of analysis is especially suited to neo-classical compounds as their meaning is often "compositional," in the sense that it is a combination—at least partial—of the meaning of the constituent parts.

Morphosemantic analysis can therefore help processes interested in semantics, such as the detection of similar terms, the generation of definition, or the retrieval of medical documents. This was for instance the aim of [2], where a tool for unsupervised learning of morphological segmentation [3] was used to contribute mappings between WHO-ART and SNOMED terms in order to cluster semantically close WHO-ART terms to group related medical conditions. The idea was to morphologically decompose WHO-ART terms into their constituent parts, and map them thanks to a table of components paired with SNOMED terms. Computation of semantic distance was then performed between the decomposition of the terms. Another application of morphosemantic analysis is cross-language document retrieval. In [4,5] for instance, document and query terms are morphologically segmented, each component being mapped to an identifier and synonymous components sharing the same identifiers. In this article we focus on the methods to perform the morphosemantic analysis useful for these applications.

It has also been observed that the morphological structure of neo-classical compounds is similar in numerous European languages [6]. It therefore seems possible to transfer a linguistic analysis dedicated to neo-classical coumpounds from one language to other related languages [7] proved it for a certain type of medical compounds by proposing an analysis of pathology names (as *hypercalciuria*) that could be applied to French, German, Spanish, Italian and English. Morphosemantic analysis of such compounds demonstrates a multilingual potential.

Several approaches have been dealing with the analysis of those complex words. Early work on medical morphosemantic analysis focused on specific morphemes such as *-itis* [8] or *-osis* [9], then on larger sets of neoclassical compounds [10]. Lovis [11] introduced the notion of morphosemantemes, *i.e.*, units that cannot be further decomposed without losing their original meanings. The Morphosaurus system [4,5] segments complex words using a similar notion called subword [12]. The UMLS Specialist Lexicon [13], with its "Lexical tools," handles derived words, *i.e.*, complex words built through the addition of prefixes or suffixes. It provides tables of neoclassical roots, but no analyzer to automatically decompose compound words. DériF [14] morphosemantically analyses complex words. In contrast to [5] or [11], it computes a hierarchical decomposition of complex words. Moreover, it produces a semantic definition of these words, which it can link to other words through a set of semantic relations including synonymy and hyponymy. In contrast to the Specialist tools or to [15], DériF handles both derived and compound words. Designed initially for French general language complex words, then extended to the medical domain, its potential for cross-linguistic application was showed in [16]. Its transposition to English would fill a gap in the set of tools currently available to process complex English medical words.

This paper[1] reports work on the adaptation of DériF to medical English complex words. It focuses on neoclassical compounds, leaving aside derived words. Our goal is to have DériF analyse English words and present its results in English, thus illustrating the similarity of structure between compounds from related languages, and obtaining a tool which is missing for the English language, or at least unpublished.

We first describe the morphosemantic analyzer and our test set of words. We explain the modifications performed on this tool and the evaluation conducted. We then expose the results, discuss the method and conclude with some perspectives.

2. Theoretical background

The principle on which this work is based is morphosemantic analysis, that is morphological analysis associated to a semantic interpretation of words. In other words, we want to obtain a decomposition into base components, and a description of the meaning of a complex word based on the meanings of these components. A complex word may consist of different types of base components:

- affixes (prefixes and suffixes), *e.g.*, *de-*, *pre-*, *-al*, *-ic*, *-ful*;
- classical roots, which are called combining forms (CFs): *gastr-*, *arthr-*, *-uria*, *-itis*;
- simplex modern-language words that cannot be decomposed: *pain*, *head*.

A complex word is built from any combination of the following two word creation rules:

- derivation, which adds affixes to base words, *e.g.*, *pain/painful*;
- compounding, which joins two (or more) words together, each of those words being either combining forms (CFs) or modern-language words, *e.g.*, *backdoor*, *arthritis*.

For this work we chose to analyze compounds, more specifically neo-classical compounds (compounds formed from at least one CF). However, as we stated above, a complex word may have been built by both compounding and derivation. A compound may be subjected to a derivation rule and a derived word may be a component of a compounding process. Therefore mixed-formation words (both derived and formed from at least one CF, such as *haemorrhagic*) must also be addressed.

Our work hypothesis to transpose morphosemantic analysis from French to English is that a same linguistic analysis may be applied to neo-classical compounds from several languages. We indeed assume that these compounds are built in a similar fashion and that the components involved in the process are similar, the main differences being orthographical (as for instance, *-algie* in French and *-algia* in English). Potential obstacles to direct transposition from French to English could arise in:

[1] This paper is an extended description of work presented in [17,18].

- the order of combination of the components: the analysis will not succeed if the combination order is not the same in the two languages. This case should be rare since we are dealing with classical compounding which adopts Latin or Greek order;
- the components themselves: French and English analyses can only match if both French and English words are formed from CFs. This is our hypothesis and the analysis will be possible because these CFs are listed and in limited number (in this work our list of CFs contains 945 elements);
- the combination of the components: the first component may vary when it is combined with a second one (allomorphy), for instance adding the linking vowel -o- to the first constituent. If these phenomena are different in the two languages, this may cause problems in the analysis. We assume they are similar, aside from orthographical modifications;
- the morphological processes of suffixation and prefixation applied to the neoclassical compounds. Affixes are indeed different in the two languages. However we assume that in the case of neoclassical compounds it is sufficient to replace French affixes by English affixes of the same "class" (for instance, suffixes used to form French relational adjectives, *e.g.*, *-ique, -al* may be replaced by their English counterparts, *e.g.*, *-ic, -al*).

3. Materials and methods

3.1. *Materials*

We started from the French version of the DériF ("Derivation in French") morphosemantic analyzer. DériF was designed both for general language and more specialized vocabularies such as medical language. Its analysis is purely based on linguistic methods and implements a number of decomposition rules and semantic interpretation templates. Resources necessary to the tool include lexicons of word lemmas tagged with their parts-of-speech and a table of combining forms (to be detailed below). When applied to biomedical vocabulary, the system goes further than simple decomposition and interpretation steps by predicting lexically related words.

As input the system expects a list of words tagged with their parts-of-speech and lemmatized (in their base form—no plural). It outputs the following elements:

1. a structured decomposition of the word into its component parts that represents the order of the rules successively applied to analyse the word;
2. a definition ("gloss") of the word in natural language, according to the meaning of the components;
3. a semantic category, inspired by the main MeSH tree descriptors (anatomy, organism, disease, etc.);
4. a set of potentially lexically related words. The relations identified are equivalence relations (*eql*), hyponymy relations (*isa*), meronymy relations and see-also relations (*see*).

For instance, the French word *acrodynie* (English *acrodynia*) is analyzed in the following way (*N* stands for *noun*, *N** is assigned to a noun CF):

Table 1 – Extract from the list of complex words containing at least one combining form (N = noun, ADJ = adjective)

arthralgia/N
atelectasis/N
blepharospasm/N
calcinosis/N
capillary/ADJ
cardiomegaly/N
cerebellar/ADJ
claustrophobia/N
clostridial/ADJ
cryptococcal/ADJ
crystalluria/N
dermatomyositis/N
dextrocardia/N
dorsal/ADJ
dysmenorrhea/N

acrodynie/N ==>

1. [[acr N*] [odyn N*] ie N].
2. douleur (de/lié(e) à) extrémité (*pain of/linked to extremity*).
3. maladie (*disease*).
4. eql:acr/algie, eql:apex/algie, see:acr/ite, see:apex/ite (the slashes are here to separate the CFs).

To test the transposition of DériF to English, we prepared a list of test words. These words were taken from the WHO-ART terminology since one of the intended applications of this work is to contribute to the pharmacovigilance domain. We selected the English terms of this terminology; since DériF works on single words and not on multi-word units, we split them into single words; and since we adapted DériF to analyze neo-classical compounds, we only retained those types of words. The selection was done both automatically by removing all words of four characters or less (these words are practically never morphologically complex), and manually by reviewing the list to look for neoclassical compounds (the work was done by a language engineer, LD). This gave us a list of 1299 words to be decomposed out of a total of 3476 words. Among those words 8.6% were composed of more than two CFs. As stated earlier, we selected both "pure" compounds (only composed of CFs) and mixed-formation words (words formed from derivation and from at least one CF). 540 (41.6%) were pure compounds and 759 (58.4%) were mixed-formation words. An extract of the list is given in Table 1. The words were lemmatized and tagged with their parts-of-speech using the TreeTagger[2] part-of-speech tagger [19]. We used a lexicon of tagged words from the UMLS Specialist lexicon[3] to help TreeTagger deal with unknown words.

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (last accessed 20.02.08).

[3] http://www.nlm.nih.gov/pubs/factsheets/umlslex.html (last accessed 20.02.08).

Table 2 – Table of combining forms (excerpt) the lexical relations between the CFs are labelled as follows: <- for a meronymy relation, ~ for a see-also relation and no sign for an equivalence relation

Root	English	Type	Relations
algia	Pain	Disease	odyn, algo, ~itis
blephar	Eyelid	Anatomy	palpebr, <- ocul, ~coro
ectomy	Surgical excision	Act	~tomy, ~stomy
gastr	Stomach	Anatomy	stomac, gaster, ~hepat, ~entero, <-abdomin, ~pancreat

3.2. *Methods*

3.2.1. *Adapting the morphosemantic analyzer*

The method of morphological analysis that we want to transpose to English is schematized in Fig. 1.

Language-specific material is located in:

1. *the lexicon* of tagged word lemmas which is used by the system to test whether a component exists and to retrieve its part-of-speech;
2. *the table of CFs*: each form is associated with a modern-language word which describes its meaning, a semantic type, a part-of-speech, and a set of CFs related through relations *eql*, *isa*, *see*, *part-of*;
3. *the decomposition rules*: these rules are triggered in a certain order according to the part-of-speech of the word and to the suffix identified (if any). They identify the head of the word and its other components, relying on the table of CFs and lexicon. They are responsible for the hierarchical structuring of the decomposition, and therefore intervene in both derivation and compounding (see [20] for examples of complex rules for compounds). Each rule may have a set of exceptions. Orthographic normalization is also performed on the components;
4. *the semantic templates*: they provide template glosses of a complex word based on the relation between its head and its other components.

We address each of these in turn.

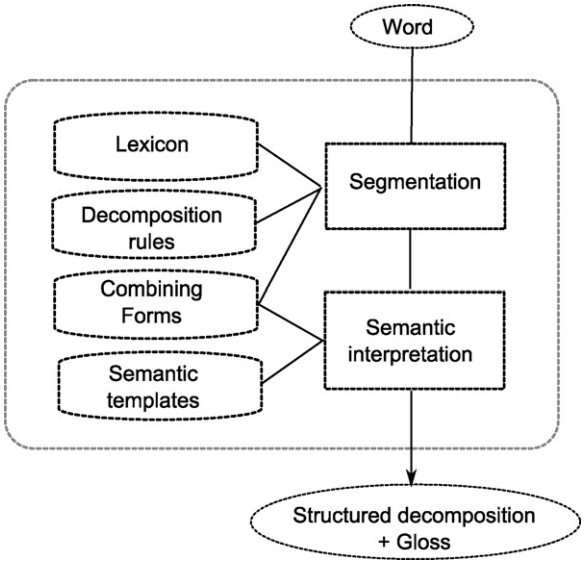

Fig. 1 – Morphosemantic parsing.

1. We replaced the French lexicon by an English lexicon derived from the UMLS Specialist lexicon.
2. Our hypothesis is that CFs are mainly the same in English and French save for a few minor orthographic differences. We handled these by making small modifications to the French CFs (*e.g.*, removing accents as in *blépharo—blepharo*) to obtain their English equivalents. Semantic types might have been left in French, but we considered that English names would be more consistent; this was done very easily since they were few. We also associated modern-English words to the CFs. They were obtained from two lists of CFs, one taken from the UMLS Specialist lexicon, the other one extracted from the Dorlands medical dictionary, where CFs were paired with modern-English words.[4] We automatically matched them to our English CFs and reviewed the results. Those CFs that could not be matched were dealt with manually. Finally, the set of semantically related CFs was replaced by the English equivalents (obtained through orthographic modifications as stated above). The parts-of-speech were kept as is. An extract of the resulting table can be seen in Table 2.
3. Intervention at the rule level remained limited since we assumed that neo-classical compounds were formed in a similar way in French and English. So this mainly involved adapting the exceptions, the orthographic normalizations performed and the affixes (*e.g.*, French suffix *-ique* was replaced by *-ic*).
4. Finally, we translated the semantic templates so that they could generate English glosses. For instance, the following template is now associated with the *-ia* suffix, the *hyper-* prefix and X/Y as nouns (components of the word occurring in the linear order YX):

   Affection of **X** linked to the excess of **Y**

where **X** and **Y** will be substituted with modern-language equivalents of CFs or with words from the lexicon, as in *hypercalcemia*, which is analyzed into *Affection of blood linked to the excess of calcium*.

3.2.2. *Evaluation of the resulting analyzer*

We ran the updated DériF on the 1299 complex words extracted from WHO-ART. The expected output for each word is a hierarchical decomposition into components, a definition in natural language, a semantic type and a set of related words, which we evaluated for coverage and validity. We define

[4] http://www.merckmedicus.com/pp/us/hcp/thcp_dorlands_content.jsp?pg=/ppdocs/us/common/dorlands/dorland/dmd-a-b-000.htm (last accessed 20.02.08).

Table 3 – Evaluation results for the English version of DériF

List of words	Number of words	Decomposed words (coverage)	Number of correct results	Precision (%)	Recall (%)
All	1299	859 (66%)	675	78.5	52
Compounds	540	368	282	76.6	52.2
Mixed-formation	759	491	393	80	51.8

coverage as the proportion of words the system was able to analyze. An analysis was considered valid if its decomposition and definition were correct; in this evaluation, we did not take into account the semantic type nor the set of potentially related words. At this point we computed two standard measures of evaluation in natural language processing: precision and recall. Precision is the ratio of correct analyses over the total number of analyses. Recall is the ratio of correct analyses over the total number of analyses that should have been produced. We computed those measures on the set of all 1299 words together, as well as on the set of 540 pure compounds and the set of 759 mixed-formation words. To have an idea of how our tool performs in term of linear unstructured decomposition (as proposed by existing tools) we also evaluated precision as regards decomposition without taking into account its structuring. Finally, because our list of words was selected manually and included only complex words, therefore reducing potential noise (*i.e.* unwanted decomposition of words not composed of CFs), we also ran the system on a random sample of 100 words from WHO-ART that had not been selected in our list. We measured the proportion of those words that were decomposed (the closest the proportion is to zero the less the system produces noise).

Besides, since the method of [2] is applied, as ours, to the WHOART terminology, we could set up a comparison. The same word list we submitted to the English version of DériF was also given to Morfessor, the morphological segmenter used by [2], in exactly the same conditions as in [2] and we evaluated coverage, precision and recall to compare the results.

4. Results

In the present state of its transposition to English, DériF was able to analyze 859 out of the 1299 words in our list (see Table 3), thus obtaining 66% coverage. Example word analyses are given in Table 4. They gave an overview of the words that can be analyzed with the system. The first four lines of the table display pure neo-classical compounds, and the following ones show derived words built on a classical basis. These mixed-formation words can be suffixed, prefixed or both. The table gives examples of the most frequent adjectival suffixes:

Table 4 – Example word analyses generated by the English DériF

Word/POS	Decomposition	Definition	Type	Related words
arthralgia/N	[[arthr N*] [algia N*] N]	Pain (of/linked to) joint	Disease	eql:arthr/algesia see:arthr/it is
appendicitis/N	[[appendic N*] [itis N*] N]	(Part of/Specific type of) inflammation related to appendix	Disease	eql:appendic/phlogo, see:appendic/algia, see:appendic/odyn
thyroidectomy/N	[[thyroid N*] [ectomy N*] N]	Surgical excision (of – towards) thyroid	Action	see:thyroid/tomy
acrocyanosis/N	[[acr N*] [[cyan A*] [osis N*] N] N]	(Part of/Specific type of) cyanosis related to extremity	–	eql:apex/cyanosis
arteriovenous/ADJ	[[arterio N*] [veno N*] ous ADJ]	Related to vein and artery	Anatomy	eql:arterio/phleb, isa:angi/veno
cardiovascular/ADJ	[[cardio N*] [vascul N*] ar ADJ]	Related to vessel and heart	Anatomy	eql:cardio/angi
gastroesophageal/ADJ	[[gastr N*] [oesophag N*] al ADJ]	Related to oesophagus and stomach	Anatomy	eql:stomac/oesophag isa:abdomin/oesophag
hepatopulmonary/ADJ	[hepat N*] [pulmon N*] ary ADJ]	Related to lung and liver	Anatomy	eql:hepat/pneumon, isa:abdomin/pneumon
psychosomatic/ADJ	[[psych N*] [somat N*] ic ADJ]	Related to body and mind	Anatomy	–
macrocephaly/N	[[macro A*] [cephal N*] y N]	Condition linked to head characterized by large	Disease	eql:mega/cephal, see:gigant/cephal
menorrhagia/N	[[meno N*] [rrhage N*] ia N]	Excessive flow of menstruation	Disease	isa:meno/rrhea
intraretinal/ADJ	[intra+ [retin N*] +al ADJ]	Located in the interior of the retina	Disease	–
adactyly/N	[a+ [dactyl N*] +y N]	Affection characterized by the absence of digit	Disease	–
hypoplasia/N	[hypo+ [plas N*] +ia N]	Affection characterized by growth below average	–	–
hyperkalemia/N	[hyper+ [kal N*][em N*] +ia N]	Affection of blood linked to the excess of potassium	Disease	isa:kal/it is

N = noun, ADJ = adjective, N* = nominal combining form, A* = adjectival combining form.

Table 5 – Examples of incorrect word analyses generated by the English DériF

Word/POS	Decomposition	Definition	Type	Related words
meningo-encephalitis/N	[[mening N*] [encephal N*] [itis N*] N] N]	(Part of – Specific type of) encephalitis related to meninges	–	–
acanthosis/N	[[acanth N*] [osis N*] N]	(Part of – Specific type of) disease related to prickle	Disease	–
alveolar/N	[[alveolar A] N]	Entity being alveolar	–	–

N = noun, ADJ = adjective, N* = nominal combining form, A* = adjectival combining form.

-ous, -ar, -al, -ic, -ary, and nominal suffixes: *-y, -ia*. The last four lines show words that are both suffixed and prefixed. Common prefixes include *hypo-, hyper-, a-, intra-*. The fourth line of the table is an example of a compound composed of more than two CFs. These examples illustrate the ability of the tool to deal with different kinds of complex words.

We measured a total precision of 78.5% (see Table 3) which is fairly good. Combined to a moderate coverage, this yields a 52% recall. The compounds obtained 76.6% precision and 52.1% recall, while the mixed-formation words obtained 80% precision and 51.8% recall. Precision for only linear decomposition was 95% which shows that high quality results are more easily obtained when taking into account neither the order of decomposition nor the definition.

We identified several causes of non-decomposition:

- Some elements were unknown, *i.e.*, they were neither in the CF table nor in the word lexicon: *e.g.*, *campt-* is not listed in the CF table so the word *camptodactyly* could not be decomposed. This accounted for 39% of the cases of non-decomposition and is easily perfectible by adding these elements to our lists.
- Certain suffixation rules were not implemented in DériF at the time of the experiment: it is the case for suffixes *-ion* and *-ism*, which prevented the decomposition of words such as *lacrimation* and *hermaphroditism*. 10% of the cases of non-decomposition were due to this. But the French version of DériF is evolving and new affixes are regularly implemented, which means that new rules could be added to the English version quite easily.
- Errors at the preprocessing level (mistagged words), *e.g.*, *corporal/N* was tagged as a noun while being an adjective in our context. 1% of non-decomposed words were mistagged.
- The rest (50%) is composed of miscellaneous cases which cannot easily be categorized into obvious classes.

Incorrect results (see Table 5) were mainly due to:

- Wrong structuring of the decomposition. An example can be seen in Table 5 with the word *meningoencephalitis*. Its correct decomposition should be:

  [[[mening $N*$][encephal $N*$]][itis $N*$]N]

  -itis should be the head of the conjunction of *mening-* and *encephal-*, which would give a definition such as "*inflammation related to head and meninges*."

Let us recall that the linear segmentation of such words into components is correct; this is the level of segmentation adopted by many existing approaches. However, as we aim at a more precise analysis which also includes hierarchical structure, our evaluation is more stringent.

25% of errors were due to this phenomenon. A way to correct the problem would be to examine the semantic type of the components: if they are of the same type, they should be grouped together (coordination). This is the case for *mening-* and *encephal-* which are both labelled as "anatomy", therefore *-itis*, which is labelled as "disease", should apply to both of them together. This modification would be quickly implementable but needs to be experimentally validated;

- Unsatisfactory definition (often due to the fact that the meaning of the word was not sufficiently reflected by the meaning of its parts). See for instance the analysis of the word *acanthosis* in Table 5. The decomposition is right but its meaning has evolved too much and cannot be derived from that of its components (lexicalization). It should nowadays be analyzed as a simplex word. This constituted the largest part of incorrect results (45%). These cases can be addressed by adding those words to exception lists, as done in [4].
- Mistagged words that could not be correctly analyzed. This is the case of *alveolar* (last row of Table 5), which was treated as a noun derived from an adjective (*alveolar/A* is correctly analyzed by DériF). 6% of errors were due to mistagging.
- The remaining 24% is made of diverse cases which are harder to categorize.

The study of the results shows that most errors are due to compounding (order of decomposition between the CFs or lexicalized meaning) rather than derivation rules, which explains why mixed-formation words obtained a slightly better precision than pure compounds. No case of noise (incorrect results) or silence (non-decomposed words) seems to be due to a specificity of the English compounds as opposed to the French compounds.

Table 6 – Evaluation results for Morfessor

Total number of words	1299
Decomposed words (coverage)	1217 (93.7%)
Number of correct results	648
Precision (%)	53.2
Recall (%)	49.9

The analysis of the 100 (non-compound) words sample gave a proportion of 2% decomposed words, which is very low and shows that the system does not produce much noise. An example of a word decomposed by mistake is the word *anomaly* which was decomposed because *ano-* was wrongly identified as the CF meaning *anus*.

Evaluation of the results of the analysis by Morfessor (the tool used in [2]) gave coverage of 93.7%, a precision of 53.2% and a recall of 49.9% (see Table 6). We also measured precision on the set of words for which both tools provided a decomposition, that is 830 words. We obtained 58.7% precision for Morfessor and 78.2% for our transposed version of DériF.

5. Discussion

The precision of the adapted system is good, especially for a first implementation, while recall is lower. Precision for simple linear segmentation is excellent (95%) as is usual for this level of processing. This system goes further by providing a hierarchical decomposition and a definition. This more complete task obtained an overall precision of 78.5%. The system is more precise with derivation (80%) than compounding (76.6%). However the study of the results shows that there is a good potential for quick improvement of the system. 70% of the errors would be easily corrected by adding new words to the list of compounds not to be decomposed (whose meanings are lexicalized) and by modifying some of the rules to take care of errors of structuring in the decomposition. Almost 50% of non-decomposed words would be analyzed if new CFs were added to our list (a rough estimate of 150 new CFs to be added, that is a 15% extension) and if a few more suffixation rules were implemented (this is being done in the French version of the system). The current results are similar to those obtained on French by [16]. In [16] the original French version of the system was applied to a lexicon of biomedical terms and evaluation was performed on a sample of 100 words by two experts (a linguist and a medical expert). They obtained a 77.3% precision for the definition of these terms which comes close to our results. Our transposed system does not generate more incorrect analyses than in the original language. Besides, the analysis of the results did not detect errors due to different mechanisms in the formation of compounds. The potential problems enumerated in the Background section do not seem to have occurred. Although our test corpus is not exhaustive, we conclude that these difficulties are pretty rare and could be handled, if need be, with rules. The present state of the system already shows that this language-dependent system could be successfully adapted to another, related language for the task of analyzing medical compounds.

The advantage of using an existing tool is that we did not have to start from scratch and implement a new system. Indeed, a certain amount of manual work is still necessary, such as preparing the root table and adapting the semantic templates; but there is a stable basis on which to work, so that we believe that this solution is overall less time-consuming. We also believe that using an automatic system as opposed to resorting to a manual approach saves time and presents important advantages. When decomposing and defining words manually, it is hard to ensure full exhaustivity and such a method cannot deal dynamically with the occurrence of new words. Neoclassical compounding is a productive phenomenon and when using morphosemantic analysis in an integrated applicative system (for information retrieval, for instance), it is desirable that such a system be robust to neologisms. Besides, as stated above, we reuse an existing system, which definitely saves time compared to a manual approach.

Using a linguistics-based morphosemantic analyzer such as DériF has a certain number of advantages. The system performs both morphological decomposition and semantic interpretation while other methods remain at the level of morphological segmentation [12] or add semantics after using a tool for decomposition [2]. We also provide a hierarchical decomposition as opposed to a linear one [2,5,11]. The results obtained using the method of [2] gave a coverage far superior to that of DériF, but a much inferior precision (20–25% less) and a slightly lower recall. This means that about the same number of words was correctly analyzed (recall), but DériF was much more to the point by proposing significantly less incorrect analyses. By relying on a statistical segmenter such as Morfessor, the method in [2] has the advantage of being language-independent; however, the implementation used in [2] also relies on a table of morphemes, so transposition to another language is not immediate either. Besides, statistical segmentation also brings certain types of errors that could be avoided with linguistic rules as implemented in DériF, *e.g.*, *chemosis* segmented as *c+hem+osis*. Moreover, as pointed out above, DériF outputs more complete information than Morfessor (simple linear segmentation).

This work also suggests the perspective of a system that could work with several languages. We transposed the system from French to English, but a possible next step would be to have it work on both languages (using the same system to analyse indifferently both French and English compounds) and even use it for translation: producing an English definition of a French word (or vice-versa). This would require a multilingual table of Combining Forms (as suggested in [16] and prepared here for French and English), multilingual semantic templates (as obtained from the present work for these two languages), and translations of words from the lexicon. Such a system could contribute for instance to cross-language information retrieval, with the same principles as [4].

The system could also be used for pharmacovigilance. The decompositions and definitions generated can be used to measure proximity between WHO-ART terms and group similar terms, but without necessarily relying on SNOMED as done in [2].

6. Conclusion

In this work, we successfully transposed a linguistics-based morphosemantic analyzer from French to English in order to provide definitions for English neoclassical medical words. This verifies our hypothesis that neoclassical compounds from different languages can be analyzed in a similar way, as can be expected in the medical vocabularies of languages in the Romance (*e.g.*, French) and Germanic (*e.g.*, English) families.

Summary points

What was already known on the topic?

- Medical language is rich with neo-classical compounds and morphosemantically analyzing those compounds help processes interested in semantics such as generating definitions and detecting semantically similar terms.
- No complete morphosemantic analyzer has been published for English.
- Neo-classical compounds are similarly built in several European languages, so transferring a linguistic analysis from one language to another should be possible with minimum effort.

What does this study add?

- A morphosemantic analyzer for English was produced, thus filling a gap in this area.
- A French analysis of neo-classical medical compounds could be transposed to English which illustrates the similarity of structure of those words and demonstrates the multilingual potential and portability of morphosemantic analysis as regards such compounds.

This work constitutes a first step towards the creation of a multilingual system which could be obtained by applying the method to other languages, a task more or less easy according to the proximity of the languages (French and English being relatively close, we can suppose that new difficulties will arise with less related languages).

Future work includes improvement of the system as well as using the results for a specific application.

REFERENCES

[1] L. Bauer, English Word Formation, Cambridge University Press, Cambridge, 1983.

[2] J. Iavindrasana, C. Bousquet, M.C. Jaulent, Knowledge acquisition for computation of semantic distance between WHO-ART terms, Studies in Health Technology and Informatics 124 (2006) 839–844.

[3] M. Creutz, K. Lagus, K. Linden, S. Virpioja, Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report, Publications in Computer and Information Science, Helsinki, University of Technology, 2005.

[4] S. Schulz, M. Romacker, P. Franz, et al., Towards a multilingual morpheme thesaurus for medical free-text retrieval, in: Proceedings of MIE 99, IOS Press, Ljubliana, Slovenia, 1999, pp. 891–894.

[5] K. Markó, S. Schulz, U. Hahn, Morphosaurus—design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain, Methods of Information in Medicine 44 (2005) 537–545.

[6] C. Iacobini, Composizione con elementi neoclassici, in: M. Grossman, F. Rainer (Eds.), La formazione delle parole in italiano, Niemeyer, Tübingen, 2003, pp. 69–96.

[7] F. Namer, Guessing the meaning of neoclassical compounds within LG: the case of pathology nouns, in: Proceedings of Generative Approaches to the Lexicon, Geneva, 2005, pp. 175–184.

[8] M.G. Pacak, L.M. Norton, G.S. Dunham, Morphosemantic analysis of -ITIS forms in medical language, Methods of Information in Medicine 19 (1980) 99–105.

[9] P. Dujols, P. Aubas, C. Baylon, F. Grémy, Morphosemantic analysis and translation of medical compound terms, Methods of Information in Medicine 30 (1991) 30–35.

[10] S. Wolff, The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding, Methods of Information in Medicine 4 (23) (1984) 195–203.

[11] C. Lovis, P.A. Michel, R. Baud, J.R. Scherrer, Word segmentation processing: a way to exponentially extend medical dictionaries, Methods of Information in Medicine (1995) 28–32.

[12] U. Hahn, M. Honeck, M. Piotrowski, S. Schulz, Subword segmentation: leveling out morphological variations for medical document retrieval, in: Proceedings of the AMIA Symposium, 2001, pp. 229–233.

[13] A.T. McCray, S. Srinivasan, A.C. Brown, Lexical methods for managing variation in biomedical terminologies, in: Proceedings of the Annual Symposium on Computer Applications in Medical Care, 1994, pp. 235–239.

[14] F. Namer, P. Zweigenbaum, Acquiring meaning for French medical terminology: contribution of morphosemantics, in: M. Fieschi, E. Coiera, Y.C.J. Li (Eds.), Proceedings of the 10th World Congress on Medical Informatics vol. 11(Pt 1), San Francisco, CA, 2004, pp. 535–539.

[15] R. Baud, A.M. Rassinoux, P. Ruch, C. Lovis, The power and limits of a rule-based morphosemantic parser, in: Proceedings of AMIA 1999 Annual Symposium, 1999, pp. 22–26.

[16] F. Namer, R. Baud, Defining and relating biomedical terms: towards a cross-language morphosemantics-based system, International Journal of Medical Informatics 76 (2–3) (2007) 226–233 (Epub 2006 Jun 30).

[17] L. Deleger, F. Namer, P. Zweigenbaum, Defining medical words: transposing morphosemantic analysis from French to English, Medinfo 12 (Pt 1) (2007) 535–539.

[18] L. Deleger, F. Namer, P. Zweigenbaum, Analyse morphosémantique des composés savants: transposition du français à l'anglais, in: Proceedings of TALN 2007, Toulouse, June, 2007, pp. 79–88.

[19] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: Proceedings of the International Conference on New Methods in Language Processing Manchester, UK, 1994, pp. 44–49.

[20] F. Namer, Automatiser les définitions des termes médicaux: qu'est-ce que le traitement automatique du langage apporte à la théorie morphologique, in: Proceedings of Journées francophones d'informatique médicale, 2005.