MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Acquiring meaning for French medical terminology: contribution of morphosemantics

Fiammetta Namer[a] , Pierre Zweigenbaum[b]

[a] *ATILF, Université Nancy 2, France*
[b] *STIM, DSI, Assistance Publique-Hôpitaux de Paris & ERM 202 INSERM, France*

Abstract

Morphologically complex words, and particularly neoclassical compounds, form more than 60% of the neologisms in the biomedical field. Guessing their definitions and grouping them into semantic classes by means of lexical relations are thus two crucial improvements for handling these words, e.g., for information retrieval, indexing and text understanding applications. This paper describes a morphosemantic linguistic-based parser called DériF, currently developed in the framework of two projects, UMLF and VUMeF, and its application to French biomedical derived and compound words. It shows how the resulting morphologically tagged lexicon is enriched by semantic relations leading both to the synthesis of pseudo-definitions and to the constitution of classes of synonyms, hypo- and hypernyms.

Keywords:

Natural Language Processing; Semantics; Language; France; Neoclassical Compounds; Morphosemantics; Semantic Relations; Biomedical lexical database.

Introduction

Morphologically complex words (MCWs), and particularly the so-called neoclassical compounds, form more than 60% of the neologisms in technico-scientific domains, and especially in the biomedical field [1]. Morphological analysis, i.e., the process of decomposing a complex word into its constituent parts, has proved useful to avoid the need for costly, repetitive maintenance of specialized dictionaries to account for these new terms [2,3,4]. Constraint-based morphological analyzers, as described in [5], can additionally enrich the decomposition of each word with semantic knowledge. However, previous work such as [4] does not provide structure to the semantic decomposition of morphologically complex words. This limits the precision of this semantic representation and of its usages for medical language processing. On the other end of the spectrum, conceptual representations such as GALEN [6] are much more precise and structured, but require human, knowledge-intensive definition of each concept.

In contrast, the present work performs hierarchical, morphosemantic decompositions of input words. Based on a linguistically sound theory [7], it orders prefixation, suffixation and compounding processes according to categorial and semantic criteria [8], consequently providing a structured decomposition of words along with various types of semantic information. This analyzer, called DériF ("Dérivation en Français")[1] , produces a morphologically tagged lexicon enriched by semantic relations. These relations provide the basis for the constitution of lexical classes (relating a word with its (quasi)synonyms, hypo- and hypernyms, etc.) and for the synthesis of pseudo-definitions. They may be used to improve the contents of thesauri or knowledge bases with new links, which is particularly useful in biomedical informatics. The exposed method and examples focus here on the morphological specificities of medical language, i.e., its massive use of neoclassical compounds, but also work on general-language words.

After exposing the data and the theoretical model we rely on, we describe our morphological analysis method and explain how it generates semantic representations and classes. We then present its application in the framework of projets aiming to develop a bio-medical specialized lexicon for French; finally we discuss the obtained results and their validation.

Material and Methods

Material

We relied on two sources of French words. First, a large French lexicon, the TLF[2] , provided us with reference information for general-language words and some medical words. It contains 80,000 lemmas (uninflected forms). Second, the list of nouns and adjectives found in the French ICD-10[3] was used as a sample of medical words to test our methods. It contains about 6000 derived and compound nouns (N) and adjectives (A). In French, as in many other languages, medical terms often contain non-autonomous combining forms (CFs) [9,10] that may incur both derivation (combining with an affix, i.e., a prefix or suffix, e.g., anhydre) and compounding (combining with lexical forms or other CFs, e.g., hydrofuge). A list of 900 CFs was assembled from two sources. A partial list of such combining forms can be

1. DériF has been designed during the MorTAL project: "Morphologie pour le Traitement Automatique des Langues", 2000-2002, supported by the French Ministry of Public Education, Research and Technology, and coordinated by G. Dal (Silex, CNRS). Extensions to medical language are supported by project UMLF (French Ministry for Research and Education grant #02C0163, 2002--2004) [11].
2. Trésor de la Langue Française, online at www.atilf.fr.
3. Kindly provided by Robert Baud.

obtained from the 984 TLF entries tagged as "forming elements", "prefixes" or "suffixes". Another comes with the ULMS Specialist Lexicon [12]. Additionally, the method and results below also rely on the use of about twenty affixes, the linguistic behavior of which has been compiled from the conclusion of linguistic studies (among them [7;13]).

Linguistic Model

The morphological analysis process starts from a part-of-speech-tagged lemma, as can be found in a term or corpus after automatic tagging, e.g., *désintoxication*/N (N = noun)[1]. It must identify the combining forms which make up complex lemmas. In a derived word, a CF may play the role of base (***pod***iste); in a neoclassical compound, it may play that of head (gastropode) or of *modifier* (*podoencéphale*). These CFs mainly come from Latin or Greek; they can be seen as corresponding to modern nouns (N), adjectives (A) or verbs (V).

At each morphological identification step, the analyzed lemma is semantically paired with its base (or head). In other words, the meaning of a derived word (*décalcifier/V*), or of a compound (*thalassothérapie/N, bactéricide/A)* can always be ob-tained through (a) that of its base (or that of its head and modifier CFs), (b) the involved morphological process, (c) the word's part-of-speech (POS), and (d) the POS of the base (or head and modifier CFs). For instance, *dé-* prefixed denominal verbs can be paraphrased with *Deprive smth/someone from what is referred to by the basenoun*, so that *décalcifier* gets the interpretation "*Deprive smth/someone from calcium*"; A nominal compound with nominal head denotes a hyponymy relation with its head, as illustrated by thalassotherapy, glossed by "*Subtype/part of a therapy characterized by the sea*"; verb-headed compound adjectives (bactéricide) create a predicative relation (realized by the head CF) between the noun they modify and their modifier CF: so, a "agent *bactéricide*" is an agent "*That kills bacterias*".

In addition to these Input-Base semantic Relations (IBRs), CFs within a given conceptual domain may be grouped according to lexical relations: quasi-synonymy, meronomy, see-also. We will see now how to take advantage of these relations to reconstruct definitions for complex word and arrange these words into lexical classes.

Morphosemantic Parsing

Given a morphologically complex word (lemma) and its part-of-speech, e.g., désintoxication/N, acroparesthésie/N, we aim to synthesize a representation of its morphosemantic structure[2] :

(1) **désintoxication/N** => [[[dé [in [toxique A] (er) V]] A]tion N],
(désintoxication/N, désintoxiquer/V, intoxiquer/V, toxique/A)
"(Action|résultat de) de désintoxiquer"
("(Action|result of) detoxicate")

(2) **acroparesthésie/N** => [[acr N*] [para [esthésie N*] N]N]
(acroparesthésie/N, paresthésie/N, esthésie/N*)

"(Partie de|Type particulier de) paresthésie en rapport avec extrémité"

("(Part of| Particular type of) paresthesia related to the extremity")

This representation includes: (i) a structured parse of the input word, where square brackets show which word component associates with which other word part; (ii) the word's corresponding morphological family, made of successively reduced bases (or heads) of the input; and (iii) a gloss describing the Input-Base Relation (IBR)[3] .

Such a structure is obtained by recursively matching word formation rules (WFRs) to the input word. For instance, the WFR "dé- noun-to-verb" we already mentioned looks like (3):

(3) déXiser V --> [dé [X' N] +iser V]

where X is extracted from the input word déXiser, and must match some noun entry in the reference lexicon. WFRs can be divided into suffixation, prefixation, conversion and compounding rules [8]. They impose categorial and semantic constraints to their input word or combining form (CF), and project categorial and semantic constraints on the resulting complex word. As showed in [14,15], these constraints enable WFRs to guess semantics for the input and/or output, even though no semantic knowledge is encoded in the entries of the input lexicon.

The correct order of decomposition must be found for words with both an initial component (prefix or CF) and a final component (suffix or CF): e.g., *désintoxication.* This is dealt with by defining constraints on formation rules (e.g., prefix *dé*-forms verbs, and suffix *-ation* combines with verbal bases), so that a suitable ordering can be computed at parsing time. Some words are actually ambiguous and must obtain multiple parses. For instance, *implantableA = implanter + able (implant + able)* but also *im+plantable (un + plantable)*. The algorithm maintains a list of all valid parses. Finally, word formation rule constraints must account for unknown words. This is performed by including default cases in each rule.

Table 1: Sample from Combining Form table

CF	trad.	POS	Semantic type	Lexical relation
gastr	*estomac* ("stomach")	N	*anatomy*	= *stomac* ← *abdomin* ~ *enter* ~ *hépat*
algie	*douleur* ("pain")	N	*disease*	= *odyn* ~ *ite* ~ *ose*

Combining Forms

The analysis process assumes the availability of a repository of combining forms with associated properties. We compiled a combining form table which lists for each CF its modern language translation, POS, semantic type, and lexical relations with other CFs of the same semantic type (see Table.1). The semantic

1. Though we focus our presentation on medical CFs, the underlying linguistic model indeed also holds for modern, general-language lexical units.

2. A POS followed by "*" means that the CF translation was extracted from the CF lexicon.

3. In this paper, we manually added an English translation of this gloss.

type is taken from a subset[1] of the 15 MeSH tree descriptors (anatomy, organisms, disease, chemicals, etc.). When the CF is a compound head, its type is projected on the compound. Lexical relations here are synonymy (*gastr=stomac*), part-of (*gastr<--abdomin*), hyponymy (*phléb<vascul, ite<pathie*) and see-also[2] (*gastr~hépat*). Each related CF also has an entry in the table.

Reconstructing Semantic Relations for MCWs

The analysis of a morphologically complex word (MCW), as seen above, yields a semantic representation of this word with respect to its base word (its IBR). Recursively substituting IBRs of complex bases leads to the synthesis of a definition of the input complex word. This is the basis for recovering semantic relations between complex words and other words.

Constructed Definitions

For instance, Table 2 shows the definitions obtained for (1) and (2). Whereas Input-Base Relations (IBRs) may refer to complex words (left column, underlined), the expanded defini-tions only contain undecomposable units (right column).

Table 2: From IBRs to reconstructed definifions

Elementary IBR	Definition obtained
désintoxication= «*action/result of* désintoxiquer» désintoxiquer= «*process opposed to that of* intoxiquer» intoxiquer= «*Put a state qualified as toxic*»	désintoxication= *«action / result of the process opposed to that of putting in a state qualified as toxic»*
acroparesthésie= «(Part of particular type of) *paresthésie* in relation with extremity» *paresthésie*= «(Entity / expression close to that of a sensation»	acroparesthésie = "*(Part of | particular type of)* a *(entity | expression) close to that of a sensation in relation with extremity*"

Lexical Relation Computation Rules

The purpose of the lexical relations in Table 1 is to infer lexical relations for MCWs. This is performed through two Lexical Relation Computation (LRC) rules (Table 3). These rules may be reminiscent of those in [16]. Given a parsed MCW, its IBR may be use to detect its quasi-synonyms. Moreover, projecting CF features from the CF table (Table 1) onto a compound word, through the control of the LRC rules (Table 3) allows to link this word to conceptually neighboring terms.

Pseudo-Synonymy

Two complex words A and B are "pseudo-synonyms" when they receive the same definition. This may happen in the following situations where at least one of A or B contains a CF:

- A and B are formed through the same suffix: e.g., the pairs *pétrifier/lithifier "Transform into stone*(=pétr=lith)", *hydrique/aquatique "Relative to water*(=hydr=aqua)"; or through concurrent suffixes, such as *-ique* and *-al* which form *gastrique/stomacal ("Relative to **stomach**")* from gastr=stomac=***stomach.***
- A and B are compounds sharing the same head CF: the *angéite/vasculite* pair and the *lipome/stéatome/adipome* triplet have respectively the IBR *"(Part of|Particular type of) **affection** related to **blood vessel**" and "(Part of|Particular type of) **tumoral pathology** related to **fat**"* (angé=vascul=***blood vessel***, lip=stéat=adip=***fat***).
- A and B are compounds sharing the same modifier CF: both *dermoïde* and *dermiforme* correspond to "*(Part of|Particular type of) **shape** related to **skin***", as derm=***skin*** and forme=oïde=***shape***.

Sometimes, compounds A and B have both formally dif-ferent heads and modifiers, but have equivalent defini-tions . For instance, adjectives *hydrophage/aquavore ("who/that **eats water**"), orthodonte/rectident ("who/that has **teeth** qualified as **right**"), ichtyoïde/pisciforme ("who/that has a **shape** related to **fish**")*.

Table 3: Lexical Relation Computation rules. A, B = compound words, H = Head, Mod = Modifier.

Rule	Examples
Let A = [Mod_A H] and B = [Mod_B H]; If Mod_A R Mod_B and R is ~ or <, then A R B; if Mod_A <--Mod_B, then A < B.	Mod_A: lomb, Mod_B: disc; Mod_A ~ Mod_B =>*lombarthrose ~ discarthrose* Mod_A: phléb, Mod_B: vascul; Mod_A<Mod_B ==> *phlébite < vasculite* Mod_A:méing,Mod_B:encéphal; Mod_A<--Mod_B =>*méningocèle<encéphalocèle*
If A and B are compound words with the respective structures: [$ModH_A$] and [$ModH_B$], where $H_A R H_B$ then A R B.	H_A:ite, H_B: pathie : H_A < H_B ==> *bronchite < bronchopathie* H_A: ome, H_B: matose : H_A ~ H_B ==> *lipome ~ lipomatose*

Lexical Classes

When a compound word is analyzed, all the possible compound words conceptually related to it are identified through (i) relations in the CF table; (ii) lexical relation computation rules; and (iii) if any, pseudo-synonymy. Moreover, it inherits the semantic type of its head CF. For instance, gastralgie is analyzed as follows, with head algie and modifier gastr:

(4) gastralgie/N = [[gastr N*] [algie N*] N],

(gastralgie/, algie/N*)

"*(Partie de|Type particulier de) douleur en relation avec le(s) estomac*"

The lexical neighbors of *gastralgie* are found by: (1) substituting *gastr* with each of its related CFs as stated in the CF table; these CFs are concatenated in turn to *algie*, and the corresponding re-

1. Some of the MeSH chapter heads, e.g. *Geographic Location*, are never instantiated by CFs.
2. That is, at the time being, mainly siblings.

lation (=, <, ~) is deduced by the lexical relation computation rules; (2) performing the same with *algie*, *gastr* being kept invariant; and (3) assigning *gastralgie* the semantic type of *algie* as found in the CF table.

Table 4: possible lexical relations for gastralgie

gastralgie : *disease*
(=:gastr/odyn, ~:gastr/ite, ~:gastr/ose, =:stomac/algie, <:abdomin/algie, ~:enter/algie, ~:hépat/algie)

Table 4 presents the results: semantic type, then a list of semantic relations to possible pairs of modifier/head combining forms. The nouns gastrodynie, gastrite, gastrose, abdominal-gie, enteralgie, hépatalgie are also in our input list. When parsed, their decompositions match the modifier/head pairs in Table 4, so that they are semantically linked to gastralgie. The semantic class of gastralgie is summarized in Table 5. .

Table 5: Semantic class of gastralgie

gastralgie : synonym of gastrodynie (disease)
gastralgie : see also gastrite, gastrose, entéralgie, hépatalgie (disease)
gastralgie : subtype of abdominalgie (disease)

Results

The parsing method and the semantics reconstruction algorithm have been applied so far to analyze 13,971 lemmas from the TLF-based general lexicon, and 2,932 nominal lemmas from the French ICD-10. Currently, more than one third of the 967 inputs in the CF table are labelled with lexical relations. The coverage of the involved WFRs is distributed as follows: suffixes *-able* (implied in the analysis of 1333 MCWs), *-ifier* (156), *-iser* (831), *-ité* (1461), *-eur* (3442), *-tion* (2821) and *-ment* (1921) are fully covered, along with verb-forming pre-fixes *dé-* (1014) and *re-* (992). The same holds for neoclassical composition dealing with the 9844 verb- and noun-headed nouns. Besides these WFRs, the following morphological processes are currently only partially covered: *a-, é-, en-* verb-forming prefixes, *anti-* and *in-* adjective-forming prefixes, *-aie* and *-aille* suffixes, both A-->V and A-->N conversion. Notice finally that among the 2,932 ICD-10 nouns, 2,065 are compounds with more than two CFs, 1,200 are both compounded and suffixed, and 159 are both prefixed and compounded (distinguishing prefixes from CFs is a debated issue we will not address here; see, e.g., the criteria proposed in [9,17]).

A human validation of the linguistic validity of parses was performed for the general-language lexicon : WFRs *-able, -ifier, -iser, -ité, dé-* and *re-* (corresponding to around 80% of the 13,971 lemmas) have been checked both for expressiveness (parsing quality) and robustness (ability to parse unknown words). The validation of derived and compound medical words is ongoing within the framework of project UMLF [11], it involves three tasks: (1) linguistic validation is performed in order to improve the parsing algorithm, (2) CF table validation by biomedical terminology experts who check CF translations and the compatibility between the CF structure and the content of structured terminologies such as MeSH, SNOMED or UMLS, (3) these experts also validate parsing results, especially the IBRs produced. This last task aims to detect (a) lexicalized (or frozen) complex words that should not be decomposed (*arthrose, leucémie*), (b) domain-specific WFRs, which should be added to account for word structures that are absent from general language (e.g., *alpha-*, *beta-*, etc., as in *alpha-foetoprotéine*), and (c) reconstructed definitions which deviate from actual meanings (*pneumothorax* is not a part of the *thorax*, but a disease). A perspective is to use

already available decompositions [1] and existing lexical relations (e.g., found in MeSH) to reduce the amount of human validation.

Discussion

There are evident limits to the presented method, which are related to the use of linguistic constraints: such use is semi-automatic (exception lists have to be maintained), it requires the collaboration of several skills (linguistics, natural language processing, domain experts), and human validation plays a crucial role given the semantic nature of the results. Finally, as implied by the results shown in the previous section, the algorithm foresees incremental WFR development, which implies that only partial coverage is currently guaranteed; to overcome this drawback, a synergy is organized between this method and [18], which is training-based, in order to ensure the largest possible coverage. Other limits can also be observed with respect to the current results. Namely, experts have noticed a distance occurring sometimes between the computed semantic features and the actual meaning of complex words. Consider for instance the pair *angiodilatation* and *vasodilatation*: they share the same IBR "*Particular type of dilatation related to blood vessels*", and thus are considered quasi-synonyms by the system. However, they refer to different entities (pathology vs. medical act) that their equivalent constructed definition cannot account for. This classical discrepancy between meaning and reference will also require some special treatment to be performed in order to assign correct semantic classes and relations.

On the other hand, acquiring semantic knowledge through a linguistic analysis is an advantage this method holds with respect to others. For instance, [4] decompose both general language and neoclassical compound words into their constituent combining forms, and use the resulting 'subwords' for improving information retrieval results. Like [1,2], they produce a flat decomposition; this proved sufficient for indexing with a bag-of-words approach [4]. The present method additionally provides a linguistically-motivated, structured decomposition which is suitable for more precise medical language processing. This decomposition can be compared to that of [16], but has also already been extensively tested on general-language affixes. Using equivalence relations among combining forms can be compared to the correspondences between morphemes in different languages proposed by [3]. Consequently, the work performed here on French neoclassical compounds should be easily reproducible in other European languages, where word formation processes are very similar [17].

Conclusion

This paper outlines a method to perform a morphosemantic analysis of French complex words, among them complex words specific of medical language. Our analysis includes a structural decomposition, a gloss defining the word with respect to its base, and, whenever relevant, its semantic type, together with various lexical relations with its conceptual neighbors. This method has been applied to parse some 3,000 complex nouns from ICD-10. The application of the resulting word decompositions for improved French term matching is planned in the UMLF project and is to be instrumental in the follow-up VUMeF project [19]. The priority for coverage extension is now to extend the algorithm to compound and derived (relational) adjectives: as indicated in [18], these complex words are massively used in medical texts. Another envisaged improvement deals with lexical relations and LRC rules. Current discussions aim to determine the opportunity to create new rules, and to which extent; namely, a new rule will enable new lexical relations between more distant complex words: is it relevant, for instance, to link *gastralgy* to *hepatitis*, where both modifier CFs and head CFs are siblings? A maximal distance must be set to keep within conceptual neighborhood.

Acknowledgments

Many thanks to the members of the UMLF project for useful insights and guidance, especially to Stéfan Darmoni, Robert Baud and Anita Burgun.

References

[1] Lovis C, Baud R, Rassinoux AM, Michel PA, Scherrer JR. Medical dictionaries for patient encoding systems: a methodology. Artif Intell Med 1998;14:201-14.

[2] Lovis C, Michel PA, Baud R, and Scherrer JR. Word segmentation processing: a way to exponentially extend medical dictionaries. In: Greenes RA, Peterson HE, and Protti DJ, eds, Proc 8th World Congress on Medical Infor-matics, 1995:28-32.

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

[4] Hahn U, Honeck M, Piotrowski M, and Schulz S. Subword segmentation: Leveling out morphological variations for medical document retrieval. J Am Med Inform Assoc 2001;8(suppl):229-33.

[5] Daille B, Fabre C, and Sébillot P. Applications of computational morphology. In: Boucher P, ed, Many morphologies. Cascadilla Press, Somerville, MA, 2002:210-34.

[6] Rogers J and Rector AL. GALEN's model of parts and wholes: Experience and comparisons. J Am Med Inform Assoc 2000;7(suppl):714-8.

[7] Corbin D. Morphologie dérivationnelle et structuration du lexique. Presse universitaire de Lille, Lille, 1987.

[8] Corbin D. French (Indo-European: Romance). In: Gooij G, Lehmann C, and Mugdan J, eds, Morphology - An International Handbook on Inflection and Word Formation, (vol1), New York. Walter de Gruyter, 2000.

[9] Warren B. The importance of combining forms. In: Dressler WU and others , eds, Contemporary Morphology, Berlin, New-York. Mouton de Gruyter, 1990:111-32.

[10]Fradin B. Combining forms, blends and related phenomena. In: Doleschal U and Thornton AM, eds, Extragrammatical and Marginal Morphology, München. Lincom Europa, 1999:11-59.

[11]Zweigenbaum P, Baud R, Burgun A, et al. Towards a unified medical lexicon for French. In: Baud R, Fieschi M, Le Beux P, and Ruch P, eds, Stud Health Technol Inform, 2003;95:415-20.

[12]McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminol-ogies. In: Proc Eighteenth Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1994:235-9.

[13]Fradin B. Nouvelles approches en morphologie. PUF, Paris, 2003.

[14]Light M. Morphological cues for lexical semantics. In: Proceedings of the 34th ACL, Santa Cruz, Ca. 1996. 25-31

[15]Namer F. Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. In: Pierrel JM, ed, Proc TALN 2002 (Traitement automatique des langues naturelles), Nancy. ATALA, ATILF, June 002. 235-44

[16]Baud R, Rassinoux AM, Ruch P, Lovis C, Scherrer JR. The power and limits of a rule-based morphosemantic parser. J Am Med Inform Assoc 1999;6(suppl):22-6.

[17]Iacobini C. Composizione con elementi neoclassici. In: Grossmann M and Rainer F, eds, La formazione delle parole in italiano, Tübingen. Niemeyer, 2004. To appear.

[18]Grabar N and Zweigenbaum P. A general method for sifting linguistic knowledge from structured terminologies. J Am Med Inform Assoc 2000;7(suppl):310-4.

[19]Darmoni SJ, Jarrousse E, Zweigenbaum P, et al. Extending the French part of the UMLS. In: Musen M, ed, Proc AMIA Symp 2003, Washington, DC. AMIA, November 2003:824 (poster).

Address for correspondence

Fiammetta Namer

ATILF/Université Nancy2 - CLSH - BP3397 -

54015 Nancy Cédex - France

email:fiammetta.namer@univ-nancy2.fr

URL:http://www.univ-nancy2.fr/pers/name