Semantic Handling of Medical Compound Words through Sound Analysis and Generation Processes

Anne-Marie Rassinoux, Ph.D., Patrick Ruch, M.S., Robert H. Baud, Ph.D., Christian Lovis, M.D.
Medical Informatics Division, University Hospital of Geneva, Switzerland

Compound words are frequently encountered in the medical domain. Their conciseness complies with the telegraphic style usually adopted by clinicians in daily practice. This amplifies the need for clarifying their semantic interpretation and representation through respectively the analysis and generation processes. While highlighting the peculiarities of medical compound words, this paper shows how model-driven linguistic tools accurately deal with the compositionality of medical language. These statements are illustrated by means of examples, stemming from the handling of surgical procedures as part of the GALEN-IN-USE project.

INTRODUCTION

Medical language makes extensive use of morpho-semantic constituents, also called morpho-semantems as they denote morphemes that carry units of elementary medical meaning. Their combinations into compound words allow a high density of information to be encapsulated in a short and succinct manner. This type of concise style is commonly adopted by health care professionals in their daily practice, as they are often facing a limited amount of space and time for communicating pertinent information. The exchange of medical compound terms is also promoted across linguistic borders, as they are constructed from a substantial number of Greek and Latin roots and affixes that are shareable between different languages. This multilingual consistency contributes to reducing the endeavors for managing compound word forms in various national languages. The morphology of medical language has already been widely investigated, with a view to automatic indexing and classification of medical documents. The automated handling of compound word forms is subordinate to the existence of large and fairly detailed dictionaries of morpho-semantic constituents. More recently, interest has grown in the automatic acquisition of multilingual sources of morphological knowledge, for subsequent use in natural language processing and information retrieval. In particular, recent publications by the authors of this paper have focused on extensive processing of morpho-semantems via a rule-based parser of medical expressions. This paper goes one step further in the semantic interpretation and representation of morpho-semantems, by considering side-by-side, the deep analysis and generation of medical compound words.

BACKGROUND

In Europe, a common reference model, for the domain of surgical procedures, has been developed as part of the GALEN-IN-USE project. In order to ensure the quality and coherence of national classifications that are represented in this model, the multilingual analysis and generation of surgical procedures have been endorsed. In particular, the French government has adopted these natural language tools (delivered by the authors), for the assessment of a new French coding system for surgical procedures named CCAM (a French acronym for Classification Commune des Actes Médicaux). For this, the semantics of individual surgical procedure rubrics is represented by a description logic that allows concepts to be linked through relationships. This representation is then directly translated into the formalism of conceptual graphs that serves as knowledge representation in the linguistic tools described in this paper. In order to conceal, from clinical end-users, the complexity of such formal and compositional representations, natural language expressions, which reflect the semantics of these structures, are utilized in two ways. On the one hand, the analysis of surgical procedure rubrics supports the task of producing relevant semantic representations. On the other hand, the generation of multilingual natural language expressions from these structured representations facilitates their interpretative reading. As compound word forms are frequently used to denote surgical procedures, they must be taken into account during the analysis process. Besides, the strategy adopted for generation is to produce noun phrases "as detailed as necessary but as concise as possible". This means that a precise and direct language style is preferred to a wordy one for expressing the given semantic structures. This confers full priority on generating compound terms, when allowable, in the target language.

The purpose of this paper is to obtain a better insight into the way medical compound words, including...
compound nouns as well as compound adjectives, can be automatically handled by model-driven linguistic tools, while being aware of cases specific to a given linguistic task. Comparisons with the processing of full sentences are outlined whenever appropriate. The development of the French CCAM classification constitutes the ground for numerous examples. These are translated into English for the sake of clarity for this paper. This type of language-shift does not question the relevance of portrayed examples, thus strengthening the multilingual principles of medical compound word construction, at least for most of European languages.

**SEMANTIC INTERPRETATION AND REPRESENTATION OF COMPOUND WORDS**

The analyzer and generator presented in this paper, and currently used to handle surgical procedures, are based on the GALEN concept model. This means that the meaning of any medical expression formulated in a given language, can be checked against the semantics of the treated domain. The latter affords a multiple-inheritance typology of concepts and relationships that supports compositional definitions. For this, sanctions, also called "sensible statements" in the GALEN project, allow the correctness of any combination of pairs of concepts to be checked. This model-driven approach appears also as relevant for the treatment of morpho-semantics, which is commonly compared to the processing of short sentences. Indeed, compound words, as well as noun phrases, are frequently used in the medical domain, as they denote circumscribed and significant units of medical meaning. However, some peculiarities are worth considering.

**Analysis of Medical Compound Words**

The analysis of compound words consists of two steps. First, it has to identify the different morpho-semantic constituents, i.e. to segment compound words into prefixes and suffixes. Second, the semantic relationships occurring between these constituents must be determined in order to provide the resulting semantic interpretation. The first step relies on a segmentation algorithm (that commonly applies the principle of the longest match, starting from right to left in English) and on the existence of morphemic dictionaries that associate one concept to each morpheme (or more than one in the case of semantic ambiguity). The second step requires a description of the semantics of the treated domain in order to clarify the correct relationships occurring between the highlighted conceptual entities. It is commonly admitted in the medical linguistic community that both syntactic and semantic information is useful to achieve a deep understanding of medical language. In the case of compound words, the syntax is limited to the linear order in which morphemes are concatenated as well as to the categorization of these as prefixes and suffixes. This last point is of paramount interest for semantic interpretation, as it is generally assumed that the suffix constitutes the head of the semantic representation around which the other prefixes will be connected. Moreover, the absence of functional particles (such as prepositions or conjunctions) that are useful to determine the kind of relationships existing between two entities, renders difficult the arrangement of semantic entities that is only implicitly reflected by the linear order of morphemes. Therefore, the final interpretation strongly depends on semantics.

The GALEN sanctions that represent semantic co-occurrence patterns, accurately define the combinatorial semantics of the treated domain. They are applied by the analyzer for setting the semantic interpretation of compound words. For instance, the analysis of the compound word *craniotomy* first identifies the prefix *crani* and the suffix *tomy*, which are respectively assigned to the concepts *Skull* and *Incising* in the dictionary. Second, the sanctioning rule *(Incising, actsSpecificallyOn, SolidStructure)* is triggered, and by inference in the concept hierarchy, the relation *actsSpecificallyOn* is selected to link the concept *Incising* with *Skull*. However, the degree of granularity of these semantic sanctioning rules is not always suitable for solving semantic ambiguities. For example, the sanctioning rule *(Pain, actsSpecificallyOn, BodyStructure)* is not specific enough to disambiguate the words *mastodynia* (*masto + odynia*) or *cystalgia* (*cysto + algia*). Indeed, both *odynia* and *algia* are suffixes that denote pain. The prefixes *masto* and *cysto* are semantically ambiguous as they both carry two meanings, which are respectively portrayed by the concepts *Breast* or *MastoidProcess* on the one hand, and *UrinaryBladder* or *Cyst* on the other. As all these concepts have *BodyStructure* as one of their ancestors in the GALEN multiple-inheritance hierarchy, the above sanctioning rule cannot decide between the different meanings. One solution to bypass such a discrepancy is to distinguish in the morphemic dictionary the meanings that are most frequently encountered in medical compound words. In the above example, the concepts *Breast* for *masto* and *UrinaryBladder* for *cysto* are marked as those preferred in the dictionary. This implies that compound words that make use of the other meanings have to be directly enumerated in the dictionary, in their full form. For instance, the compound word
cystogastrostomy should be directly specified as an English annotation of a composite concept, the meaning of which expresses an internal drainage of an adjacent cyst into the stomach. Finally, as in natural language processing, general ambiguities that require additional contextual information to be solved, also exist at the level of compound words. This is the case for ulectomy (ulo + ectomy), which means either an excision of scar tissue or an excision of the gingivae. The latter meaning can be unambiguously expressed through the compound word gingivectomy, thus encouraging the use of the unambiguous prefix gingivo instead of ulo for denoting the concept Gingiva.

Generation of Medical Compound Words

The generation of compound words also consists of two steps. First, each adjacent concept in the input representation, for which a morphological unit exists in the dictionary, is elected. Second, the set of selected morphological constituents must be concatenated, according to a correct order, to yield the corresponding compound word. These two steps strongly rely on the input structured representation given to the generator. The latter structure is assumed to describe, in an unambiguous and accurate way, the nested combinations of concepts. It follows that the head concept of the selected structure requires a suffix, whereas the other concepts are systematically represented by prefixes. Moreover, when the same relationship occurs between the head concept and the other conceptual entities, the sequential order in which these entities are enumerated in the representation, is selected for concatenation (see example (1) extracted from the otorhinolaryngology chapter (Orl) of the CCAM classification). Otherwise, the order is fixed by the combinatorial semantics of the various relationships. This is shown in example (2) that describes an abnormal passage communicating from the rectum to the urinary bladder to the vagina. An extensive description of the various morpho-semantic distribution patterns and their associated semantic interpretation, as encountered in the context of surgical procedures, is given by Norton & Pacak.

Compared to the generation of noun phrases, the syntax of generated compounds is simplified, as no agreement in gender and number is required. The only checking concerns the elision of a final vowel when the following morphological constituent also starts with a vowel. For example, concatenating osteo plus itis leads to osteitis, likewise, osteo plus ectomy leads to ostectomy, as duplicate vowels are also removed. Moreover, the choice of inserting a hyphen to mark the morphological units, does not follow any universal rule, and thus is essentially a question of preference. In the generated compound words, it is basically inserted for the sake of readability when more than two prefixes are present. For instance, the generation of glosso-pharyngo-laryngectomy will be preferred to the whole compound word glosso-pharyngo-laryngectomy.

Finally, whereas the structured input given to the generator defines in an unambiguous way the semantics to be expressed, the question of choosing among different synonyms also appears at the level of morpho-semantics. Indeed, several medical concepts can be expressed with more than one morpheme that can be interchangeable or not. For instance, both vagino and colpo annotate the concept Vagina and are indifferently used to generate an inflammation of the vagina (e.g. colpitis or vaginitis), an excision of the vagina (e.g. colpectomy or vaginectomy), a pain of the vagina (e.g. colpalgia, colpodinia or vaginodynia), etc... It is worth noting that the English term vaginalgia, while semantically correct, is not commonly used to express a vaginal pain. Such preferences are difficult to apprehend and, in addition, are inconsistent between languages. The previous example is preferably phrased in French by the compound words colpodynie and vaginodynie, whereas in Italian the words colpalgia and colpodinia are utilized. This results in awkward forms (i.e. semantically correct but never used in standard language) being created through the automated generation of compound words. For example, the noun spondylopexy (respectively spondylopectie in French) was generated as a literal interpretation for "fixation of vertebrae". While neologisms can be discovered in this way, the forms that are deemed clumsy from a medical point of view can always be explicitly maintained in a list of exclusions.

<table>
<thead>
<tr>
<th>COMPOUND WORDS VERSUS COMPOSITIONAL MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the GALEN model, each composite concept has a corresponding semantic definition expressed with more elementary concepts. Besides, the semantic interpretation of compound terms, which is built from</td>
</tr>
</tbody>
</table>
morphy-semantic constituents, can also be regarded as semantic definitions that are expressed and maintained at the conceptual level. This establishes a strong connection between compound semantic concepts and compound words, thus leading to jointly consider the compositionality of the domain model with the morphological processing of compound words. However, this interdependence must be carefully tuned, both at the linguistic and conceptual levels.

**Concept Annotation and Concept Definition**

The strategy adopted, is to keep the size of dictionaries minimal, while ensuring smooth lexical variations. For this, static lexical expressions that can be automatically handled through linguistic tools, are systematically excluded from dictionaries. For instance, the GALEN composite concept Nephrectomy is deliberately not annotated in the various national dictionaries. However, the English expressions "nephrectomy", "excision of the kidney" or "renal excision" (respectively "néphrectomie", "excision du rein" or "excision rénale" in French) are automatically recognized and generated. This is performed thanks to the semantic definition associated with the concept Nephrectomy, which describes an "Excising which actsOn Kidney". However, sometimes the compositionality of the model does not fit with the compositionality of the language. Let us consider the CCAM example (3).

In the above original CCAM rubric, both the adjective outer and the morpheme cantho are used at the linguistic level to reflect the meaning of the concept OuterCanthus. These two lexical units do not constitute a well-formed annotation for the whole concept OuterCanthus, which is more correctly annotated by the English expression "outer canthus". The lack of definition for this concept in the GALEN model as is, prevents the generation of compound terms. Adding a definition that dissects the concept OuterCanthus as being a "Canthus which hasInnerOuterSelector outersclection" allows the annotation process to be redirected from the composite concept OuterCanthus to its primary constituents. The different English lexical expressions, as shown in (3), thus also become manageable by linguistic tools. This example also emphasizes the fact that the full interpretation of certain medical compound words requires the analysis scope to be extended to the surrounding words. For instance, analyzing the following expressions "a mammary capsulotomy", "an urethral meatotomy" or "a temporomandibular arthroscopy" requires the adjective mentioned to be linked, at the linguistic level, to a morpho-semantic constituent that is embedded in the adjacent compound word. At the conceptual level, semantic definitions are needed in order to represent the above expressions by the concept Incising, that acts respectively on the compound semantic concepts CapsuleOfBreast, UrethralMeatus, and TemporoMandibularJoint.

**Linguistic and Conceptual Idiosyncrasies**

The modeling style adopted for representing information at the conceptual level has strong implications when shifting to the linguistic level. In particular, linguistic tools must be aware of abstract constructions that are used at the conceptual level. An illustration is given in example (4) where the abstract entities ArbitraryBodyConstruct and hasArbitraryComponent must be concealed at the linguistic level. Indeed, they only reflect the GALEN modeling style used to conceptually clarify the various body parts on which the inspection occurs.

In the above original CCAM rubric, both the adjective outer and the morpheme cantho are used at the linguistic level to reflect the meaning of the concept OuterCanthus. These two lexical units do not constitute a well-formed annotation for the whole concept OuterCanthus, which is more correctly annotated by the English expression "outer canthus". The lack of definition for this concept in the GALEN model as is, prevents the generation of compound terms. Adding a definition that dissects the concept OuterCanthus as being a "Canthus which hasInnerOuterSelector outersclection" allows the annotation process to be redirected from the composite concept OuterCanthus to its primary constituents. The different English lexical expressions, as shown in (3), thus also become manageable by linguistic tools. This example also emphasizes the fact that the full interpretation of certain medical compound words requires the analysis scope to be extended to the surrounding words. For instance, analyzing the following expressions "a mammary capsulotomy", "an urethral meatotomy" or "a temporomandibular arthroscopy" requires the adjective mentioned to be linked, at the linguistic level, to a morpho-semantic constituent that is embedded in the adjacent compound word. At the conceptual level, semantic definitions are needed in order to represent the above expressions by the concept Incising, that acts respectively on the compound semantic concepts CapsuleOfBreast, UrethralMeatus, and TemporoMandibularJoint.

**Linguistic and Conceptual Idiosyncrasies**

The modeling style adopted for representing information at the conceptual level has strong implications when shifting to the linguistic level. In particular, linguistic tools must be aware of abstract constructions that are used at the conceptual level. An illustration is given in example (4) where the abstract entities ArbitraryBodyConstruct and hasArbitraryComponent must be concealed at the linguistic level. Indeed, they only reflect the GALEN modeling style used to conceptually clarify the various body parts on which the inspection occurs.

In the above original CCAM rubric, both the adjective outer and the morpheme cantho are used at the linguistic level to reflect the meaning of the concept OuterCanthus. These two lexical units do not constitute a well-formed annotation for the whole concept OuterCanthus, which is more correctly annotated by the English expression "outer canthus". The lack of definition for this concept in the GALEN model as is, prevents the generation of compound terms. Adding a definition that dissects the concept OuterCanthus as being a "Canthus which hasInnerOuterSelector outersclection" allows the annotation process to be redirected from the composite concept OuterCanthus to its primary constituents. The different English lexical expressions, as shown in (3), thus also become manageable by linguistic tools. This example also emphasizes the fact that the full interpretation of certain medical compound words requires the analysis scope to be extended to the surrounding words. For instance, analyzing the following expressions "a mammary capsulotomy", "an urethral meatotomy" or "a temporomandibular arthroscopy" requires the adjective mentioned to be linked, at the linguistic level, to a morpho-semantic constituent that is embedded in the adjacent compound word. At the conceptual level, semantic definitions are needed in order to represent the above expressions by the concept Incising, that acts respectively on the compound semantic concepts CapsuleOfBreast, UrethralMeatus, and TemporoMandibularJoint.

**Linguistic and Conceptual Idiosyncrasies**

The modeling style adopted for representing information at the conceptual level has strong implications when shifting to the linguistic level. In particular, linguistic tools must be aware of abstract constructions that are used at the conceptual level. An illustration is given in example (4) where the abstract entities ArbitraryBodyConstruct and hasArbitraryComponent must be concealed at the linguistic level. Indeed, they only reflect the GALEN modeling style used to conceptually clarify the various body parts on which the inspection occurs.
"osteotomy" (i.e. "ostéotomie orbitaire" in French).

RESULTS

The assessments, through the linguistic tools, of more than 5500 surgical procedures belonging to the French CCAM classification, have already been described in detail. Only the automated handling of compound words is considered here. Given the CCAM urology chapter, it currently contains 400 original rubrics expressed with 534 different words. Among these, 147 are compound words and are utilized in more than 67% of rubrics. In 46% of the cases, these compound words are nominal, the others being compound adjectives. The generation process performed from the GALEN semantic interpretations has reproduced the same compound words in 58% of the cases. In 16% of additional cases, a different compound term was generated and in the remaining 26%, no compound terms were built. This was essentially due to lack of definitions in the GALEN model, thus hampering the use of individual medical morphemes. Among the 1270 affixes (i.e. 1101 prefixes and 169 suffixes) currently described in our dictionaries, a subset of 240 (i.e. 22 suffixes and 218 prefixes) have been effectively used for generating compound words denoting surgical procedures. These morphemic dictionaries are simultaneously maintained for French, English and Italian.

CONCLUSION

This paper has described the added values of model-driven understanding tools for the automatic handling of medical compound words denoting surgical procedures. The success of this approach is based on the ability to manage, in parallel, the peculiarities of medical language and the characteristics of compositional modeling. This tuning is the price to pay for a thorough interpretation and representation of medical compound words. In return, the benefit, on an international basis, is obvious, as such techniques can be applied to all languages that share the same set of Greek and Latin constituents.

Acknowledgments

This work was funded by the Swiss government (OFES - Office Fédéral de l'Education et de la Science) as part of the European GALEN-IN-USE project.

References