A Medical Multilingual Information Retrieval

Edson José Pacheco[23], Percy Nohama[23], Stefan Schulz[1], Kornél Markó[1]

[1]Freiburg University Hospital, Department of Medical Informatics, Freiburg, Germany
[2]Paraná Catholic University, Health Informatics Laboratory, Curitiba, Brazil
[3]CEFET-PR, Graduate Program in Electrical Engineering and Industrial Informatics, Curitiba, Brazil

Abstract. *The Web is full of documents and resources. Users employ different strategies to find information they need: by browsing, using search engines, by following existing categories in a Web catalog. For technical sublanguages such as the medical one, document indexing based on lexical entities at a subword level has proved useful. However, it still remains challenging to identify and to delimit the meaningful lexical entities, as well as to group them in synonymy classes. We present a lexicographic and semantic foundation underlying the multilingual* MORPHOSAURUS *lexicon.*

Resumo. *A Web é repleta de documentos e outros recursos. Usuários utilizam diferentes estratégias para encontrar as informações que eles necessitam: navegando entre sites, usando máquinas de busca ou usando catalógos de domínio. Para linguagens técnicas como a linguagem médica, a indexação de documentos usando entidades lexicais em nível estruturante menor que a palavra tem se mostrado útil. Porém, ainda há desafios com relação à indentificação e delimitação de entidades lexicais apropriadas, assim como ao agrupamento em conjuntos de sinônimos. Apresentamos os fundamentos lexicográficos e semânticos do léxico multilíngue* MORPHOSAURUS.

1. Introduction

The problem of Information Retrieval is that users have to spend a lot of time and effort to navigate but may not find the information required, and sometimes, the information is lost during the document navigation process. In the medical context, the problem is exacerbated because the information sources are too numerous. Various works are related to medical information retrieval, normally supported by a full or a semi-automated indexation of documents. However, these are works aiming only the acquisition of information and not the evaluation of these data. In this work, we want to treat efficient recovery of medical information, allowing the recuperation in many formats (like medical documents, guidelines, knowledge representations, etc), extracting the knowledge guiding users when they browse and search information, represents the meaning elements of the textual sources by means of controlled vocabularies, thesauri and ontologies, based upon a rich inventory of biomedical terminologies such as provided by the UMLS, OBO and others. To do that we need a multilingual thesaurus, in our case a medical domain thesaurus, to process documents wrote in one language and builds documents without language dependence (using an artificial language). To specify an artificial language is necessary deeply knowledge about syntactic and semantic structure of human languages. The conventional view on human language is word-centered, at least for written language where words are clearly delimited by spaces. It builds on the hypothesis that words are the basic building blocks of phrases and sentences. In syntactic

theories words constitute the terminal symbols. To break down natural language to the word level appears, therefore, straightforward. When we look at the sense of natural language expressions, however, we find much evidence that semantic atomicity frequently does not coincide with the word level. As an example, in the English term *high blood pressure* the word limits reflect quite well the semantic composition, whereas this is not the case in its literal translations *verhoogde bloeddruk* (Dutch) or *bluthochdruck* (German). Especially in technical sublanguages we encounter atomic senses at different levels of fragmentation or granularity. An atomic sense may correspond to word stems (e.g., *hepat-*), prefixes (e.g., *anti-*), suffixes (e.g., *-logy*), larger word fragments (*hypophys-*), words (*spleen*) or even combination of words (*yellow fever*). The possible combinations of these word-forming elements are immense and ad-hoc term formation is common. As a consequence, a high coverage of a domain-specific lexicon can only be expected if lexical units are restricted to units of atomic sense, which then can be used as building blocks for composed terms at any level of granularity. Extracting atomic sense units from texts in order to achieve a basis for cross-language semantic document indexing is an important goal for many applications in the fields of information extraction, text mining and document retrieval [Schulz and Hahn 2000]. The latter is the main application context of the MORPHOSAURUS system[1] which builds upon a multilingual lexicon of semantically atomic lexical units covering the domain of medicine. In the following we will give a semi-formal account to lexical atomicity as the theoretical basis of the MORPHOSAURUS subword lexicon. We will then turn to an empirically founded scheme for the delimitation of words and lexical items at a sub-word level. Our application domain is medicine; we use examples in English, Spanish, and Portuguese language. Finally, we will present the MORPHOSAURUS lexicon and its lexicographic guidelines as a concrete instance of the implementation of our theory.

2. Semantic Atomicity

We here introduce the notion of "semantic atomicity" which will guide our further argumentation in this article.

A sequence of characters is semantically atomic if the sense conveyed[2] (in a given language and a given domain context) is not univocally derivable from the sense of its constituents. In linguistic terms, the constituents of words are morphemes, and they are tied together by word-forming operations such as inflexion, derivation and composition. *Inflexion* conveys number, gender, tense, or aspect information, thus combining the lexical sense of the word stem with the grammatical function of the affix. *Derivation*, instead, covers different phenomena. A derivational affix may simply affect the part of speech without any semantic implication (*patient with a severe injur-y = severe-ly injur-ed patient*). Or it may add an additional sense, such as *hepatitis = hepat* (liver) + *itis*[3] (inflammation). However, cases in which the derived form has gained sense of its own are frequent. For instance, *neurosis* is the result of linking *neur* (nerve) with *osis* (disease). However, the sense of *neurosis* is not really a disease of nerves (at least in modern scientific medicine). As a consequence, the derivation *neurosis* would be considered an atomic lexical unit. (Single-word) *composition*, finally, combines two or more

[1] http://www.morphosaurus.net

[2] We understand by the *sense* of a linguistic expression the mental construction associated with this expression, in contrast to the words' referents (concrete objects in the world) [Eco et al. 1988].

[3] cf. discussion of related work on domain-specific suffixes in [Schulz and Hahn 2000]

stems in one and the same word. It is a very frequent phenomenon in Germanic languages, but also in technical sublanguages where words like *adenosintriphosphat*, *prebetalipoproteinemia*, *osteoartrose*, *inmunodeficiencia*, referred to as "neoclassical compounds" [McCray et al. 1988], are common.

Lexical units may have multiple senses (homonymy, in a broad sense); and one sense can be expressed by different expressions (synonymy). Although domain specific terminologies are constructed in order to control the use of a specialized language and to avoid ambiguous expressions, non-standardized terminology is widely used in any domain. For instance, *molar* has a completely different sense in obstetrics (*molar pregnancy*) than in lab medicine (*molar mass*), or in dentistry (*fractured molar*). *Head* has a different sense in *headache* than in *head of femur* or *head of department*. *Operation* means "surgical procedure" in a medical domain, opposed to different senses in mathematics or business. In such cases, the local context (the surrounding words) generally helps us select the right sense. Furthermore, the restriction to a well-defined domain (e.g. clinical medicine, in our case) allows us to ignore word senses which are definitely outside that domain (e.g. the sense *head* as the role of a word in grammar theory).

Besides ambiguity, lexical units may have overlapping senses. Quasi-synonymy relations can hold between terms of different language (*caput*, *head*) or different levels of erudition (*belly*, *abdomen*). Complete identity in sense (true synonymy) which holds throughout all possible uses of a word is rare. If we want to establish classes of synonymous expressions we have to make, firstly, a clear commitment to the environment in which the expressions are considered synonymous, *viz.* what we call the *domain context*, and secondly, convene upon a tolerance in sense deviation which is still compatible with the formal properties of an equivalence relation[4]: If we agree on considering *disease* a synonym of *illness* and *illness* a synonym of *sickness*, then *disease* and *sickness* are synonyms, as well. The tolerance depends also on the relevance of subtle sense distinctions in the chosen domain context. In the domain of clinical medicine, e.g., *neoplasm-*, *cancer*, *carcinom-* would hardly be considered synonyms but a different decision may, however, be taken in another domain. A counterexample would be to create an equivalence class {*excis-, exstirp-, remoc-, -ectom-*} in a domain of general medicine, neglecting subtle distinctions of surgical technique. Translation is a special case of synonymy in which words of different languages are linked. Here we can define equivalence classes, as well, e.g. {*disease, illness, enfermedad, doença*}. Not only the *grouping* of lexical units into synonymy classes, but also their proper *delimitation* depends on the domain context. *Leukemia*, e.g., literally means "white blood", and *neurosis* literally means "nerve disease". This may be plausible in a historic medical context, but it provides an incomplete description when related to modern medicine. Thus, a composite sense may be ascribed in the historic context, and an atomic one in the present one.

In order to represent atomic senses of lexical units we define a semantic layer, which contains language-independent identifiers, so called MIDs (**M**orphoSaurus **ID**s). MIDs can be roughly compared to concepts in thesauri (such as CUIs in the UMLS metathesaurus [UMLS 2004] or to synsets in WordNet [Fellbaum 1998])[5]. However, there

[4]reflexivity, transitivity, symmetry

[5]In terms of notation, MID will be represented by the composition of the # sign with one of its non-ambiguous English lexemes, e.g. #liver = {*hepar, hepat, liver, figad, higad* } or #caput = {*caput, cabec,*

are two major differences between MIDs and UMLS CUIs or WordNet synsets: Firstly, MIDs can represent disjunctions of different senses. This is the case when ambiguous lexical units are addressed. To take the above example, the disjunction of the different senses of *molar* is represented by one MID, and each of the non-ambiguous senses by another MID, each. Secondly, all lexical units which are assigned to one MID must be fully interchangeable. For example, {*head, caput, cabec, cabez, cefal, cephal* } would not be a proper representation of one MID, since *head* has additional senses, at least in a domain context which includes the meaning of *head* as "person in charge of sth.".

A different view on MIDs is to regard them as non-ambiguous words of an interlingua, since each synonym class is uniquely identified by one MID. This perspective emphasizes our preference of representing lexical meaning abstracting away from the variety of human language, an exercise that must not be mistaken for the construction of a domain ontology (cf. [Hirst 2004]).

We now introduce the notion of a subword as the minimal meaning-bearing constituent of a domain-specific term. Its defining property is that its sense is not composite. This rules out, for instance, to consider *hepatitis* a valid subword because its sense can be derived from its constituents, in contradistinction to, e.g., *hypophysis* (composing the sense of its components *hypo* and *physis* does not lead to the proper sense of *hypophysis*), i.e. *hypophysis* is semantically underdeterminate. For each subword there exists at most one MID where the assignment of the MIDs depends on the domain context d and the language under consideration i. If no meaning is assigned to a subword, it is a *stop entry* (it has only a grammatical function), such as auxiliary verbs or inflection endings. The relation between lexical unit, sense, domain context[6] and language can therefore be expressed by the quadruple (LU, MID, D, L). Let us now consider some typical examples:

- (l_1, m, d, i), (l_2, m, d, i), (l_3, m, d, i)
  l_{1-3} are synonyms in domain d and language i since they refer to the same MID m. Example: *nephr-*, *ren-*, *kidney*
- (l_1, m, d, i_1), (l_2, m, d, i_2)
  l_1 in language i_1 is the translation of l_2 in language i_2 in domain d which is expressed by the reference to the same MID m. Example: *nephr-*, *riñon*.
- (l, m_1, d, i), (l, m_2, d, i)
  l has the two senses m_1 and m_1 in domain d and language i. Example: *head* refers both to body parts and to persons who are in charge of something.
- $(l, , d, i_1)$, (l, m_2, d, i_2)
  l is a stop entry in language i_1 and it has the sense m_2 in language i_2. Example: *era* is an auxiliary verb form in Spanish and Portuguese and a noun in English.
- (l_1, m_1, d_1, i_1), (l_2, m_1, d_1, i_1), (l_1, m_2, d_2, i_1), (l_2, m_3, d_2, i_1)
  l_1 and l_2 are synonyms in language i_1 and domain d_1 but not in domain d_2.
  Example: *sildenafil* and *viagra* can be considered synonyms in clinical medicine but not in the context of pharmaceutical industry.

MIDs can be linked by two lexical relations, *viz.* the horizontal (syntagmatic) relation *expands-to* , and the vertical (paradigmatic) relation *has-sense*:

cabez, defal, cephal}.

[6]We will not need an elaborated theory of domain contexts for the following examples. For a detailed discussion cf. [Buvač et al. 1994].

- The relation *expands-to*($m_0, [m_1, m_2, ..., m_n]$) relates a MID m_0 to an ordered list of MIDs (at least 2 elements). This relation is used in order to make a hidden semantic compositionality explicit. Example: The MID assigned to the lexical item *short* is is expanded to the sequence of the MID representing {*length, longitud, comprimento*} and the MID representing the meaning of "high value". The relation *expands-to* is also used to deal with composed meanings in compounds which exhibit omission of characters, e.g. *urinalysis* (see below).
- The relation *has-sense*($m_0, \{m_1, m_2, ..., m_n\}$) relates an ambiguous MID to a set of MIDs (at least 2 elements). This relation is used to relate an ambiguous MID to each of its (non-ambiguous) senses. Example: The MID assigned to the ambiguous word *head* is related via *has-sense* to the non-ambiguous MIDs for "upper part of the body" and "person in charge of sth.".

Both relations are transitive. Insertions into lists or sets create expanded lists or sets, not nested ones, e.g.:

- *expands-to*($m_0, [m_1, m_2]$) & *expands-to*($m_1, [m_3, m_4]$) is equivalent to *expands-to*($m_0, [m_3, m_4, m_2]$)
- *has-sense*($m_0, \{m_1, m_2\}$) & *has-sense*($m_1, \{m_3, m_4\}$) is equivalent to *has-sense*($m_0, \{m_3, m_4, m_2\}$)

Cycles are not allowed. A set of inter-MID relations is called normalized if all possible substitutions are realized. A set of quadruples, together with a set of inter-MID relations defines a multi-context multilingual dictionary $\mathcal{D}$. Other than in many thesauri such as the UMLS [UMLS 2004] or WordNet [Fellbaum 1998], we do not define semantic relations between equivalence classes such as hypernymy, hyponymy, mereonymy etc. Encoding these richer relations is left to domain thesauri or ontologies such as MeSH [MESH 2004] or SNOMED CT [sno 2004]. MIDs can be linked to external vocabularies or ontologies by the following triple:
(*MID*, *ONT*, *EID*+). *ONT* is the identifier of the external source, *EID* is the identifier of the term / class / concept of the external source (conjunctions of identifiers are possible). If the *MID* is ambiguous with regard to the external vocabulary, there will be one record for each *EID*+ per *MID*.

3. The MORPHOSAURUS Lexicon

In the following, we describe a concrete implementation of the lexicon model as introduced above, *viz.* the structure of the MORPHOSAURUS lexicon, a multilingual subword repository covering the domain of clinical medicine. The MORPHOSAURUS lexicon provides the data base for the MORPHOSAURUS indexer, a tool which extracts meaningful items from texts and maps them to MIDs, resulting in a language-independent abstraction of text contents. The MORPHOSAURUS lexicon, so far, does not manage multiple contexts. Rather it is committed to one, well-defined domain context, *viz.* clinical medicine. We introduce further specifications and conventions which characterize the MORPHOSAURUS lexicon and from which guidelines for lexicon construction and management can be derived. This lexicon is mainly a lexicon of subwords, as introduced above, but it contains – for reasons to be explained in the following – a limited number of multi word entries. We therefore refer, in the following to the broader term "lexical unit", rather than "subword".

3.1. Attributes of lexicon entries

Every lexical unit is classified according to one of the following categories:

- Language: English (en), Spanish (sp), German (ge), Portuguese (pt), French (fr), Swedish (sw)... The language attribute refers to the real-world occurrence of lexemes, including common foreign words. This means that English lexemes which commonly occur as foreign lexemes in a certain domain (e.g. *shunt*, *round*, *feedback*) are considered lexemes of the respective host language.
- Lexical units are word stems, prefixes, suffixes, infixes, proper prefixes, proper suffixes, or invariants:
  Stems (ST), like *gastr*, *hepat*, *enferm*, *diaphys*, *head* are the primary content carriers in a word. They can be prefixed, linked by infixes, and suffixed, some of them may also occur without affixes; **Prefixes** (PF), like *de-*, *re-*, *in-*, *an-* precede a stem once or more [7]; **Proper Prefixes** (PP) like *peri-*, *hemi-*, *down-* are prefixes that themselves cannot be prefixed; **Infixes** (IF), like *-o-*, in *gastr-o-intestinal*, or *-r-*, in *hernio-r-rafia* are used as a (phonologically motivated) glue between stems; **Suffixes** (SF) such as *-a*, *-io*, *-ion*, *-tomy*, *-itis*[8] follow a stem or another suffix; **Proper Suffixes** (PS) (e.g. verb endings such as *-ing*, *-ieron*, *-ão*, *-iésemos*) are suffixes that cannot be suffixed. All these lexeme types are used for segmentation of inflected, derived and composed words, taking into account their compositional constraints. In contradistinction, **Invariants** (IV), like *ion*, *gene* coincide with words and are not allowed as word parts. In most cases, these are short words which would cause artificial ambiguities if they could be used as building blocks for complex words.

We use the following notation for lexical items: The languages are added as superscripts, the lexeme type as subscript, e.g. $ectom_{SF}^{[en,sp,pt]}$ means that the string "ectom" acts as a suffix in English, Portuguese, and Spanish. An MID represents the sense of a group of lexemes which are considered synonymous in the given domain context, e.g. #remove = $\{ectom_{SF}^{[en,sp,pt]}, exstirp_{ST}^{[en,pt]}, estirp_{ST}^{[sp]}, remov_{ST}^{[en,sp,pt]}, \ldots\}$ Meaningless lexemes (stop entries), e.g. grammatical suffixes like *-ation*, *-s*, *-ed*, *-ación*, auxiliary and modal verb forms like *is*, *have*, *would*, *tuvieron*, *es*, *era*, *soy* , *são* are not assigned to an MID, since they are ignored for indexing.

3.2. Equivalence Class Relations

At introduced above, we link MIDs by two semantic relations, viz. *has-sense*, and *expands-to*. Groups of lexemes which have (the same) multiple senses are assigned a MID of their own. The *has-sense* relation then connects such ambiguous MIDs to each of its senses. Example: #lobo=$\{lobo_{IV}^{[sp,pt]}, lobos_{IV}^{[sp,pt]}\}$ is linked by *has-sense* to both #wolf =$\{wolf_{ST}^{[en]}, wolves_{ST}^{[en]}, \ldots\}$ and #lobe =$\{lob_{ST}^{[en]}, \ldots\}$. #cold =$\{cold_{IV}^{[en]}\}$ is linked to both #lowtemp =$\{frio_{IV}^{[sp,pt]}, fria_{IV}^{[sp,pt]}, \ldots\}$ and #commoncold =$\{\text{'common cold'}_{IV}^{[en]}, \ldots\}$.

The *expands-to* relation links one or more non-atomic lexemes (which are also grouped by a MID) to their atomic senses. There are mainly three reasons for this:

[7] E.g. in *hemi-an-opsia* the prefix *an* is prefixed by *hemi*

[8] The classification of subwords like *-logia* or *-itis* as suffixes may be controversial. For the applications supported by the MORPHOSAURUS lexicon, this is, however, of minor relevance. As a rule of thumb, our criterion for stems is that they do not require any other stem in order to build well-formed words.

1. Utterly short morphemes are not permitted as word constituents in order to prevent improper segmentation of compounds. Words which contain these morphemes must therefore have their semantic decomposition pre-coded. For example, #myalg = $\{myalg_{ST}^{[en]}, mialg_{ST}^{[sp,pt]}\}$ is linked by *expands-to* to the sequence of #muscle = $\{muscul_{ST}^{[en,sp,pt]}, muscle_{ST}^{[en]}, \ldots\}$ and #pain = $\{algy_{PS}^{[en]}, algia_{SF}^{[sp,pt]}, pain_{ST}^{[en]}, \ldots\}$, thus avoiding the occurrence of *my* or *mi* in the lexicon;
2. An indecomposable lexeme in one language has a composed sense in the reference language[9]. For example, #esparadrapo = $\{esparadrap_{ST}^{[sp,pt]}\}$ is linked by *expands-to* to the sequence of #adhesive = $\{adhesiv_{ST}^{[en,sp,pt]}, \ldots\}$ and #tape =$\{tape_{IV}^{[en]}, \ldots\}$;
3. Compounds exhibit ellipsis (omission of characters): For example, #urinalise = $\{urinalise_{ST}^{[pt]}\}$ is linked by *expands-to* to the sequence of #urine = $\{urin_{ST}^{[en,sp,pt]}, \ldots\}$ and #analysis =$\{analys_{ST}^{[en]}, analyis_{ST}^{[sp,pt]}, \ldots\}$;
4. Words are nondecomposable but have an inherent composite semantic structure, e.g. #broad = $\{broad_{ST}^{[en]}, larg_{ST}^{[sp,pt]}\}$ is linked by *expands-to* to the sequence of #breadth = $\{largur_{ST}^{[sp,pt]}, breadth_{IV}^{[en]}, \ldots\}$ and #highgrade.

3.3. Delimiting subwords

A comprehensive list of standard and domain-specific affixes is the starting point of subword dictionary building. Sources for affixes and infixes are the morphological grammar specification for the respective languages.[10] As a consequence, the main criterion for the delimitation of a word stem is its compatibility with existing prefixes and suffixes *in+compat+ibility*, *aprend-izaje*, *ventricul-i*. Wherever derivation causes a clear change of word sense which goes beyond the combined sense of the compounds, the derivate gains status of new lexeme with a different MID, e.g. *decubit-* in addition to *cubit-*, *neurot-* in addition to *neur-*. Many words of Latin and Greek origin come with stem variants (e.g., *corpus, corpor+is; abdomen, abdomin+al, diagnos-is, diagnost-ico*). Here, a reduction to the common substring (*corp-* or *abdom-*) would cause the proliferation of pseudo-suffixes (here *-oris, -inal*) on the one hand and the generation of short word stems on the other hand. In these cases stem variants are accounted for.

A high performance extraction of subwords from large amounts of text is best achieved by the use of finite-state techniques for lexicon-based decomposition, dederivation and deflection such as described in [Schulz and Hahn 2000]. Lexicon builders' decisions of subword delimitation are therefore driven not only by formal linguistic criteria, but also by the proper function of segmentation. This is especially relevant with long and composed words where different valid segmentations are possible. For example, *nephrotomy* may be segmented into $nephr_{ST}^{[en]}$ (#kidney) + $o_{IN}^{[en,sp,pt]}$ + $tomy_{PS}^{[en]}$ (#incision), but also into $nephr_{ST}^{[en]}$ + $oto_{ST}^{[en]}$ (#ear) + $my_{ST}^{[en]}$ (#muscle). If the word segmentation routine, here, prefers a long match from the left, the second (erroneous) segmentation would be preferred. Only costly knowledge and language processing routines (which are not available, in general) would be expected to detect this kind of errors. A pragmatic solution is to include additional synonymous lexeme variants. In our example, this means that the

[9]Reference Language is English. Therefore expansions into other languages are not allowed, since the intelligibility of the semantic structure of this dictionary would be restricted to the speakers of that language.

[10]Common agglutination of suffixes may be pre-coded (e.g., *-igkeiten*, *-izations*, *-ivelmente*, *-ectomies*, *-ingness*, *-ationally*).

sense #kidney is not only represented by $nephr_{ST}^{[en]}$ but also by $nephro_{ST}^{[en]}$ (as well as by $nefr_{ST}^{[sp,pt]}$ and $nefro_{ST}^{[sp,pt]}$).

3.4. Corpus-based validation of string specificity

Especially short or ambiguous word stems, such as *gen, my, mi, ship* are prone to side effects as described above. The shorter they are, the more frequently they arbitrarily occur as accidental substrings, producing erroneous segmentation results. In order to empirically assess this risk, we match them against word lists built from domain-specific text corpora. Here we distinguish between two cases:

- The number of accidental matches is high: First, all correct matches have to be checked. Here, in many cases, the short stem will occur at the beginning of a word. If this does not lead to false matches, we can add (unorthodoxly) this stem as a proper prefix in order to make use of the position constraint on this class of lexemes. If there are still many occurrences in the inside of words left, then, the pertaining compounds or prefix-stem combinations have to be added to the lexicon and linked to their components by expansion. An example therefore is the stem *ship*. We must avoid that the sense of *ship* (vessel, to send) is extracted from any word with the suffix -*ship*, e.g. *relationship*. Therefore *ship* is added both as an invariant and a prefix (!) instead of a stem, together with usual inflections. For each excluded short stem, the most frequent compounds and derivatives have to be included, together with their inflections (e.g. *#eat* = $\{eat_{IV}^{[en]}$, $eats_{IV}^{[en]}$, $eating_{IV}^{[en]}$, $ate_{IV}^{[en]}$, $eaten_{IV}^{[en]}$, $eater_{ST}^{[en]}\}$). In order not to preclude synonymy match, e.g., #egg = $\{ov_{ST}^{[en,sp,pt]}$, $ovo_{ST}^{[en,sp,pt]}$, $huev_{ST}^{[sp]}$, $egg_{ST}^{[en]}\}$, the syntagmantic expansion link can be used, e.g. #oocyte = $\{oocyte_{ST}^{[en]}$, $oocit_{ST}^{[sp,pt]}$,$\}$ is linked by *expands-to* to #egg and #cell.
- There are relatively few accidental matches. Here the strategy is the opposite one. The stem is added to the lexicon, and the erroneously matching words are segmented. Wherever the erroneous stem happens to extracted, adjustments have to be made at the components of these words. An example for this is the *nephrotomy* example. Instead of eliminating *oto* as a stem, the stem variant *nephro* is added (see above) and thus false segmentation results are avoided.

3.5. Criteria for Inclusion of Subwords in the Dictionary

The selection of lexical units should reflect the language use in the domain of interest. Again, we use word statistics extracted from extensive, language specific corpora in order to measure the relevance of terms. Ideally, each lexicon entry should correspond to an atomic (indivisible) entity of semantic reference. However, there are borderline cases, especially where a composed lexeme may have an atomic synonym. As a consequence, the atomic lexeme is either related to the components of its synonym by the relation *expands-to* (a), or the composed lexeme is entered as a whole and equaled with its atomic synonym. Example:

1. #ascorb = $\{ascorb_{ST}^{[en,sp,pt]}$, *vitamin* $c_{IV}^{[en]}$, *vitamina* $c_{IV}^{[sp,pt]}\}$.
2. #ascorb is expanded to the sequence of #vitamin = $\{vitamin_{ST}^{[en,sp,pt]}\}$ and #C = $\{c_{IV}^{[en,sp,pt]}\}$

The latter case is preferred if the components of the composed lexeme are semantically relevant, the first one if the components are semantically weak.

In contrast to the general rule, semantically underdetermined complex lexemes or noun groups need not to be included in the dictionary as long as there exists a strict mapping through all languages of interest. As an example, the sense of the term *yellow fever* is not derivable from its components, but its components literally translate to all languages (*fiebre amarilla, febre amarela, gelbfieber*).

Proper names are entered into the lexicon under the following circumstances: (i) they are needed for synonym linkage, e.g. between different product names, e.g. #diclofenac = $\{diclofenac_{ST}^{[en,sp,pt]}, voltaren_{ST}^{[en,sp,pt]}, cataflam_{ST}^{[en,sp,pt]}\}$; (ii) they are used as eponyms, i.e. they belong to the domain terminology (e.g. *crohn*, *parkinson*); (iii) translations exist, especially with regard to geographic terms (*suiza* = *switzerland*).

3.6. Aspects of lexicon construction

Finally we outline the process of lexicon construction as it underlies the MORPHOSAURUS lexicon. It is based upon the view that the delimitation of classes of semantic equivalence is mainly an intellectual task which cannot be fully automatized. Therefore, as a starting point, each lexicon entry has its own MID. If the lexicon designer concludes that two lexicon entries have identical sense, then the two MIDs are fused. The incremental fusion of lexemes, however, leads repeatedly to a class of decisions which we can consider the main dilemma of the lexicon engineering process. Let K, L, and M be atomic lexical items. Two users group these items in different ways, according to slightly different subdomain contexts, here represented by D_1 and D_2, respectively. In D_1 the lexical items K and L are considered synonyms. In D_2, however, M and not L is considered a synonym of K. The fusion of these two subcontexts gives raise to the two solutions, *viz.* closure and sum. Whereas the closure operation merges the synonym classes, the sum operation preserves the context-related distinction and introduces two senses for the ambiguous equivalence class. The decision of whether following the one or the other strategy is complex. On the one hand, we end up with a tight network of ambiguous senses when pursuing the latter strategy. On the other hand, the transitive closure tends to yield numerous synonym classes in which pairs of lexemes are far from being synonymous. As an example, a user may assert synonymy between *head* and *caput* in an anatomy subdomain. Another one equalizes *head* with *chief*, when modeling terms in a subdomain of administration. Applying the closure operation, *chief* would become synonym to *caput*, and all literal and figurative senses of *head* would be represented by one MID. Applying the sum operation, *head* would be assigned an ambiguous MID which then would be related to its non-ambiguous senses.

4. Conclusion and Further Work

The construction of multilingual dictionaries which account for the variety of meanings in different domain contexts and languages constitutes a major challenge, even if restricted to a technical sublanguage such as the medical one. We have presented an approach which concentrates on the economic encoding of subwords as lexical units. The main criterion for the inclusion of a subword entry in the lexicon is semantic atomicity, since semantically composed entries can be reconstructed out of atomic ones. Beside the proper delimitation of lexical items, which should optimize both generality (to warrant a high recall)

and specificity (to warrant a high precision), the grouping of lexical items in domain-specific equivalence classes has posed problems which have required the formulation of rigid editing guidelines for lexicon developers and are currently guiding the development of benchmarking and validation tools. Presently, the MORPHOSAURUS lexicon contains about 80,000 lexical items which are related to about 20,000 equivalence classes. Due to its compositional character it has a high coverage for English, Portuguese, Spanish, and German. French and Swedish lexicons are currently under construction.

Acknowledgements: This research was sponsored by CNPq, the Brazilian Research Council.

References

(2004). SNOMED *Clinical Terms*. Northfield, IL: College of American Pathologists.

Buvač, S., Buvač, V., and Mason, I. A. (1994). The semantics of propositional contexts. In *Proceedings of the 8th International Symposium on Methodologies for Intelligent Systems*. Berlin: Springer.

Eco, U., Robering, K., Scheffczyk, A., and Habermeier, R. (1988). Metamorphoses of the semiotic triangle. *Zeitschrift für Semiotik*, 10(3).

Fellbaum, C., editor (1998). WORDNET*: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hahn, U., Markó, K., Poprat, M., Schulz, S., Wermter, J., and Nohama, P. (2004). Crossing languages in text retrieval via an interlingua. In *RIAO 2004 – Conference Proceedings: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, pages 100–115. Avignon, France, 26-28 April 2004. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaire (CID).

Hirst, G. (2004). Ontologies and the lexicon. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 209–229. Berlin: Springer.

McCray, A. T., Browne, A. C., and Moore, D. L. (1988). The semantic structure of neo-classical compounds. In Greenes, R. A., editor, *SCAMC'88 – Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care*, pages 165–168. Washington, D.C., November 1988. New York, N.Y.: EEE Computer Society Press.

MESH (2004). *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.

UMLS (2004). *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

Schulz, S. and Hahn, U. (2000). Morpheme-based, cross-lingual indexing for medical document retrieval. *International Journal of Medical Informatics*, 59(3):87–99.

Schulz, S., Markó, K., Sbrissia, E., Nohama, P., and Hahn, U. (2004). Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics*, volume 2, pages 813–819. Geneva, Switzerland, August 23-27, 2004. Association for Computational Linguistics.