Semantic Atomicity and Multilinguality in the Medical Domain: Design Considerations for the MORPHOSAURUS Subword Lexicon

Stefan Schulz[1,2], **Kornél Markó**[1,5], **Philipp Daumke**[1],
Udo Hahn[5], **Susanne Hanser**[1], **Percy Nohama**[2,3],
Roosewelt Leite de Andrade[2,3], **Edson Pacheco**[2,3], **Martin Romacker**[4]

[1]Department of Medical Informatics, Freiburg University Hospital, Freiburg, Germany
[2]Health Informatics Laboratory, Paraná Catholic University, Curitiba, Brazil
[3]Graduate Program in Electrical Engineering and Industrial Informatics, CEFET-PR, Curitiba, Brazil
[4]Text Mining in Life Sciences Informatics, Novartis, Basel, Switzerland
[5]Jena University Language & Information Engineering (JULIE) Lab, Jena, Germany

Abstract

We present the lexico-semantic foundations underlying a multilingual lexicon the entries of which are constituted by so-called subwords. These subwords reflect semantic atomicity constraints in the medical domain which diverge from canonical lexicological understanding in NLP. We focus here on criteria to identify and delimit reasonable subword units, to group them into functionally adequate synonymy classes and to relate them by two types of lexical relations. The lexicon we implemented on the basis of these considerations forms the lexical backbone for MORPHOSAURUS, a cross-language document retrieval engine for the medical domain.

1. Introduction

The form of lexicalization of nominal compounds, ranging from multi-word noun phrases to complex uni-words, varies grossly among different languages. Whereas in English the consituent words in the noun phrase *high blood pressure* directly reflect the building blocks of its semantic interpretation, this is not the case with its literal translations *verhoogde bloeddruk* (Dutch) or *Bluthochdruck* (German). Especially in scientific sublanguages we encounter atomic senses at different levels of lexicalization. An atomic sense may well correspond to word stems (*hepat-*), prefixes (*anti-*), suffixes (*-logy*), but also to word fragments composed of at least two independent stems (*hypophys-*), straight words (*spleen*) or even combinations of words whose compositional interpretation does not match the intended meaning of the term (*yellow fever*).

Ad-hoc term formation is common so that for the growing number of combinations a high lexical coverage can only be achieved when lexical units are restricted to units of atomic sense. Extracting these basic units from texts is an important goal for many applications, e.g., for cross-language document indexing and retrieval such as in the MORPHOSAURUS system (Markó et al., 2005a). It builds upon a multilingual lexicon of semantically atomic lexical units, so-called subwords, covering the domain of clinical medicine. In the following, we shall give a semi-formal account of lexical atomicity as the basis of the MORPHOSAURUS medical subword lexicon.

2. Semantic Atomicity

We consider a lexical form to be semantically atomic, if its sense(s) (in a given language and a given domain context) cannot univocally be derived from the sense(s) of its lexical constituents. Non-atomicity is usually due to word forming operations such as inflection, derivation and composition. *Inflection* combines the lexical sense of the word stem with the grammatical function of the affix. *Derivation*, however, covers various phenomena. A derivational affix may simply change the part of speech of the basic form without any semantic implication (e.g., *patient with a severe injur-y = severe-ly injur-ed patient*). But it may also add an additional sense, such as with *hepatitis = hepat* (liver) + *itis* (inflammation). However, cases in which the derived form has gained an autonomous sense are frequent as well. *Neurosis*, e.g., is the result of linking *neur* (nerve) with *osis* (disease), but *neurosis* is still different from a disease of nerves (at least in modern medicine). Hence, the latter derivation should be considered as an atomic lexical unit. (Uni-word) *composition*, finally, combines two or more stems in one word. It is a frequent phenomenon in scientific sublanguages where words like *adenosintriphosphat*, *inmunodeficiencia*, *prebetalipoproteinemia*, referred to as "neoclassical compounds" (McCray et al., 1988), are ubiquitous.

Lexical units may have multiple senses (*polysemy*, in a broad sense); and one sense can be expressed by different lexical forms (*synonymy*). Although domain specific terminologies should enforce controlled lexicalization in a specialized language and, thus, avoid lexical ambiguity, non-standardized terminology is widely used in any domain.

Besides ambiguity, lexical units may have overlapping senses. Quasi-synonym relations may hold between terms of different language (*caput*, *head*) or different levels of erudition (*belly*, *abdomen*). Complete sense identity (i.e., true synonymy) in all possible uses of a lexical form is a rare phenomenon. So when we establish classes of synonymous expressions we, firstly, have to make a clear commitment to the context in which the expressions are considered synonymous, i.e., their *domain context*, and secondly, we have to convene upon a degree of tolerance in sense deviation which is still compatible with the algebraic properties of an equivalence relation. Hence, if we agree on considering *disease* a synonym of *illness* and *illness* a synonym of

sickness, then (by transitivity) *disease* and *sickness* must be synonyms, as well. The tolerance level depends also on the relevance of subtle sense distinctions in the chosen domain context.

In order to represent atomic senses of lexical units we define a fundamental semantic layer in terms of language-independent identifiers, so called MIDs (**M**orphoSaurus **ID**s). Language-independency is achieved by treating translations in different languages as synonyms, thus grouping them together in a single MID class (e.g., {*disease, illness, maladie, enfermedad, doença*}). MIDs can roughly be compared to concepts in thesauri (such as CUIs in the UMLS metathesaurus (UMLS, 2004), or synsets in WORDNET (Fellbaum, 1998)).[1] However, there are two major differences between MIDs and UMLS CUIs or WORDNET synsets. Firstly, MIDs can represent disjunctions of senses. This is the case when ambiguous lexical units are addressed. Following up on the above example, the disjunction of the different senses of *molar* is represented by one MID, and each of the non-ambiguous senses by yet another, different MID. Secondly, all lexical units which are assigned to one MID must be fully interchangeable within the given domain context. For instance, {*head, caput, cabec, cabez, cefal, cephal* } would not be a proper representation of one specific MID because of the additional senses that can be attributed to the English noun *head*. The interchangeability is an important goal in order to create language-independent concepts as a basis for semantic interoperability.

We now introduce the notion of *subword* as the minimal meaning-bearing constituent of a domain-specific term. Its defining property is that its sense be non-decomposable. *Hepatitis*, e.g., is not considered a subword because its sense can be derived from its constituents, *hepat* and *itis*. By contrast, the decomposition of *hypophysis* into presumed sense components, *hypo* and *physis*, does not lead to the shared sense of *hypophysis*. For each subword there exists at most one MID, where the assignment of the MIDs depends on the domain context d and the language l under consideration. If no meaning can be assigned to a subword, it constitues a *null entry* (it has only a grammatical function), such as auxiliary verbs or inflectional suffixes. The relation between a subword sw, a MID m, a domain context d and the specific language l can then be expressed by quadruples of the following form:

- $(sw_1, m, d, l), (sw_2, m, d, l), (sw_3, m, d, l)$

  sw_{1-3} are synonyms in domain d and language l since they refer to the same MID m.

  Example: *nephr-*, *ren-*, *kidney*

- $(sw_1, m, d, l_1), (sw_2, m, d, l_2)$

  sw_1 in language l_1 is a translation of sw_2 in l_2 in domain d which is expressed by the reference to the same MID m.

  Example: *nephr-*, *riñon*

- $(sw, m_1, d, l), (sw, m_2, d, l)$

  sw has two senses, m_1 and m_2, in domain d and language l.

  Example: *head* refers both to body parts and to top-level staff.

- $(sw, null, d, l_1), (sw, m_2, d, l_2)$

  sw is a null entry in language l_1, while it has the sense m_2 in language l_2.

  Example: *era* is an auxiliary verb in Spanish and Portuguese but a noun in English.

- $(sw_1, m_1, d_1, l_1), (sw_2, m_1, d_1, l_1), (sw_1, m_2, d_2, l_1), (sw_2, m_3, d_2, l_1)$.

  sw_1 and sw_2 are synonyms in language l_1 and domain d_1 but not in domain d_2.

  Example: *sildenafil* and *viagra* are considered synonyms in clinical medicine but not in the context of pharmaceutical industry.

MIDs can be linked by two lexical relations, *viz.* the syntagmatic relation *Expands* and the paradigmatic relation *Has-Sense*:

- *Expands*$(m_0, [m_1, m_2, ..., m_n])$ relates the MID m_0 to an ordered list of MIDs (composed of at least two elements). One of the uses of this relation is to deal with composed meanings in compounds which exhibit contractions, e.g., *urinalysis*.

- *Has-Sense*$(m_0, \{m_1, m_2, ..., m_n\})$ relates an ambiguous MID to a set of MIDs (composed of at least two elements), which constitute its (non-ambiguous) senses. For 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".

Both relations are transitive. Insertions into lists or sets create expansions, not nestings:
Expands$(m_0, [m_1, m_2])$ &
Expands$(m_1, [m_3, m_4])$ is equivalent to
Expands$(m_0, [m_3, m_4, m_2])$;
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 $\mathcal{D}$. Diverging from many thesauri such as the UMLS (UMLS, 2004) or WORDNET (Fellbaum, 1998), we do not define additional semantic relations such as hypernymy, meronymy. Such semantic enhancements can be obtained by linking the lexicon to external thesauri or ontologies (such as MESH (Markó et al., 2004)).

[1] MIDs are represented by composing the '#' symbol with one of its non-ambiguous English lexemes, e.g., #liver = {*hepar, hepat, liver, figad, higad*}.

3. The MORPHOSAURUS Lexicon

We have implemented the lexicon structure from the previous section in the MORPHOSAURUS lexicon. This serves as the lexical repository for the MORPHOSAURUS indexer which determines indexing units from input texts and maps them to interlingual MIDs. The MORPHOSAURUS lexicon is currently committed to a single, well-defined domain context, *viz.* clinical medicine.

3.1. Attributes of lexicon entries

MORPHOSAURUS lexemes are classified according to language and lexeme type: The languages currently supported are English (en), Spanish (sp), German (ge), Portuguese (pt), French (fr), and Swedish (sw). The language type reflects the grounding of lexemes in a particular language. We distinguish the following lexical types:

Stems (ST), like *hepat*, *enferm*, *diaphys*, *head* are the primary content carriers in a word. They can be prefixed, suffixed, or linked by infixes; some of them may also occur without any affix;

Prefixes (PF), like *de-*, *re-*, *in-*, precede a stem;[2]

Proper Prefixes (PP) like *peri-*, *hemi-*, *down-* are prefixes that cannot be prefixed;

Infixes (IF), like *-o-*, e.g., 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* 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* or *gene* are not allowed as word parts. In most cases, these are short words which would cause artificial ambiguities if they were used as building blocks for complex words.

As a notational convention, the language type is indicated by superscripts, while the lexeme type occurs in subscripts, e.g., $ectom_{SF}^{[en,sp,pt]}$ means that the string "ectom" acts as a suffix in English, Portuguese, and Spanish. With this convention, an MID represents the sense of a group of lexemes which are considered synonymous in the given domain context such as with #remove = $\{ectom_{SF}^{[en,sp,pt]}, exstirp_{ST}^{[en,pt]}, estirp_{ST}^{[sp]}, remov_{ST}^{[en,sp,pt]},\dots\}$. Null entries, e.g., grammatical suffixes like *-ation*, *-s*, *-ed*, *-ación*, auxiliary and modal verb forms are not assigned to an MID.

3.2. Equivalence Class Relations

Additionally, we link MIDs by two lexical relations, *viz. Has-Sense* and *Expands*. Groups of lexemes which share multiple senses are assigned to an MID of their own. The *Has-Sense* relation then connects such ambiguous MIDs to each of its senses. For example, #lobo = $\{lobo_{IV}^{[sp,pt]}, lobos_{IV}^{[sp,pt]}\}$ is linked by *Has-Sense* to both #wolf = $\{wolf_{ST}^{[en]}, wolves_{ST}^{[en]},\dots\}$ and #lobe = $\{lob_{ST}^{[en]},\dots\}$. #cold = $\{cold_{IV}^{[en]}\}$ is linked to #lowtemp = $\{frio_{IV}^{[sp,pt]}, fria_{IV}^{[sp,pt]},\dots\}$ and #commoncold = $\{$*'common cold'*$_{IV}^{[en]},\dots\}$.

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

1. Utterly short morphemes are not permitted as word constituents to prevent improper segmentation of compounds. So, words which contain these morphemes must have their semantic decomposition precoded. #myalg = $\{myalg_{ST}^{[en]}, mialg_{ST}^{[sp,pt]}\}$, e.g., is linked by *Expands* to the sequence of #muscle = $\{muscul_{ST}^{[en,sp,pt]}, muscle_{ST}^{[en]},\dots\}$ and #pain = $\{algy_{PS}^{[en]}, algia_{SF}^{[sp,pt]}, pain_{ST}^{[en]},\dots\}$, thus avoiding the occurrence of *my* or *mi* in the lexicon;
2. A non-decomposable lexeme in one language has a composed sense in another. #esparadrapo = $\{esparadrap_{ST}^{[sp,pt]}\}$, e.g., is linked by *Expands* to the sequence of #adhesive = $\{adhesiv_{ST}^{[en,sp,pt]},\dots\}$ and #tape = $\{tape_{IV}^{[en]},\dots\}$;
3. There are contractions in compounds, e.g., #urinalise = $\{urinalise_{ST}^{[pt]}\}$, which is then linked by *Expands* to the sequence of #urine = $\{urin_{ST}^{[en,sp,pt]},\dots\}$ and #analysis = $\{analys_{ST}^{[en]}, analyis_{ST}^{[sp,pt]},\dots\}$;

3.3. Delimiting subwords

We start with building the subword dictionary with a comprehensive list of general and sublanguage-specific affixes. The main criterion for the delimitation of a word stem is its compatibility with the given prefixes and suffixes (assuming regular morphological processes). Wherever derivation causes a clear change of word sense which goes beyond the combined sense of the constituents involved, this derivate explicitly gains the status of new lexeme with a different MID, e.g., *decubit-* in addition to *cubit-*, *neurot-* in addition to *neur-*.

Subword delimitation is therefore not only driven by purely formal linguistic criteria, but also by functional considerations (what kind of effects have alternative segmentations on the performance of the underlying CLIR system?), especially 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 is geared for a longest match from the left, the second (erroneous) segmentation would be preferred. A pragmatic solution is to include linguistically unorthodox variants, such as $nephro_{ST}^{[en]}$, in addition to $nephr_{ST}^{[en]}$.

Especially short or ambiguous word stems, such as *gen, my, mi, ship* are prone to unwarranted side effects, as they may arbitrarily occur as accidental substrings. 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:

[2] Prefixes can be prefixed, as well, e.g., in *hemi-an-opsia* the prefix *an* is prefixed by *hemi*.

The number of accidental matches is high: All pertaining compounds of the considered word stem have to be added to the lexicon and linked to their components by expansion. Consider the term *ship*. We must avoid that the 'stem' sense of *ship* (i.e., *vessel, to send*) is extracted from any word with the suffix *-ship*, e.g., *relationship*. So, rather than being a stem, *ship* is added as an invariant, together with its inflectional forms. For each excluded short stem, the most frequent compounds and derivatives then have to be included, together with their inflections.

The opposite case is given when there are relatively few accidental matches. Here the stem is added to the lexicon, and adjustments have to be made to the components of these words. Consider *nephrotomy* as an example. Instead of eliminating *oto* as a stem, the stem variant *nephro* is added and thus false segmentation results are avoided.

3.4. Criteria for Subword Inclusion

The selection of lexical units should reflect the language use in the chosen domain. We use word frequency in corpora in order to measure the relevance of terms. Ideally, each lexicon entry should correspond to an atomic sense. However, there are borderline cases, especially where a composed lexeme may have an atomic lexeme as a synonym. As a consequence, the atomic lexeme is either related to the components of its synonym by the relation *Expands*, such as with #ascorb which is expanded to the sequence of #vitamin = $\{vitamin_{ST}^{[en,sp,pt]}\}$ and #C = $\{c_{IV}^{[en,sp,pt]}\}$, or the composed lexeme is entered as a whole and equaled with its atomic synonym #ascorb = $\{ascorb_{ST}^{[en,sp,pt]}$, *vitamin* $c_{IV}^{[en]}$, *vitamin a* $_{IV}^{[sp,pt]}\}$. The latter case is preferred, if the components of the composed lexeme are semantically relevant, the first one if they are semantically weak.

3.5. Lexicon Engineering

The MORPHOSAURUS lexicon construction task is based upon the view that the formation of equivalence classes is currently still an intellectual process (cf. (Markó et al., 2005b) for an automated approach). Whenever a new lexicon is created, each lexicon entry has its own MID. If the lexicon designer observes that two lexemes have identical senses, then the two MIDs associated with these lexemes are fused. Figure 1 illuminates this situation. Let K, L, and M be atomic lexical items. Two lexicon designers group these items in different ways, according to slightly different subdomain contexts, here represented by d_1 and d_2. 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 rise to the two solutions, *viz.* closure and sum. Whereas closure simply merges the synonym classes, sum preserves the context-related distinction and introduces two senses for the ambiguous class. The decision which way to go 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, closure tends to produce large synonym classes in which pairs of lexemes can hardly be considered as synonymous. As an example, one designer might assert synonymy between *head* and *caput* in the anatomy subdomain.

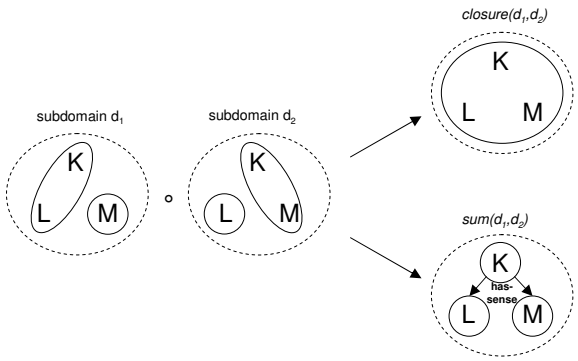

Figure 1: Fusing subdomains

Another one joins *head* with *chief*, when modeling terms in a subdomain of administration. Applying the closure operation, *chief* becomes a 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 to an ambiguous MID which then would be related to its non-ambiguous senses.

4. Conclusion and Further Work

We have presented an approach for the parsimonious encoding of lexical units as subwords . 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 generality (to warrant a high recall) and specificity (to warrant a high precision), the grouping of lexical items in domain-specific equivalence classes poses problems which require the formulation of rigid guidelines for lexicon curators. Presently, the MORPHOSAURUS lexicon contains roughly 80,000 lexemes which are related to about 20,000 equivalence classes. Due to its compositional character the lexicon has a high coverage for English, Portuguese, Spanish, and German.

Acknowledgments: This work was supported by the European Network of Excellence "Semantic Mining" (NoE 507505), the Brazilian federal grant CNPq 550240/2003-9, and the German federal grant (BMBF IB) BRA 03/013.

5. References

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

Kornél Markó, Udo Hahn, Stefan Schulz, Philipp Daumke, and Percy Nohama. 2004. Interlingual indexing across different languages. In *Proceedings RIAO 2004*, pages 82–99.

Kornél Markó, Stefan Schulz, and Udo Hahn. 2005a. MORPHOSAURUS: Design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4):537–545.

Kornél Markó, Stefan Schulz, Alyona Medelyan, and Udo Hahn. 2005b. Bootstrapping dictionaries for cross-language information retrieval. In *Proceedings 28th SIGIR* , pages 528–535.

Alexa T. McCray, Allen C. Browne, and D. L. Moore. 1988. The semantic structure of neo-classical compounds. *Proceedings 12th SCAMC'88*, pages 165–168.

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

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