Defining and Relating Biomedical Terms: towards a Cross-language Morphosemantics-based System

Fiammetta Namer[a], **Robert Baud**[b]

[a]UMR ATILF CNRS & University of Nancy2, Nancy, France

[b]Hôpitaux Universitaires de Genève, Geneva, Switzerland

Keywords

Natural Language Processing; Semantics; Language; Multilingualism; Neoclassical Compounds; Morphosemantics for French; Semantic Relations; Biomedical Lexical Database.

Abstract

This paper addresses the issue of how semantic information can be automatically assigned to compound terms, i.e. both a definition and a set of semantic relations. This is particularly crucial when elaborating multilingual databases and when developing cross-language information retrieval systems. The paper shows how morphosemantics can contribute in the constitution of multilingual lexical networks in biomedical corpora. It presents a system capable of labelling terms with morphologically related words, i.e. providing them with a definition, and grouping them according to synonymy, hyponymy and proximity relations. The approach requires the interaction of three techniques: (1) a language-specific morphosemantic parser, (2) a multilingual table defining basic relations between word roots, and (3) a set of language-independent rules to draw up the list of related terms. This approach has been fully implemented for French, on an about 29,000 terms biomedical lexicon, resulting to more than 3,000 lexical families. A validation of the results against a manually annotated file by experts of the domain is presented, followed by a discussion of our method.

1. Introduction

The approach and the results presented here[1] contribute to the development of a structuration of biomedical lexicons by the use of a morphosemantics-based approach, i.e. which provides morphologically complex words with a definition as well as with lexical relations to other words ([1], [2]). By morphosemantic, we mean morphological analysis of derived and compound words and semantic interpretation of the whole from the meaning of the parts and their relations. Our objective with such semantically tagged terms is to enrich thesauri, terminologies and ontologies, to enable cross-language question-answering and to extend information retrieval requests to neighbour concepts.

Moreover, our linguistic-based method contributes to solve the general issue of multilingual terminology and cross-language information retrieval (IR) in the medical domain, due to the fact that the roots are common to most of the Western languages; this issue is addressed, eg, in [3], [4] and [5].

The fundamental principle is that semantic information is acquired on morphologically complex words through the following actions: morphosemantic analysis, collection of lexical data about Latin and Greek-based roots and derivation of lexical relations using computation rules. The typical lexical relations inferred by this process are: first, each complex word is related to other words build on the same basic root (i.e. its morphologically related word); second, pairs of complex word may be bound by links of synonymy, hyponymy or proximity.

The developed system relies on three hypotheses. First, complex words form more than 60% of the new terms found in techno-scientific domains, and especially in the field of biomedecine [6][7]. It is therefore difficult to permanently update dictionaries in order to collect all neologisms. On the other hand, linguistic-driven constraint-based morphosemantic systems are suitable to define words meaning with respect to the meaning of their parts: for

[1] The here reported methods and results are supported by the projects UMLF (coordination: P. Zweigenbaum, grant from French Ministry for Research and Education, 2002-2004) [1], and VumeF (coordination: S. Darmoni, grant from French Ministry for Research, National Network of Health Technologies, 2003-2005) [2].

instance, whereas Dorland's medical dictionary ([8]) proposes the following definition for the adjective *anticephalalgic*[2]*:* "inhibiting headache", a morphosemantic parser as presented in this article is able to provide it with the following definition: "which is against brain pain".

Second, we observe that whatever the involved Western language[3], complex words in biomedical field make use of Latin and Greek roots, which will be called here combining forms (CF) following [9]. CFs inherit their part-of-speech tag from the modern language words they substitute for (stomach,N → gastr,N). Additionally, CF realizations are simple graphic variants from a language to another. Very similar word formation rules are at play in all these languages to build words belonging to specialized terminologies. Both CFs and complex word structures are therefore likely to be identified by neutral representations, which abstract away differences between languages: VASCUL--ITE[4] = *vascul--ite*$_{FR}$ = *Vascul--itis*$_{GE}$ = *vascol--ite*$_{IT}$ = *vascul--itis*$_{ES/EN}$.

The third assumption deals with biomedical classifications: just like words they substitute for, abstract CFs can be ranked according to sound hierarchies (SNOMED, MeSH…), in such a way that they can be labelled by descriptors such as *anatomy* (GASTR), *physiology* (TAXI) or *pathological process* (ALGI). On this basis the CFs may be combined by different links.

We are currently processing 4 types of links:

- **synonymy** represented by "**=**" (e.g. OPT=OPHTALM, vision); this link is strong and pairs of CFs shares all their properties; synonyms are usually different by their quality (preferred term, rare term, old fashion term, jargon, acronym, etc).
- **hyponymy** represented by "<" (e.g. BLAST, embryonal cell < CYT, cell); this link is also very strong and by definition the descendant CF inherits all the properties on its

[2] Though we said the system currently runs for French, examples are given in English whenever necessary, for sake of readability.
[3] This claim is illustrated here with examples in German (GE), English (EN), French (FR), Italian (IT) and Spanish(ES)
[4] Throughout the paper, abstract CFs are written in small capitals, CF to CF boundaries are represented by '--'

ancestor. Morphological compounding we are dealing with is another way to produce hyponymy links, namely between a compound noun (e.g. *dacryoadenitis*) and its rightmost nominal component (e.g. *adenitis*): the former denotes a specialisation with respect to the latter.

- **meronymy** (part-to-whole relation between entities) represented by "←" (e.g. CORO, pupil ← OCUL, eye); this link is not really strong, nor weak, but it does not preserve the properties between the whole and the parts; it is less useful than the two types of links above[5].
- **proximity** represented by "~" (e.g. DISC, intervertebral disk ~ SPONDYL, vertebra); proximity may be of various types : physical (example above), functional (SIAL : "salivary gland" and STOMAT "mouth"), procedural (TOMI : "incision" and ECTOMI : "ablation"), etc. This link is undoubtedly weak and no inheritance of properties is possible through it. No transitivity can be inferred for this link: if A ~B and B~C, it is not possible to say that A~C. Therefore this link is quite difficult to use. Nevertheless, semantic proximities often represent evidences and cannot be ignored for language disambiguisation, typically for query expansion when noise is acceptable to a cost of a loss in specificity.

These background hypotheses are exploited in order to reach our current objectives to be developed in this article: to elaborate a multilingual grouping of similar words in different languages; to supply biomedical terms with semantic information; to prove the feasibility of the method by implementing it for French.

2. Materials and Methods

2.1 Language specific resources

As we shall see, the quality of the results mainly depends on lexicon size. Therefore

[5] The use of the « meronymy » notion is here linguistically motivated. Its meaning differs from its definition in formal ontologies.

large-scale monolingual lexica are required in order to optimize lexical content and coverage. On the other hand, only one resource among the three of which our methodology relies on is fully language-specific, namely the word formation parser. Linguistic-driven constrained-based morphological analysis, i.e., the process of decomposing a complex word into its constituent parts, is a language dependent task. It has been proved useful to avoid the need for costly, repetitive maintenance of specialized dictionaries to account for new terms ([10][11][12]); moreover, it can additionally enrich the decomposition of each word with semantic knowledge, as described in [13][14].

For biomedecine, units that compose complex words often are CFs: in suffixation *epatico*$_{IT}$ (hepatic), prefixation *Hypothermie*$_{GE}$ (hypothermia), as well as in compounding *thermo*--*taxy*$_{EN}$, *stomac*--*odynie*$_{FR}$ (stomach--odynia). Unlike affixation, which builds a complex word by applying a suffix (*-ico*$_{IT,ES}$) or a prefix (*hypo-*$_{GE,FR,EN}$) to a base word or CF, compounding constructs a new word by associating two words or CFs. Among compounds, the paper focuses on single word "neoclassical" compounds, in which at least one of its component is a CF: *thermoregulation*, *hypophysectomy*, *thermotaxy*. In these compounds, the rightmost component, noted X, semantically heads the lefthand (modifier) component, noted Y.

2.2 Multilingual CF Table

The second and third hypotheses of section 1 lead to the design of a 900 rows table sampled in Table 1. Part-of-speech tag (3), SNOMED head chapters (4) and basic lexical relations (5) refer to CF abstract representations whereas instantiation (2) deals with CFs (1) respective realizations[6] and translations in each language (see [15][16]).

2.3 Language independent Lexical Relation Computation Rules

The third technique required to perform cross-language semantic tagging on

[6] According to the language in consideration, CFs may have ambiguous written forms: so in French 'aur' means either *gold* (*aurithérapie*$_{N}$: "gold therapy" or *ear* (*auriforme*$_{A}$: "ear shaped"). Abstract CF disambiguisation is ensured by the other field values in the CF Table

compounds is a set of rules capable of propagating basic lexical relations attached to CFs, and encoded in the CF Table, onto words which are composed with these CFs. There are currently four rules that only deal with compound words, as indicated in Table 2. Extensions are in progress to account for affixed words as well.

- The first rule (**R1**) performs the following task: assume two compound words A and B, headed by the same X component (as e.g. in *procto**rrhagia*** and *colo**rrhagia***). When Y_A and Y_B are related, according to the content of Multilingual CF Table (see Table1), this relation is used to compute the one which holds between A and B: namely, A and B share the same relation as Y_A and Y_B (ABDOMIN=LAPAR implies that ***abdomino**scopy* means the same as ***laparo**scopy*), except for the meronymy relation. If Y_A is a part of Y_B (PROCT ← COLO), then A (***proct**ectomy*) is an hyponym of B (***col**ectomy*)[7].
- Let us now consider the rule **R2**. Here, A and B are headed, respectively, by two synonymous components X_A and X_B (ALGI equals ODYNI). The reasoning is the same as with Rule **R1**. So, if Y_A refers to a part of Y_B (ENTER ← ABDOMIN) then A is a hyponym of B (*enteralgia* is a special type of *abdominodynia*). Otherwise, if Y_A and Y_B hold any other basic relation (ABDOMIN equals LAPAR, ALBUMIN is a subtype of PROTEIN, XER is an approximation of SCLER) then A and B share the same relation as Y_A and Y_B, whatever the language.
- Rules **R3** and **R4** are symmetric to, respectively, **R1** and **R2**: X' and Y'roles are swapped. In **R3**, Y components (e.g. ARTHR) are equal in A and B, and in **R4**, they are synonymous (e.g. LIP=ADIP). Then, the semantic relation between A and B depends on that of X_A and X_B . For instance, by means of **R3**ITIS~ALGI involves for *arthritis* to have to do with *arthralgia*; and, through **R4**, the similarity between *lipomatose* and

[7] As we shall see in section 4, such an inference may lead to wrong hyponymy predictions, due to discrepancies between linguistic meaning and real one.

adiposis comes from that of MATOSE and OSE.

The interaction between morphosemantic parser, CF table and lexical relation computation rules results in a processing chain which leads to the tagging of compounds by means of the =, < and ~ lexical relations. (1) The parser analyses an input word and provides it with a definition with respect of the word components; it also feeds lexical rules with CFs the input word is composed with. (2) Lexical rules try to match the CFs against the CF Table content, in order to identify the abstract roots basically related to them. (3) According to these collected basic relations, lexical rules predict all the "possible words" the input may be lexically linked to. (4) The last task is then to filter out unattested words from this candidates list. This is what has been fully realized for French, as described in the next section.

3. Results

Results for French have been obtained on a 29,000 nouns, verbs and adjectives specialized medical lexicon. The language specific morphosemantic parser for French (§2.1) is DériF ("Dérivation en Français") ([17]), which makes use of the CF Table content (§2.2) to provide each input lemma with a linguistic-based ([18]) recursive and hierarchical analysis, whose result is threefold: it includes the parsing trace, under square brackets, the ordered list of results, from the input to an indecomposable unit, and the definition of the input, expressing in natural language the semantic relation between the input and its morphological base (Table 3 (1)). Lexical relation computation rules (§2.3) are then applied to (Y,X) CF pairs, which correspond to DériF analysis of each compound word input A, in order to collect all possible (Y',X') links, by matching Y and X against the appropriate entries in the CF Table. Each computed relation is displayed together with the corresponding candidate (Y',X') pairs (Table 3 (2)). To identify which of the candidate (Y',X') relations are actual words, the system first instantiates (Y',X') into the French pair (Y'_{FR}, X'_{FR}) according to CF Table. Then it examines each of the parsing results from the input lexicon. For each compound word B,

(Y'_{FR},X'_{FR}) is compared to B components. In case of Y'_{FR}/Y_B and X'_{FR}/X_B identity, B is added to the semantic family of A, with the appropriate relation (Table 3 (3)).

From a quantitative point of view, the following can be said about results obtained so far for French. First, DériF implement various word formation processes, which comprise about 30 suffixations, prefixations and compounding rules. It analyses 17,240 lemmas as complex words, out of the 29,000 lemmas, and each of them is provided with a definition relating it to its base, or to its components. Finally, the processing chain defines more than 3,000 lexical families among compound nouns and adjectives that are included in the 29,000 entries lexicon.

4. Assessment

A quantitative evaluation of the method against a Gold Standard has been performed in the following way: a sample of 100 words and their family have been randomly chosen among these 3,000 results. Two blind manual assessments (the first one by a linguist, the second one by a medical expert) have been performed in order to check their validity. Each expert has access to all known medical dictionaries and similar resources available today, when performing their task and checking the definitions.

Each assessment validated (1) the definition assigned to each term, and (2) the lexical relation predicted between the term and each member of its family. As far as definition is concerned, three results (and thus three values) were expected for each validation, and by each expert:

1) Full agreement : the prediction entirely meets the expectation

2) The result is wrong, but perfecting it only requires small DériF data or program modifications

3) The result is wrong, and any program improvement would be too costly.

Adding 1 and 2 positive results amounts to estimate the system limits.

As for lexical relations, validation is straightforwards for synonymy or hyponymy: the value

is “yes” or “no”. Proximity is harder to assess. It currently approximates any lexical relation other than synonymy and hyponymy. So, it may relate both antonymous terms (*héméralopie “night blindness”/ nyctalopie “day blindness”*[8]) and words whose definition truly relates them back to eachother (*ovariectomie / salpingectomie*). For sake of simplicity, a positive mark is given only when the latter case is met, according to expert judgement.

Assessment results are summarized in Table 4. It can be seen that the system is able to produce up to 77.3% correct definitions to unknown terms. Moreover, we reached a prediction of almost 70% percent of correct synonymy links. On the other hand, the hyponymy links are very poorly represented in our sample and no figure can be presented. Finally, the proximity links have raised a problem of interpretation between the experts, and it was difficult to reach a common understanding of what means a proximity between lexical entries; such a notion is indeed largely dependant of the intended use of proximity links. Therefore we had to renounce to the validation of this point, until a more concise basis is available, as well as further developments.

Assessments also confirm that the system main drawback lies in the delta that may occur between linguistic predictions and real meanings or lexical links. For instance, the current, lexicalized interpretation of *microstomie* (“*stenostomia”*), (“congenital pathology characterized by an abnormal narrowness of the mouth”), is no longer computable from that of STOMI (“opening”), unlike e.g. *périnéostomie* (“(surgical) STOMI= opening of PERINEE = pelvic floor”). In the same way, the wrong hyponymy relation computed between *appendicite* and *enterite* (via rule **R1**, cf. section 2.3) is due to the specialization of the former term. Such a situation makes necessary the preparation of a list of exceptions for the consolidation of the rule-based system, as shown in [19].

5. Discussion

[8] The meanings of „nyctalopie/nyctalopia“ and „héméralopie/hemeralopia“ are opposite in French and English.

The here presented approach enables the morphological grouping of medical compound words into semantic families, according to a basic multilingual classification (the CF table) set up on the basis of international terminologies of the biomedical domain. A few language-independent computation rules defining lexical relations are required to project the basic CF relations onto the words containing the CFs. Our approach distinguishes from that of [11], [20] and [12], although their work also makes use of word decomposition and CF (called 'subwords') in order to enhance (cross-language) IR recall. However, they do not provide structure to the semantic decomposition of morphologically complex words. This limits the precision of its semantic representations and of its usages for medical language processing. On the other end of the spectrum, conceptual representations such as GALEN [21] are much more precise and structured, but require human, knowledge-intensive definition of each concept.

The quality of the prediction of lexical relations did not met our expectations, except for the synonymy links. The reason is certainly the lack of a more formal definition of what is a proximity relation and the difficulty to provide a judgement outside of a context of usage of such relations. Therefore, we had to postpone this point until further developments are performed.

Results obtained so far for French are used in the framework of the UMLF and VumeF projects. Literal definitions for morphologically constructed terms are in the process of being integrated in the CiSMeF[9] gateway to French-language Health Internet resources [22].

As far as information retrieval and terminology-morphology synergy are concerned, the following can be said:

- Terminology descriptions, and hence extraction of term variants are already enhanced by new links. For instance, having via morphological analysis a relation between

hepatique$_{ADJ}$ (‘hepatic’) and *foie*$_{NOM}$ (‘liver’) enables to relate *maladie du foie* (‘liver disease’) et *maladie hépatique* (‘hepatic disease’). Such a link could not be envisaged from terminology acquisition methods based on pure stemming rules. Other additional direct links are obtained with synonymy, hyponymy and proximity relations from rules in Table 2.

- Another way to (indirectly) connect terms, and thus to enhance term matching systems, is through the literal definition computed by the French morphological analyser DériF: having e.g. *antigastralgique*$_{ADJ}$ (‘antigastralgic’) automatically formulated as “*contre la gastralgie*” (‘against gastralgy’) and thus “*contre la douleur à l’estomac*” (‘against stomach pain’), comes to draw up the following synonymy links: “*traitement* (‘treatment’) *contre la douleur à l’estomac*” = “*traitement contre la gastralgie*” = “*traitement antigastralgique*”.
- Currently, we are crossing compound lexemes literal definitions (*gastralgie* : “*douleur à l’estomac*”, *abdominodynie* : “*douleur à l’abdomen*” (‘abdomen pain’)) with lexical relations (*gastralgie* : hyponym of *abdominodynie*), in order to check synonymy, hyponymy and proximity links reusability between pluriwords terms. For instance, we wish to test to which extend pairing such as the following are valid: “*gastralgie* (and thus *douleur à l’estomac*) is a subtype of “*douleur à l’abdomen*”.
- Another possible application is to use hyponymy links in order to solve discourse anaphora (*"… gastralgy. Treating this pain …"*). By supplying this technique with features taken from Multilingual CF Table, we increase candidate anaphoras : *hysterectomie*$_{NOM}$ can be anaphorically linked to either *ablation*, French translation of ECTOMI, or *acte medical*, i.e. ECTOMI semantic type, according to SNOMED classification (see Table 1).

  Extending the approach to other languages requires only the availability of language-

[9] URL : http://www.cismef.org/

dependent morphosemantic parsers**,** that can be reduced e.g. to simple stemmers in a first approach, as their primary task is to recognize Y and X components for compound words[10]. As soon as it is available for EN, GE, IT and ES, together with monolingual lexical resources necessary to feed the system, the processing chain, already operative for French, will produce a set of cross-language lexical families as illustrated in Fig.1: abstract labels (GASTRALGI) identify groups of words that hold one of the =, <, ~ relations with other multilingual groups of words, when attested in each language-specific lexicon.

6. Conclusion

The originality of the linguistic-based approach we have presented lays both in the computation of cross-linguistic lexical relations, and in the computed definition each decomposed word is provided with. Of course, in counterpart, its strongest drawback is that it requires lists of exception to be accounted for: a human validation is therefore necessary, first to check the appropriateness of morphological decompositions, second to validate computed definitions, and finally to verify the basic relations in the CF Table[11].

However, the performed assessment of the method has proved that the provided definitions are correct in more than 77% of the cases (and probably partially correct in other cases), and that 70% of computed synonymy relations give right predictions. Ongoing researches by the authors show how useful is this method when applied to the discovery of corresponding words between several languages in a multilingual lexicon, allowing thus cross-linguistic question-anwering or information retrieval, multilingual terminological enhancing, multilingual translations.

7. Acknowledgments

[10] The fact that German makes less use of CFs than the other European languages (e.g. *Schädigung* is employed instead of *pathie*, in *Aderhautschädigung* (*choriopathy*)) may imply in the end a less important amount of lexical links with this language.

[11] Among them, proximity relations have to be dealt with carefully in the CF Table, in order to reduce disputable projections, e.g. *gastr ~ hépat* implying *gastritis ~ hepatitis*.

Thanks to UMLF project members for insights and guidance: S. Darmoni and P. Zweigenbaum.

8. Address for correspondence

Fiammetta Namer, UMR 7118 "ATILF" & Université Nancy2 - CLSH – 23 Boulevard Albert 1er, BP3397 – 54015 Nancy Cedex. email: *Fiammetta.Namer@univ-nancy2.fr*; URL:http://www.univ-nancy2.fr/pers/namer

9. References

[1] Zweigenbaum, P., et al., *UMLF: a unified medical lexicon for French.* International Journal of Medical Informatics, 2005. **74**(2-4): p. 119-124.

[2] Darmoni, S.J., et al. *VumeF: Extending the French part of the UMLS.* in *the American Medical Informatics Association (AMIA) Symposium*. 2003. Washington, DC: AMIA.

[3] Volk, M., et al., *Semantic Annotation for concept-based cross-language medical information retrieval.* International Journal of Medical Informatics, 2002. **67**(1-3): p. 97-112.

[4] Fabry, P., et al., *Amplification of Terminologica anatomica by French language terms using Latin terms matching algorithm: A prototype for other languages.* International Journal of Medical Informatics, 2005.

[5] Marko, K., S. Schulz, and U. Hahn, *MorphoSaurus -- design and evaluation of an interlingua-based, cross-language docuyment retrieval engine for the medical domain.* Methods of Information in Medecine, 2005. **44**(4): p. 537-545.

[6] Lovis, C., et al., *Medical dictionaries for patient encoding systems: a methodology.* Artificial Intelligence in Medecine, 1998. **14**: p. 201-214.

[7] Lovis, C., *Trends and pitfalls with nomenclatures and classifications in medecine.* International Journal of Medical Informatics, 1998. **52**(1-3): p. 141-148.

[8] Dorlands, W.A.N., *Dorland's Illustrated Medical Dictionary, 30th edition,* Saunders, Editor. 2002, Saunders: London.

[9] Iacobini, C., *Distinguishing derivational prefixes from initial combining forms*, in *Proceedings of First Mediterranean Morphology Meeting (19-21 sept. 1997)*, G. Booij, A. Ralli, and S. Scalise, Editors. 1999: Mytilene (Greece). p. 132-140.

[10] Lovis, C., et al. *Word segmentation processing: a way to exponentially extend medical dictionaries.* in *8th World Congress on Medical Informatics*. 1995.

[11] Schulz, S., et al. *Towards a multilingual morpheme thesaurus for medical free-text retrieval.* in *Proceedings of MIE'99*. 1999. Ljubliana, Slovenia: IOS Press.

[12] Hahn, U., et al., *Subword segmentation: Leveling out morphological variations for medical document retrieval.* Journal of American Medical Informatics Ass, 2001. **8(suppl)**: p. 229-233.

[13] Daille, B., C. Fabre, and P. Sébillot, *Applications of computational morphology*, in *Many Morphologies*, P. Boucher, Editor. 2002, Cascadilla Press: Somerville, MA. p. 210-234.

[14] Namer, F. and P. Zweigenbaum. *Acquiring meaning for French Medical Terminology: contribution of Morphosemantics*. in *11th MEDINFO*. 2004. San Francisco, CA.

[15] Namer, F. *Acquiring Lexical Classes in Biomedical Lexicons: a Morphosemantics-based Multilingual Approach*. in *International Colloquium on "Word Structure and Lexical Systems: models and applications", (nov. 16-17th, 2004)*. 2004. Università di Pavia, Pavia, Italy.

[16] Namer, F. *Morphosémantique pour l'appariement de termes dans le vocabulaire médical: approche multilingue*. in *TALN 2005 (6-10 juin 2005)*. 2005. Dourdan: ATALA.

[17] Namer, F., *Automatiser l'analyse morpho-sémantique non affixale: le système DériF*, in *Cahiers de Grammaire*, N. Hathout, et al, Eds. 2003, ERSS: Toulouse. p. 31-48.

[18] Corbin, D., *Morphologie dérivationnelle et structuration du lexique.* 1987, Lille: Presses Universitaires de Lille.

[19] Baud, R., et al., *The power and limits of a rule-based morpho-semantic parser.* Journal of American Medical Informatics Association, 1999. **6(suppl)**: p. 22-26.

[20] Schulz, S. and U. Hahn, *Morpheme-based, cross-lingual indexing for medical document retrieval.* International Journal of Medical Informatics, 2000. **58-59**: p. 87-99.

[21] Trombert-Paviot, B., et al., *GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures.* International Journal of Medical Informatics, 2000. **58-59**: p. 71-85.

[22] Douyère, M., et al., *Doc'CISMEF: un outil de recherche Internet dirigé vers l'enseignement de la médecine.* Document Numérique, 2003. **7**(1-2): p. 129-140.

Table 1 - Multilingual Combining Forms Table

CF (1)		Instantiation (2)					POS (3)	Semantic Type (4)	Lexical relation (5)
		English	**German**	**French**	**Italian**	**Spanish**			=STOMAC,
GASTR	realization translation	gastr stomach	Gastr Magen	gastr estomac	gastr stomaco	gastr estomago	N	*ANATOMY*	←ABDOMIN, ~ENTER , ~HEPAT, ~PANCREAT
ALGI	realization translation	algia/algy pain	algie Schmerz	algie douleur	algia dolore	algia dolor	N	*SYMPTOM*	=ODYN, ~ITE, ~OSE
ITE	realization translation	itis inflammation	ite Inflammation	ite inflammation	ite infiammazione	itis inflamación	N	*SYMPTOM*	~ALGI, ~ODYN
PHLEB	realization translation	phleb vein	Phleb Vene	phléb veine	fleb vena	fleb vena	N	*ANATOMY*	=VEN, <ANGI, <VASCUL
ANGI	realization translation	angio blood vessel	Angio Blutgefäß	angio vaisseau sanguin	angio vaso sanguigno	angio vaso sanguíneo	N	*ANATOMY*	=VASCUL, ~VAS
ECTOMI	realization translation	ectomy ablation	ektomie Ablation	ectomie ablation	ectomia ablazione	ectomía ablación	N	*MEDICAL ACT*	~TOMI

Table 2 - Language independent Lexical Relation Computation Rules

Rule	Example		
	Y	**X**	**$[Y_AX_A]$ R $[Y_BX_B]$**
R1 A = $[Y_AX]$ and B=$[Y_BX]$ If $Y_A \leftarrow Y_B$ then A < B else if Y_A R Y_B and R is { =, <, ~} then A R B	PROCT ← COLO LEUCO ~ HÉMATO ABDOMIN=LAPAR ALBUMIN<PROTEIN XER ~SCLER	ECTOMI GRAMME SCOPIE EMIE OPHTALMIE	**EN:** proctectomy < colectomy **GE:** Leukogramm ~ Hämatogramm **FR:** abdominoscopie = laparoscopie **IT:** albuminemia < proteinemia **ES:** xerophthalmia ~sclerophthalmia
R2 A = $[Y_AX_A]$ and B=$[Y_BX_B]$ and $X_A = X_B$ If $Y_A \leftarrow Y_B$ then A < B else if Y_A R Y_B and R is { =, <, ~} then A R B	ENTER←ABDOMIN MORT = THANAT API < ENTOMO CANCER ~CARCIN	**$X_A = X_B$** ALGIE = ODYNIE FERE = GENE VORE = PHAGE FORME = OÏDE	**EN:** enteralgia < abdominodynia **IT:** mortifero = tanatogeno **FR:** apivore < entomophage **GE:** cancriformis ~Karzinoïd
R3 A = $[YX_A]$ and B=$[YX_B]$ if X_A R X_B and R is {=, <, ~} then A R B	BACTÉR OTO ARTHR	OÏDE = FORME RRAGIE < RRHEE ALGIE ~ITE	**FR:** bactérioïde = bactériforme **GE:** Otorrhagie < Otorrhö **ES:** artralgia ~artritis
R4 A = $[Y_AX_A]$ and B=$[Y_BX_B]$ and $Y_A = Y_B$ if X_A R X_B and R is {=, <, ~} then A R B	**$Y_A = Y_B$** ORTHO = RECTI MÉTR = HYSTÉR LIP = ADIP	DONTE = DENT RRAGIE < RRHEE MATOSE ~OME	**FR:** orthodonte = rectident **FR:** métrorragie < hystérorrée **EN:** lipomatosis ~adipoma

Table 3 - Analysis and lexical family for hystérroragie (hysterrorrhagia)

(1)	hystérorragie/N=>[[hystéro N*] [rragie N*] N] , (hystérorragie/N, rragie/N*), " uterus bleeding "
(2) R_{POSS}:Y',X'	Constituents = /hystéro/rragie/ Type = symptom Poss. Rels : (=:HYSTER/RRHAGI), (<:HYSTER/RRHE), (=:METR/RRAGI),(=:UTER/RRAGI), (~:COLP/RRAGI,) (~:FALLOP/RRAGI)
(3) R_{ACT}:Y_B,X_B	hystérorragie/N (symptom) , " bleeding of the uterus " Actual Relations : synonym of métrorragie/N; subtype of hystérorrhée/N, métrorrhée/N; see also colporragie/N

Table 4 - 100 results sample assessment against a Gold Standard

Definitions			**Synonymy Relations**		
Full agreement	**Small data/rules enhancement required**	**Wrong**	**Full agreement**	**Small data/rules enhancement required**	**Wrong**
52.6%	24.7%	22.7%	69.8%	15.8%	14.4%
Minimal expected rate of correct definitions : 77.3%					

Fig.1 - Abstract cross-language family of GASTRALGI