Acquisition of elementary synonym relations from biological structured terminology

Thierry Hamon[1] and Natalia Grabar[2]

[1] LIPN – UMR 7030, Université Paris 13 – CNRS, 99 av. J-B Clément,
F-93430 Villetaneuse, France
thierry.hamon@lipn.univ-paris13.fr
[2] Université Paris Descartes, UMR_S 872, Paris, F-75006 France;
INSERM, U872, Paris, F-75006, France
natalia.grabar@spim.jussieu.fr

Abstract. Acquisition and enrichment of lexical resources have long been acknowledged as an important research in the area of computational linguistics. Nevertheless, we notice that such resources, particularly in specialised domains, are missing. However, specialised domains, *i.e.* biomedicine, propose several structured terminologies. In this paper, we propose a high-quality method for exploiting a structured terminology and inferring a specialised elementary synonym lexicon. The method is based on the analysis of syntactic structure of complex terms. We evaluate the approach on the biomedical domain by using the terminological resource `Gene Ontology`. It provides results with over 93% precision. Comparison with an existing synonym resource (the general-language resource `WordNet`) shows that there is a very small overlap between the induced lexicon of synonyms and the `WordNet` synsets.

1 Background

Acquisition and enrichment of lexical resources have long been acknowledged as an important research in the area of computational linguistics. Indeed, such resources are often helpful for the deciphering and computing semantic similarity between words and terms within tasks like information retrieval (especially query expansions), knowledge extraction or terminology matching.

We make the distinction between terminological and lexical resources. The aim of terminological resources is collecting terms used in a specialised area, describing and organizing them. Within terminologies, terms can be simple (*reproduction*) but mostly complex (*formation of catalytic spliceosome for first transesterification step*; *cell wall mannoprotein synthesis*). They can be linked between them with semantic relations (hierarchical, synonymous, ...). Other features of terms (*i.e.*, definitions, areas of usage) can be precised. As for lexical resources, they gather mostly simple lexical units (*i.e.*, synonyms like *formation*, *synthesis* and *biosynthesis*). These units can belong to common language or be specific to some specialised languages. They can receive descriptions (syntactic, phonetic, morphological, ...) or propose relations between them. Our observation is that units from lexical resources, being simpler linguistic units, are often parts of terms and are spontaneously used during their creation. If terms, usually complex (*i.e.* synonyms like *aromatic amino acid family biosynthesis* and *aromatic amino acid family formation*), can be hardly generalized for being used in various tasks of computational linguistics, their components (*biosynthesis* and *formation* in the given example) are more suitable candidates for the building of a lexicon and their use in natural language processing applications.

Synonym lexicon, as well as lexicon of morphological or orthographic variants, can be used for the task of deciphering semantic relations between terms or words. But not all of these resources are equally well described for various specialised languages and this observation is also true for specialised domains. We are concerned with this remark as our special interest is related to the biomedical domain.

Thus, the morphological description of languages is the most complete and several languages are provided with at least inflectional lexica (widely used within syntactic tools for POS-tagging and lemmatisation [1–3]), or even specific databases (such as `Celex` base [4] for inflectional and derivational description of English and German, or `MorTal` [5] for French). As for the biomedical domain, we can mention the widely used `UMLS` Specialized Lexicon [6] for English, and similar resources for German [7] and French [8].

But when one looks for the description of synonymous or orthographic relations, little available resources can be found. If `WordNet` [9] proposes synonym relations for English, the corresponding resources for other languages are not freely available; while the initiative for tuning this resource for the medical area [10] is still ongoing. Moreover, it has been shown that general lexica, for instance `Wordnet`, are insufficient for specialised knowledge extraction [11]. Indeed, additional specialised information is crucial to improve the coverage and the completeness of the extraction based on general-language resources. To find a solution for this, we propose to use specialised terminologies, as several of them are created and continuously updated in biomedical area. In this work, we propose a novel high-quality method for the acquisition of lexical resources of synonyms from structured terminologies. This method is language-independent. It is based on the identification of syntactic invariants. As indicated, we position our research in the domain of biology.

In the following of this paper, we start with the presentation of the material used (sec. 2), we present then the undergoing hypothesis and various steps of the method proposed (sec. 3). We describe and discuss the obtained results (sec. 4) and conclude with some perspectives to this work (sec. 5).

2 Material

2.1 Structured terminology of biology: `Gene Ontology`

In the current work, we use the `Gene Ontology` (*GO*) [12] as the original resource from which elementary synonym relations are inferred. The goal of the *GO* is to produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and their products in any organism. The project started in 1998 as a collaboration between databases of three model organism: fly *Drosophila*, yeast *Saccharomyce* and mouse. Since then, *GO* is used and enriched by other databases (genomes of plants, animals and micro-organisms).

GO terms describe one of three types of biological meanings, structured into three hierarchical trees: biological processes, molecular functions and cellular components. These trees have been chosen because they represent knowledge useful for the functional annotation of the majority of organisms and can be used for the description of genes and their products from various species. Terms are structured through three types of relations: subsumption `is-a`, partonomy `part-of` and synonymy.

The used version of *GO* contains 18,315 terms linked with 24,537 `is-a` relations and 2,726 `part-of` relations. These terms have 13,850 synonyms. The whole set of terms contains 23,899 terms, both preferred and synonyms.

In our work, we use the synonymous relations between terms.

2.2 General-language resource: `WordNet`

`WordNet` [9] is a large lexical database of English, developed and maintained at Princeton University since 1985, and adapted to other languages. Within this database, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (called *synsets*), each expressing a distinct concept. Synsets are interlinked by means of semantic and lexical relations. The version used provides 81,426 noun synsets, 13,650 verb synsets, 18,877 adjectival and 3,644 adverbial synsets.

`WordNet` synsets are used for the evaluation of coverage of the inferred resource.

3 Methods

3.1 Preliminary observations: compositionality of *GO* terms

Often within *GO*, terms are coined on the same scheme which can be exploited in order to induce the elementary relations between words or simple terms. For instance, the *GO* concept GO:0009073 contains the following series of terms, which show the compositionality through the substitution of one of their components (underlined in the examples):

aromatic amino acid family biosynthesis
aromatic amino acid family anabolism
aromatic amino acid family formation
aromatic amino acid family synthesis

On the basis of this set, it is possible to exploit the compositional structure of terms and thus to induce the following paradigm of synonymous words (or simple terms):

biosynthesis, *anabolism*, *formation*, *synthesis*

In the following, we call the series of synonym terms *original* synonym relations; and series of their substituted components *induced* or *elementary* synonym relations.

We propose a method for the generalization of this observation in order to allow acquiring specialised lexicon of elementary synonymous relations. Like in the given examples, the method exploits the compositional structure of terms and relies on existence of structured terminologies. The notion of compositionality, central for this method, assumes that the meaning of a complex expression is fully determined by its syntactic structure, the meaning of its parts and the composition function [13]. We propose to apply this principle for building a lexicon of elementary synonym relations. Moreover, it has been observed that a large part of *GO* terms indeed verify the compositionality principle [14].

In order to be able to exploit this principle, terms are first analysed syntactically into head and expansion components (sec. 3.2), then specific inference rules are applied (sec. 3.3), and the obtained results are evaluated (sec. 3.4).

3.2 Preprocessing of terminology

The aim of terminology preprocessing step is to provide the syntactic analysis of terms. Such analysis is crucial for our work: the method we propose exploits syntactic dependency relations and is based on syntactic invariants. Hence, each *GO* term must be linguistically analysed in order to prepare and perform syntactic analysis.

In our work, we use the `Ogmios` platform [15], which is suitable for the processing of large amount of data and, moreover, can be tuned to a specialised domain. Through the platform, several types of linguistic processing are performed. First, the `TagEN` [16] tool is applied for the recognition on named entities. Its use at the beginning of linguistic pipeline helps the forthcoming segmentation into words and sentences. Indeed, the recognition of named entities (*i.e.*, gene names, chemical products) allows disambiguating special characters, such as punctuation marks, dashes, slashes, etc, widely used within named entities in biology and often altering the segmentation into word and sentence. After the segmentation, the POS-tagging and lemmatisation are performed with the `GeniaTagger` [17] tool, specifically trained for the biomedical domain.

The step of syntactic parsing of terms is carried out thanks to the rule-based term extractor YATEA [18]. The syntactic dependency relations between term components are computed according to assigned POS tags and parsing rules implemented within YATEA. Thus, each term is considered as a syntactic binary tree (see figure 1) composed of two elements: head component and expansion component. For instance, *anabolism* is the head component of *acetone anabolism* and *acetone* is its expansion component.

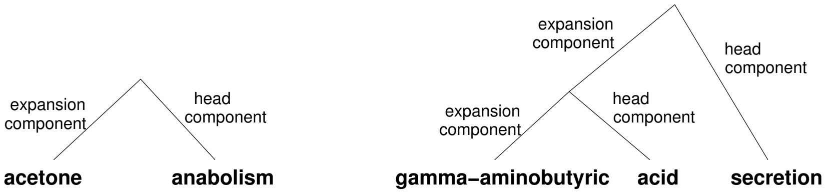

Fig. 1. Parsing tree of the terms *acetone anabolism* and *gamma-aminobutyric acid secretion*.

3.3 Acquisition of synonym lexicon

The present method is inspired by our previous work [11], where we proposed to apply the semantic compositionality principle for inferring synonymy relations between complex terms. We then postulated that the composition process preserves synonymy and that the compositionality principle holds for complex terms. Roughly, this means that if the meaning $\mathcal{M}$ of two complex terms $A\ rel\ B$ and $A'\ rel\ B$ are given by the following formulas :

$$\mathcal{M}(A\ rel\ B) = f(\mathcal{M}(A), \mathcal{M}(B), \mathcal{M}(rel))$$

and

$$\mathcal{M}(A'\ rel\ B) = f(\mathcal{M}(A'), \mathcal{M}(B), \mathcal{M}(rel))$$

for a given composition function f, and if A and A' are synonymous ($\mathcal{M}(A) = \mathcal{M}(A')$), then the synonymy of the complex terms can be inferred:

$$\begin{aligned} \mathcal{M}(A'\ rel\ B) &= f(\mathcal{M}(A'), \mathcal{M}(B), \mathcal{M}(rel)) && (1) \\ &= f(\mathcal{M}(A), \mathcal{M}(B), \mathcal{M}(rel)) && (2) \\ &= \mathcal{M}(A\ rel\ B) && (3) \end{aligned}$$

In the current work, we assume that the inverse function f^{-1} exists and can be applied for deducing elementary synonym relations given synonymous complex terms. As in the cited work [11], our approach takes into account the internal structure of the complex terms. We assume that the syntactic dependency relation between components is preserved through the compositionality principle. Thus, we can infer elementary synonym relations between components of two terms if:

- parsed terms are synonymous;
- these components are located at the same syntactic position (head or expansion);
- the other components within terms are either synonymous or identical.

The fully parsed terms are represented as a terminological network, within which the deduction of the elementary synonym relations is based on the three following rules:

Rule 1 If both terms are synonymous and their expansion components are identical, then an elementary synonym relation is inferred. For instance, we can infer the synonym relation {*B-lymphocyte*, *B-cell*} from the original synonym relation between terms:
peripheral B-lymphocyte and *peripheral B-cell*
where the expansion component *peripheral* is identical in both terms.

Rule 2 If both terms are synonymous and their head components are identical, then an elementary synonym relation is inferred. For instance, we infer the synonym relation {*endocytic*, *endocytotic*} from the synonym relation between terms:
endocytic vesicle and *endocytotic vesicle*
where the head component *vesicle* is identical.

Rule 3 If both terms are synonymous and either their head components or expansion components are synonymous, then an elementary synonym relation is inferred. For instance, we infer the synonym relation {*nicotinamide adenine dinucleotide*, *NAD*} from the synonym relation between terms:
nicotinamide adenine dinucleotide catabolism and *NAD breakdown*
where the head components {*catabolism*, *breakdown*} are already known synonyms.

The method is recursive and each inferred elementary synonym relation can then be propagated in order to infer new elementary relations, which allows to generate a more exhaustive lexicon of synonyms.

3.4 Evaluation

We perform manual validation of the inferred elementary relations between words and simple terms. For this, each pair is examined, as well as its source series of synonyms. Accuracy of the inferred pairs is thus computed. Moreover, we make an attempt to compare the inferred resource with `WordNet` synsets and compute the overlap between them. The both sets of synonyms are compared once lemmatised.

4 Results and Discussion

4.1 Preprocessing of terminology: Ogmios platform

23,899 *GO* terms have been fully parsed through the platform Ogmios. Thus, 15,863 original synonym relations could be used for inferring elementary relations.

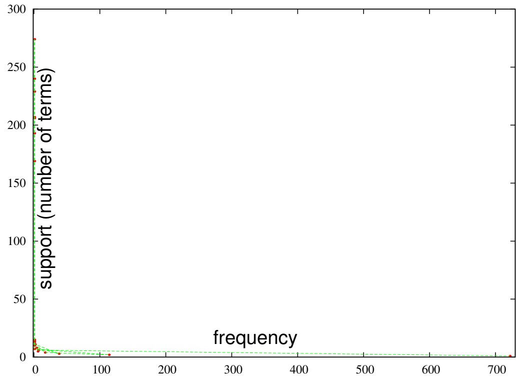

Fig. 2. Support and frequency observed when inferring elementary synonymous relations from *GO* terms.

4.2 Acquisition of synonym lexicon

The three rules defined for inferring elementary relations have been applied to the terminological network formed with 15,863 original *GO* synonym terms. In this way, 921 pairs of elementary synonym relations have been induced.

Our general observation is that, among these inferred pairs, very few (around ten) are induced from a large number of original *GO* synonyms, while the most of inferred pairs are supported by a small number of *GO* terms. For instance, 274 *GO* synonymous series allow to infer the pair {*breakdown*, *catabolism*}, which thus appears to be a fundamental notion in biology. The pair {*formation*, *synthesis*} is acquired on 240 original terms, the pair {*catabolism*, *degradation*} on 229 term pairs, etc. Pairs like {*adrenaline*, *epinephrine*}, {*gallate*, *gallic acid*}, {*formation*, *growth*}, {*flagella*, *flagellum*}, {*F-actin*, *actin filament*}, {*eicosanoid*, *icosanoid*} are acquired on small number of original term pairs (1 to 3). Such pairs correspond mainly to chemical products, to Latin inflected words or to orthographic variants. These represent nearly 80% (n=722) of the whole number of inferred synonym pairs. They may show smaller semantic acceptance of their paradigms, but this should be verified through their implementation within corpora and applications.

Figure 2 represents this observation graphically by combining figures of support of inferred pairs (number of original *GO* synonyms that allow to infer them) and of frequency of each support value. We can see that the inducing of elementary relations from *GO* terms follows a hyperbolic distribution. Although such distribution is observed in language and is often referred to as Zipf law, we assume that in our experience this situation is also due to the strong policy used by `Gene Ontology` Consortium. As a matter of fact, creation of new terms is governed by *GO* guidelines[3] and the vocabulary (*GODict.DAT*) of words already used within *GO* terms.

Another observation we can make on the basis of this data is that the compositionality is indeed a widely verified principle within *GO* terms, as it was observed in previous work [19, 14, 20]. In our experience, the large values of support confirm this and attest that the compositionality is indeed applied at large scale for coining new *GO* terms.

Additionally, the acquired synonym pairs can be classified according to their linguistic or semantic types.

Linguistic typology of synonym pairs. The linguistic types of elementary synonymous relations can be defined further to a manual analysis:

- Orthographic variants:
  {*synthase*, *synthetase*}, {*leucocyte*, *leukocyte*}, {*sulfate*, *sulphate*}
- Hyphenation variants, which can be considered as part of orthographic variants but have the specificity of being always concerned with the same type of variation (presence or absence of the hyphen):
  {*B-cell*, *B cell*}
- Word ordering: {*gamma-delta T-cell*, *T gamma-delta cell*}
- Abbreviations, which are widely used in biological domain, apply various tactics for the coining of abbreviated terms and words:
  - standard abbreviation through acronym formation:
    {*ER*, *endoplasmic reticulum*}
  - acronym formation at morphological level:
    {*DPH*, *dehydropeptidase*}, {*IL-10*, *interleukin-10*}
  - syllabic abbreviation: {*Eph*, *ephrin*}, {*Gly*, *glycine*}
  - combined abbreviation: {*TGase*, *transglutaminase*},
- Use of symbols, which is also very frequent in biological domain:
  {*1-b-glucosyltransferase*, *1-beta-glucosyltransferase*},
  {*D-isomerase*, *delta-isomerase*},
  {*omega-amidase*, *w-amidase*}
- However, most of the induced synonym pairs link entities for which no common formal features can be observed:
  {*hydroxylase*, *monooxygenase*}, {*vitamin Bh*, *myo-inositol*}, {*cell*, *lymphocyte*}, {*apyrase*, *nucleoside-diphosphatase*}, {*myrosinase*, *sinigrinase*}, {*invertase*, *saccharase*}, {*regulator activity*, *modulator*}, {*Valium*, *diazepam*}

The method we propose is specifically useful for the acquisition of this last type of synonyms which are difficult to detect otherwise, *i.e.* on the basis of their formal feature (internal structure, morphology, etc.).

[3] *http://www.geneontology.org/GO.usage.shtml*

Semantic typology of synonym pairs. Semantic types of the inferred pairs of synonyms could be defined according to the hierarchical trees of the `Gene Ontology` (biological processes, molecular functions and cellular components) within which the elementary relations have been inferred.

For instance, the synonym series which show an important support:

- *biosynthesis*, *synthesis*, *formation*, *anabolism*
- *breakdown*, *degradation*, *catabolism*

correspond to fundamental biological processes and remain specific to this hierarchical tree of *GO*.

The pair {*cell*, *lymphocyte*} is specific notion of cellular component tree but is, in fact, widespread over all the *GO* terms: the majority of biological processes and molecular functions are located at the cell level. The same observation can be done for {*ER*, *endoplasmic reticulum*} pair, which stands for a cellular component, but is the place of many biological processes and molecular functions.

As for {*DPH*, *dehydropeptidase*}, {*Eph*, *ephrin*} or {*Gly*, *glycine*} pairs, they are molecular functions inferred from few *GO* series of synonyms.

Such semantic typology, based on the hierarchical organization of *GO*, gives some insights into the language usage within biological domain and, more specifically, within *GO*. We assume a more fine-grained typology can be proposed, distinguishing in addition semantic types like phenotype, chemical products, pathological processes, etc.

Contextual nature of synonymous relations. An additional remark can be made on the nature of the synonymous relations. Like in in [21], we consider synonymy as a Boolean rather than as a scaling property. Thus, we define synonymy as a sort of *contextual cognitive synonymy*: X is a cognitive synonym of Y relatively to a context C if (i) X and Y are syntactically identical, and (ii) any grammatical declarative sentence S containing X in the context C has equivalent truth-conditions to another sentence S', which is identical to S except that, in C, X is replaced by Y [21, p.88]. In this way, our synonymy definition is close to that of `WordNet`: as far as we can observe at least one context within which a pair of words is synonymous we record these words as true synonyms in the inferred lexicon.

For instance, the pair {*cell*, *lymphocyte*} can be observed in several synonymous terms within *GO*:

establishment of B cell polarity; *establishment of B lymphocyte polarity*
T cell homeostatic proliferation; *T lymphocyte homeostatic proliferation*
B-cell homeostasis; *B-lymphocyte homeostasis*
T cell mediated cytotoxicity; *T lymphocyte mediated cytotoxicity*

For this reason, this inferred pair of synonyms is counted as correct, even if the common feeling about it would be that *cell* is a more general term than *lymphocyte*.

The validity of such pairs within other contexts should be verified.

4.3 Evaluation

The manual evaluation, performed by a computational scientist, shown that 93.1% (n=857) are correct, 5.4% (n=50) rejected and 1.5% (n=14) remain undecided. This evaluation is supported by the analysis of both inferred and initial synonym pairs.

The efficiency of the proposed method is very high. This is due to the fact that the acquisition is performed on controlled terminological data. Moreover, the inferring rules strongly exploit syntactic scheme within syntactically analyzed terms. Finally, as we observed, the compositionality principle is widely applied for the coining of new *GO* terms. All these factors can but contribute to the acquisition of high-quality synonym pairs.

We attempted a comparison between the induced elementary synonymous pairs and the synsets provided by `WordNet`. Unsurprisingly, the overlap is very low. As a matter of fact, any of the inferred synonym sets can be completely matched with any of the synsets. Although we can find partial overlapping between these two resources. For instance, the inferred set *biosynthesis*, *synthesis*, *formation*, *anabolism* partly overlaps with the following synsets:

- *biosynthesis*
- *biosynthesis*, *biogenesis*
- *constitution*, *establishment*, *formation*, *organization*, *organisation*
- *formation*, *shaping*
- *formation*
- *anabolism*, *constructive metabolism*
- *synthesis*

and shows to have no common meaning with other synsets, for instance *deduction*, *deductive reasoning*, *synthesis*.

Otherwise, there is difference in namings, for instance {*cell*, *lymphocyte*} in the inferred resource and {*lymphocyte*, *lymph cell*} as proposed by `WordNet`. But, usually, the inferred resource proposes more specialised notions: {*ER*, *endoplasmic reticulum*} and *endoplasm* in `WordNet`. Finally, many of these notions do not occur within `WordNet`.

The difference between the two compared resources is not surprising as the purpose, as well as aimed applications, of the `WordNet` and of the `Gene Ontology` are different. `WordNet`, being the only available resources of synonyms, is sometimes applied in specialised domains. For instance, its use for terminology structuring and knowledge extraction shown that such general lexica are insufficient for specialised domains [11] and should be completed with specialised resources. Indeed, specialised domains make use of concepts too specific to occur within a general language lexicon. The common-language resources have been proposed to be adapted to a given domain through corpus-based filtering, even though they do not represent the richness of this specialised language [22]. Another experiences demonstrated that although the suitability of the general-language resources for biomedical area is low [23, 24], they can be used as layer which could adapt high technical level information to lay people understanding. In this case, definitions as those proposed by `WordNet` are helpful.

5 Conclusions and Perspectives

Although there is a huge need in various types of linguistic resources, some types of such resources are missing especially in specialised domains. For instance, in many areas of the natural language processing synonym resources are widely needed. In this work, we propose a novel method for filling in the gap and inferring elementary synonymous relations. This method exploits the compositionality principle, when it is verified, and relies on existence of structured terminologies. It applies set of rules based on syntactic dependency analysis within terms.

The proposed method has been applied to Gene Ontology, a terminological resource of biology. It provides high-quality results: over 93% of inferred relations prove to be correct. However, the synonymy is as contextual relation and the validity of some inferred pairs should be tested on corpora. The attempted comparison with the available resource of synonym relations, proposed by the WordNet, is very low: in the best case scenario, the overlap is partial. But often the inferred notions are missing in WordNet.

In the next future, we plan to use the inferred synonym relations for enriching and extending the Gene Ontology. We will also test their efficiency with other biomedical terminologies. But we assume this resource can be used in many other applications of computational linguistics.

As we noticed, the method is language-independent, and it is possible to apply it to other languages as far as (1) the required linguistic processing can be realised and (2) synonym relations between complex terms are available. For this purpose, we can use for instance the UMLS resource [6], or more specifically the MeSH [25] or Snomed [26] terminologies which are available in several languages.

References

1. Brill, E.: A Corpus-Based Approach to Language Learning. PhD thesis, University of Pennsylvania, Philadelphia (1993)
2. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK (1994) 44–49
3. Namer, F.: FLEMM : un analyseur flexionnel du français base de règles. Traitement Automatique des Langues (TAL) **41**(2) (2000) 523–547
4. Burnage, G.: CELEX - A Guide for Users. Centre for Lexical Information, University of Nijmegen (1990)
5. Hathout, N., Namer, F., Dal, G.: An experimental constructional database: the MorTAL project. In Boucher, P., ed.: Morphology book. Cascadilla Press, Cambridge, MA (2001)
6. NLM: UMLS Knowledge Sources Manual. National Library of Medicine, Bethesda, Maryland. (2007) www.nlm.nih.gov/research/umls/.
7. Schulz, S., Romacker, M., Franz, P., Zaiss, A., Klar, R., Hahn, U.: Towards a multilingual morpheme thesaurus for medical free-text retrieval. In: Medical Informatics in Europe (MIE). (1999)

8. Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch, P., Duff, F.L., Thirion, B., Darmoni, S.: Towards a Unified Medical Lexicon for French. In: Medical Informatics in Europe (MIE). (2003)
9. Fellbaum, C.: A semantic network of english: the mother of all WordNets. Computers and Humanities. EuroWordNet: a multilingual database with lexical semantic network **32**(2-3) (1998) 209–220
10. Smith, B., Fellbaum, C.: Medical wordnet: a new methodology for the construction and validation of information. In: Proc of 20th CoLing, Geneva, Switzerland (2004) 371–382
11. Hamon, T., Nazarenko, A.: Detection of synonymy links between terms: experiment and results. In: Recent Advances in Computational Terminology. John Benjamins (2001) 185–208
12. Gene Ontology Consortium: Creating the Gene Ontology resource: design and implementation. Genome Research **11** (2001) 1425–1433
13. Partee, B.H. In: Compositionality. F. Landman and F. Veltman (1984)
14. Ogren, P., Cohen, K., Acquaah-Mensah, G., Eberlein, J., Hunter, L.: The compositional structure of Gene Ontology terms. In: Pacific Symposium of Biocomputing. (2004) 214–225
15. Hamon, T., Nazarenko, A., Poibeau, T., Aubin, S., Derivière, J.: A robust linguistic platform for efficient and domain specific web content analysis. In: RIAO 2007, Pittsburgh, USA (2007)
16. Berroyer, J.F.: Tagen, un analyseur d"entits nommes : conception, dveloppement et valuation. Mémoire de D.E.A. d'intelligence artificielle, Universit Paris-Nord (2004)
17. Tsuruoka, Y., Tateishi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S., , Tsujii, J.: Developing a robust part-of-speech tagger for biomedical text. LNCS **3746** (2005) 382–392
18. Aubin, S., Hamon, T.: Improving term extraction with terminological resources. In Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T., eds.: Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006). Number 4139 in LNAI, Springer (2006) 380–387
19. Verspoor, C.M., Joslyn, C., Papcun, G.J.: The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. In: SIGIR workshop on Text Analysis and Search for Bioinformatics. (2003) 51–56
20. Ogren, P., Cohen, K., Hunter, L.: Implications of compositionality in the Gene Ontology for its curation and usage. In: Pacific Symposium of Biocomputing. (2005) 174–185
21. Cruse, D.A.: Lexical Semantics. Cambridge University Press, Cambridge (1986)
22. Grabar, N., Zweigenbaum, P.: Utilisation de corpus de spécialité pour le filtrage de synonymes de la langue générale. In: Traitement Automatique de Langues Naturelles (TALN). (2005)
23. Bodenreider, O., Burgun, A.: Characterizing the definitions of anatomical concepts in WordNet and specialized sources. In: Proceedings of the First Global WordNet Conference. (2002) 223–230
24. Bodenreider, O., Burgun, A., Mitchell, J.A.: Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. In: Medical Informatics in Europe (MIE). (2003) 379–384
25. National Library of Medicine Bethesda, Maryland: Medical Subject Headings. (2001) http://www.nlm.nih.gov/mesh/meshhome.html.
26. Côté, R.A.: Répertoire d'anatomopathologie de la SNOMED internationale, v3.4. Université de Sherbrooke, Sherbrooke, Québec. (1996)