# Discovering Knowledge in Texts for the learning of DOGMA-inspired ontologies

**Marie-Laure Reinberger**[1] and **Peter Spyns**[2]

**Abstract.** Ontologies in current computer science parlance are computer based resources that represent shared conceptualizations for a specific domain. This paper first introduces ontologies in general and subsequently, in particular, shortly outlines the DOGMA ontology leaning approach. The paper also introduces the reader in the field of Knowledge Discovery in Text before, in the main part, work in progress is described and experimentally evaluated. It concerns a potential method to automatically extract concepts and conceptual relationships from texts. Preliminary outcomes are presented based on the clustering of nominal terms and prepositional phrases according to co-occurrence frequencies in the verb-object syntactic context.

## 1 INTRODUCTION

A recent evolution in the areas of artificial intelligence, database semantics and information systems is the advent of the Semantic Web [5]. It evokes "futuristic" visions of intelligent and autonomous software agents including mobile devices, health-care, ubiquitous and wearable computing. E.g., a heartbeat monitoring device integrated in a person's shirt could trigger, in case of observed rhythm deviations, via the mobile network a web agent that schedules an appointment with his/her doctor.

An essential condition to the actual realisation and unlimited use of these smart devices and programs is the possibility for interconnection and interoperability, which is currently still lacking to a large extent. Indeed, intelligent agents have to be able to exchange "meaningful" messages [3] while continuing to function autonomously (interoperability with local autonomy as opposed to integration with central control). Exchange of meaningful messages is only possible when the intelligent devices or agents share a common conceptual system representing their "world" [4], as is the case for human communication. Meaning ambiguity should be, by preference, eliminated. Nowadays, a formal representation of such (partial) intensional definition of a conceptualisation of an application domain is called an ontology [22].

The development of ontology-driven applications is currently slowed down due to the knowledge acquisition bottleneck. Indeed, the process of conceptualising an application domain and its formalisation need substantial human resources and efforts. Therefore, techniques applied in computational linguistics and information extraction (in particular machine learning) are used to create or grow ontologies in a period as limited as possible with a quality as high as possible. Sources can be of different kinds including databases and their schemas - e.g. [42], semi-structured data (XML, web pages), ontologies [5] and texts. Activities in the latter area are grouped under the label of Knowledge Discovery in Text (KDT), while the term "Text Mining" is reserved for the actual process of information extraction [26].

This paper wants to report on a joint research effort on the learning of ontologies from texts by VUB STAR Lab and UA CNTS during the Flemish IWT OntoBasis project [6]. The experiments concern the extraction and clustering of natural language terms into semantic sets standing for domain concepts as well as the detection of conceptual relationships. For this aim, the results of shallow parsing techniques are combined with unsupervised learning methods.

The remainder of this paper is organised as follows. The next section (2) gives an overview of research in the same vein (section 2.1). Methods and techniques including others than the ones applied for this paper are mentioned ( section 2.2). In section 3, a short overview of the DOGMA ontology engineering framework is given as it is the intention that the experiments described in this paper lead to a less time consuming process to create DOGMA-inspired ontologies. The objectives are presented in section 4.1, while the methods and material are explained in section 4.2. The experiments themselves are described in section 4.3 after which the results (section 4.4 ) and related work (section 4.5) are discussed. Indications for future research are given in section 5, and some final remarks conclude (section 6) this paper.

## 2 BACKGROUND

### 2.1 Overview of the field

Several centres worldwide are actively researching on KDT for ontology development (building and/or updating). An overview of 18 methods and 18 tools for text mining with the aim of creating ontologies can be found in [19]. A slightly older, more limited but complementary overview is provided by [29] [7]. It is worth to mention that in France important work (mostly applied to the French language) is being done by members of the TIA ("Terminologie et Intelligence Artificielle") working group of the French Association for Artificial Intelligence (AFIA) [8]. TIA regroups several well known institutes and researchers included in the overview mentioned above [19] and organizes at a regular basis Ontologies and Texts (OLT) workshops linked

[1] CNTS/University of Antwerp - Belgium email: marielaure.reinberger@ua.ac.be

[2] STARLab/Vrije Universiteit Brussel - Belgium email: Peter.Spyns@vub.ac.be

[3] We make abstraction here of the feasibility of physically connecting these devices and services or agents to a (global) network.

[4] See [41] for more details on the semantics of the Semantic Web.

[5] This is called ontology aligning and merging - e.g. [34]

[6] see http://wise.vub.ac.be/ontobasis

[7] We refer the interested reader to these overviews rather than repeating all the names of people and tools here.

[8] http://www.biomath.jussieu.fr/TIA

to major AI-conferences (e.g., EKAW2000 [1], ECAI2002 [2]). Other important workshops on ontology learning were linked to ECAI2000 [40] and IJCAI2001 [28].

In addition to tools and researchers listed in the two overviews, there are the EU IST projects Parmenides [9] and MuchMore [10]. These projects have produced interesting state-of-the-art deliverables on KDT [25] - in particular section 3 - and related NLP technology [33]. The NLP groups of the University of Sheffield and UMIST (Manchester) are also active in this area [7, 26]. A related tool is SOOKAT, which is designed for knowledge acquisition from texts and terminology management [32]. A specific corpus-based method for extracting semantic relationships between words is explained in [15]. Mining for semantic relationships is also - albeit in a rather exploratory way - addressed in the Parmenides project [39].

## 2.2 Overview of methods

In essence, one can distinguish the following steps in the process of learning ontologies from texts (that are in some way or another common to the majority of methods reported):

1. collect, select and preprocess an appropriate corpus
2. discover sets of equivalent words and expressions
3. validate the sets (establish concepts) with the help of a domain expert
4. discover sets of semantic relations and extend the sets of equivalent words and expressions
5. validate the relations and extended concept definitions with the help of a domain expert
6. create a formal representation

Not only the terms, concepts and relationships are important, but equally the circumscription (gloss) and formalisation (axioms) of the meaning of a concept or relationship. On the question how to carry out these steps, a multitude of answers can be given. Many methods require a human intervention before the actual process can start (labelling seed terms - supervised learning, compilation/adaptation of a semantic dictionary or grammar rules for the domain ,...). Unsupervised methods don't need this preliminary step - however, the quality of their results is still worse. The corpus can preclude the use of some techniques: e.g., machine learning methods require a corpus to be sufficiently large - hence, some authors use the Internet as additional source [13]. Some methods require the corpus to be preprocessed (e.g., adding POS tags, identifying sentence ends, ...) or are language dependent (e.g., compound detection). Again, various ways of executing these tasks are possible (e.g., POS taggers can be based on handcrafted rules, machine-induced rules or probabilities). In short, many linguistic engineering tools can be put to use. To our knowledge no comparative study has been published yet on the efficiency and effectiveness of the various techniques applied to ontology learning.

Selecting and grouping terms can be done by means of tools based on distributional analysis, statistics, machine learning techniques, neural networks, and others. To discover semantic relationships between concepts, one can rely on valency knowledge, already established semantic networks or ontologies, co-occurrence patterns, machine readable dictionaries, association patterns or combinations of all these. In [26] a concise overview is offered of commercially available tools that are useful for these purposes. Due to space restrictions, we will not discuss in this paper how the results can be validated (e.g., see [23]) and transformed in a formal model (e.g., see [3] for an overview of ontology representation languages).

[9] http://www.crim.co.umist.ac.uk/parmenides/

[10] http://muchmore.dfki.de/demos.htm

# 3 DOGMA

Before presenting the actual text mining experiments, we want to shortly discuss the framework for which the results of the experiments are meant to be used, i.e. the *DOGMA* (Developing Ontology-Guided Mediation for Agents) ontology engineering approach [11].

To be retained for this paper is the preference within the DOGMA approach given to texts as objective repositories of domain knowledge instead of referring to domain experts as exclusive knowledge sources [12]. Apparently, this preference is rather recent [1] and probably more popular in language engineering circles. The linguistic notion of "representative corpus" can be re-introduced. However, the problems raised earlier within the linguistics community about the criteria to determine (and maintain) the representative character of a corpus might find an easier solution. For ontology engineering purposes, a text is representative if it embodies by definition (e.g., law, norm or imposed reference) or by facts (e.g., product catalogue with descriptions agreed upon by the relevant stake-holders in a business situation) relevant domain knowledge. The corresponding ontology can be considered as descriptive (de facto standard) or prescriptive (de iure standard).

Additionally, notice that also restrictions on a semantic relationship, e.g. indicating its mandatory aspect or its cardinality, should be mined from the corpus. These constraints are called in general "axioms" and serve to define more precisely the concepts and relations in the ontology. This is a step that should be added before the formal model is created, and that currently is hardly mentioned in the KDT literature. But one will easily agree that, e.g. when modelling a law text, there can be a huge difference between "must" and "may".

Finally, it should be noted that, in the near future, a strict distinction in the implementation of the DOGMA ontology server will be made between concept labels and natural language words or terms [12]. In many cases, "term" is interpreted in the ontology literature as "logical term" (or concept) of the ontology first order vocabulary and, at the same time, as a natural language term. Without going too much in detail here, we separate the logical level from the linguistic level (by using WordNet-like synsets - see also [18]), which has its impact on the KDT process, namely in step (3) mentioned in section 2.2. One of the rather rare KDT methods that also takes this distinction into account is described in [30].

# 4 UNSUPERVISED TEXT MINING

In the following sections, we will report on experiments with unsupervised machine learning techniques based on results of shallow parsing.

## 4.1 Objectives

Our purpose is to build a repository of lexical semantic information from text, ensuring evolvability and adaptability. This repository can be considered as a complex semantic network. We assume that the method of extraction and the organisation of this semantic information should depend not only on the available material, but also on the intended use of the knowledge structure. There are different ways of

[11] see http://www.starlab.vub.ac.be/research/dogma

[12] This does not imply that texts will be the sole source of knowledge.

organising this knowledge, depending on its future use and on the specificity of the domain. In this paper, we deal exclusively with the medical domain, but one of our future objectives is to test our methods and tools on different (but specific) domains.

Currently, the focus is on the discovery of concept and their conceptual relations although it is the ultimate aim to discover semantic constraints as well. We have opted for extraction techniques based on unsupervised learning methods [36] since these do not require specific external domain knowledge such as thesauri and/or tagged corpora [13]. As a consequence, the portability of these techniques to new domains is expected to be much better [33, p.61].

## 4.2 Material and methods

The ***linguistic assumptions*** underlying this approach are

1. the principle of selectional restrictions (syntactic structures provide relevant information about semantic content), and
2. the notion of co-composition [35] (if two elements are composed into an expression, each of them imposes semantic constraints on the other).

The fact that heads of phrases with a subject relation to the same verb share a semantic feature would be an application of the principle of *selectional restrictions*. The fact that the heads of phrases in a subject or object relation with a verb constrain that verb and vice versa would be an illustration of *co-composition*. In other words, each word in a noun-verb relation participates in building the meaning of the other word in this context [16, 17]. If we consider the expression "write a book" for example, it appears that the verb "to write" triggers the informative feature of "book", more than on its physical feature. We make use of both principles in our use of clustering to extract semantic knowledge from syntactically analysed corpora.

In a specific domain, an important quantity of semantic information is carried by the nouns. At the same time, the noun-verb relations provide relevant information about the nouns, due to the semantic restrictions they impose. In order to extract this information automatically from our corpus, we used the ***memory-based shallow parser*** which is being developed at CNTS Antwerp and ILK Tilburg [8, 9, 11] [14]. This shallow parser takes plain text as input, performs tokenisation, POS tagging, phrase boundary detection, and finally finds grammatical relations such as subject-verb and object-verb relations, which are particularly useful for us. The software was developed to be efficient and robust enough to allow shallow parsing of large amounts of text from various domains.

Different methods can be used for the ***extraction of semantic information*** from parsed text. Pattern matching [4] has proved to be a efficient way to extract semantic relations, but one drawback is that it involves the predefined choice of the semantic relations that will be extracted. On the other hand, clustering only requires a minimal amount of "manual semantic pre-processing" by the user. We rely on a large amount of data to get results using pattern matching and clustering algorithms on syntactic contexts in order to also extract previously unexpected relations. Clustering on terms can be performed by using different syntactic contexts, for example noun+modifier relations [10] or dependency triples [27]. As mentioned above, the shallow parser detects the subject-verb-object structures, which gives us the possibility to focus in a first step on the term-verb relations with the term appearing as the head of the object phrase. This type of structure features a functional relation between the verb and the term appearing in object position, and allows us to use a clustering method to build classes of terms sharing a functional relation. Next, we attempt to enhance those clusters and link them together, using information provided by prepositional structures.

The choice of the specific ***medical domain*** has been made since large amounts of data are freely available. In particular, we decided to use Medline, the abstracts of which can be retrieved using the internal search engine. We have focused on a medical subject that was specific but common enough to build a moderately big corpus. Hence, the first corpus is composed of the Medline abstracts retrieved under the queries "hepatitis A" and "hepatitis B". It contains about 4 million words. The shallow parser was used to provide a linguistic analysis of each sentence of this corpus, allowing us to retrieve semantic information of various kinds. The second corpus has been extracted from Medline abstract also, using the string "blood" on the search engine. It contains about 7M words, and is less specific than the hepatitis corpus.

## 4.3 Experiments

The study we are reporting here has taken place after previous experiments we have made in the field of unsupervised clustering. We have carried out different comparative studies involving several clustering algorithms applied to parsed text or raw text, on different corpora of various size and specificity [37, 38, 36]. The results of those studies have shown that the verb-object dependency tends to be more informative than the subject-verb dependency, and that we improve our results by applying a hard clustering method on nominal terms selected according to their frequency. This hard clustering allows one term to belong to one and only one cluster.

Therefore, the ***first step*** of the experiment reported here consists in applying a clustering algorithm on a set of terms retrieved from the output of the shallow parser. As the shallow parser provides us with syntactic structures subject-verb-object, we select from these structures the relation verb-object, and more precisely the association verb-term, the term being here the head of the object phrase and composed of a string of adjectives and nouns. What we get is a list of verb-term relations, from which we select the most frequently co-occurring relations. We use for this selection a probabilistic measure $P(t \mid v)$ that considers, given a verb v, the probability of occurrence of a dependency verb-term (v-t):

$$P(t \mid v) = f(t, v)/f(v) \quad (1)$$

The relations selected are organized in classes. Each term is associated to the set of verbs which are the most relevant according to the statistical measure. Hence we get a set of classes of verbs, each of them associated to a different term. Note that a verb may be associated to more than one term, and therefore appear in more than one class.

Subsequently, a clustering algorithm is applied to the classes of verbs. As each class of verbs is associated to a term, this clustering will build at the same time classes of terms, but a term will only belong to one cluster. By performing this clustering, we mean to exploit the functional relation that lies between a verb and its direct object. This naive clustering algorithm is based on the similarity between two classes of verbs. It will join terms two by two during the first pass. In the next passes, the sets of terms are joined two by two. The similarity depends simply on the number of common verbs and the number of differing verbs in two sets.

[13] Except the training corpus for the general purpose shallow parser - see below.

[14] See `http://ilk.kub.nl` for a demo version.

Many clusters obtained with the hepatitis corpus tend to be very specific, and they gather terms terminologically related :

- liver transplantation, transplantation, orthotopic liver transplantation
- immunoadsorbent, immunosorbent, immunoassay, immunospot, immunosorbent assay

On the other hand, the clusters of terms obtained with the blood corpus generally contain less specific terms:

- day, h, month, hour min
- Banker, den Hollander, Knudsen, Tanner

The clustering of the terms retrieved from the verb-object dependencies provides us with classes on terms sharing a semantic relation of a functional kind. But at this moment, labeling a relation *between* the sets of terms is impossible.

For that reason, in a ***second step*** of this experiment, we used another syntactic structure on the same parsed corpora. This second syntactic structure must carry an other kind of semantic information than the verb-objet dependencies, and must give us the possibility to label the relations between the sets of terms. The prepositional structures answer to both those constraints; they stand for metonymic, part of and other kind of semantic relations, and the prepositions themselves provide a label for the links. The syntactic structure we have used on the corpus had the form "term preposition term". Among the structures retrieved, we have selected the most frequent ones and we have organised them in classes : [term *preposition* set-of-terms]. Each of the clusters obtained previously is compared with each of those sets of terms. When a cluster presents enough similarity with a set of terms, the cluster is augmented with the terms that belong to the set of terms but do not belong to the cluster, and the prepositional information is attached to the new structure. The new structure has the form: [term *preposition* cluster_augmented].

We give here an example of this mechanism in structures taken from the hepatitis corpus:

**prepositional structure:** transmission *of* infection viral_infection disease
**viral_hepatitis cluster:** hepatitis_B_virus viral_infection HCV hepatits_B HCV_infection HBV HBV_infection
**viral_hepatitis resulting structure:** transmission *of* infection disease hepatitis_B_virus viral_infection HCV hepatitis_B HCV_infection HBV HBV_infection viral_hepatitis

The process is iterated as long as two structures can be merged according to the similarity measure. What we get eventually is not a network, but a collection of labeled relations between classes of terms.

For example, the resulting structure mentioned in the example above will evolve and become:

[recurrence transmission] *of* [infection hepatitis_B_virus viral_infection HCV hepatits_B HCV_infection disease HBV HBV_infection viral_hepatitis]

Some resulting structures will involve a very general preposition, like *of*, that does not carry a highly specific semantic information. However, the second example given below shows that the association of the term preceding the preposition and the preposition (here *use of* can produce an interesting link:

1. [dose injection vaccination] *of* [hepatitis_B_vaccine HBV_vaccine vaccine]
2. [use] *of* [face_mask mask glove protective_eyewear]

On the other hand, some prepositions carry a more specific information. *During*, for example, associates a notion of temporal event to the set of terms it precedes:

1. [vaccination vaccine] *against* [disease virus virus_type]
2. [heparin blood_pressure blood blood_loss] *during* [aortic_surgery operation apostosis surgery coronary_angiography hemipathectomy coronary_artery_bypass emergency-surgery cardiac_surgery surgical_resection hemodialysis procedure dialysis transplantation]

In some cases, a set of terms will appear with different *[terms preposition]* structures. It induces a strong link between the terms of this set, as the sematic relations they share allow them to gather in several structures. This happens in particular with the set of terms: [transcriptase transcription transcriptase_activity transcription-polymerase_chain_reaction]

1. [level expression] *of* [transcriptase transcription transcriptase_activity transcription-polymerase_chain_reaction]
2. [effect] *on* [transcriptase transcription transcriptase_activity transcription-polymerase_chain_reaction]
3. [increase] *in* [transcriptase transcription transcriptase_activity transcription-polymerase_chain_reaction]

The next step in this study consisted in finding an efficient evaluation method for the clusters.

## 4.4 Evaluation

As we deal with medical data, we perform an evaluation of the classes and clusters we obtain with UMLS (Unified Medical Language System [24]). The evaluation of extracted clusters is problematic, as we do not have any reference or model for the clusters that we want to build. At the same time, we want this evaluation to be automatic.

We retrieve from UMLS every pair of terms for which:
the two terms share a semantic relation in UMLS,
each of the two terms appear in at least one cluster.

Then, we check how many of those pairs of terms appear together in a cluster. Using this number, we compute a recall and a precision value. It is important to point out that we cannot evaluate exhaustively the content of our clusters, as some of the terms they contain are unknown in UMLS. This evaluation must therefore be considered as a partial evaluation of about 60-70 % of the clusters.

The recall value R is obtained with the number of UMLS pairs found in the clusters and the total number of UMLS pairs:

$$R = \#UMLSpairsintheclusters/\#UMLSpairs \quad (2)$$

To compute the precision value P, we need also the total number of pairs in the set of clusters :

$$P = \#UMLSpairsintheclusters/\#pairs \quad (3)$$

The results of the first evaluation, after the clustering on verb-object dependencies, show low values of recall and precision. This is a consequence of the fact that we use an unsupervised method. Hence, at each step of the process, some mistakes happen, and despite the filtering we are performing, some of those mistakes remain in the final clusters. At the same time, as we perform an automatic

| | Nb of words | Recall | Precision |
|---|---|---|---|
| clustering | 250 | 9% | 11% |
| prep structures | 100 | 33% | 17% |
| | 350 | 18% | 8% |
| | 500 | 17% | 5% |

**Table 1.** *Hepatitis Corpus - Number of words clustered at the end of the process, recall and precision values obtained after the similarity based clustering on classes verb-object extracted from the hepatitis corpus, and after the addition of information provided by the prepositional structures to the clusters*

| | Nb of words | Recall | Precision |
|---|---|---|---|
| clustering | 250 | 9% | 8% |
| prep structures | 150 | 26% | 8% |
| | 400 | 27% | 3% |
| | 900 | 41% | 3% |

**Table 2.** *Blood Corpus - Number of words clustered at the end of the process, recall and precision values obtained after the similarity based clustering on classes verb-object extracted from the blood corpus, and after the addition of information provided by the prepositional structures to the clusters*

evaluation with UMLS, we cannot evaluate all our clusters, and there is a possibility that we miss the evaluation of correct clusters. Considering both initial corpora, we observe better results on the hepatitis corpus (see Table 1), although this corpus is smaller than the blood corpus (see Table 2). But due to its higher specificity, we could collect higher occurrences of structures verb-object in the hepatitis corpus.

The evaluation carried out at the end of this study [15], after the addition of the prepositional information, shows that this new information has improved the recall, but the precision values remain very low, especially when we increase the number of terms clustered. Here again, the hepatitis corpus allows better performances than the blood corpus.

It appears that this unsupervised method, used on a corpus of several million words concerning a very specific subject, allows us to get satisfying results for the clustering of a small set of frequently occurring terms (100-200), if we consider the clustering as a preliminary step in the learning of an ontology.

### 4.5 Related work

A similar approach has been described in [20], where raw text corpora are tokenized and syntactically analysed before the extraction of attributes based on syntactic structures, in order to build automatically a first-draft thesaurus. Related work in the medical area happens in the context of the MuchMore project [33]. However, the UMLS is used as an external knowledge repository to discover additional terms on basis of attested relations between terms appearing in a text. Relations themselves are not the focus of the research. Earlier work on creating medical ontologies from French text corpora has been reported on by [31]. Instead of using shallow parsing techniques, "full parse" trees are decomposed into elementary dependency trees. The aim is to group bags of terms or words according to semantic axes.

[15] Previous work on the clustering methods reported on in this paper as well as a preliminary evaluation have been presented in [37, 36].

Another attempt involving clustering on specific domains, including the medical domain, is described in [6]. Term extraction is performed on a POS-tagged corpus and followed by a clustering operation that gathers terms according to their common components, in order to build a terminology. An expert provides some help in the process, and performs the evaluation.

Unsupervised clustering has been performed also on general domains. In [27], a thesaurus is built by performing clustering according to a similarity measure after having retrieved triples from a parsed corpus. Here, a big corpus (64M words) was used, and only very frequently occurring terms were considered.

## 5 DISCUSSION AND FUTURE WORK

Unsupervised clustering allows us to build semantic classes. The main difficulty lies in the creation of a semantic network as the core, or the basic layer of an ontology, and especially in the systematic labelling of the relations of this semantic network. The ongoing work consists in part in improving the performance of the shallow parser by increasing its lexicon and training it on passive sentences taken from medical corpora, and in part in improving the results of the semantic information extraction methods. With respect to this, we are planning to apply the same experiments to a much bigger corpus, to work on the terminology of the medical domain in order to perform a filtering of this terminology that could lead to an improvement of the quality of the clusters.

In order to perform unsupervised clustering, external help is often required (expert, existing taxonomy...). However, using more data seems to increase the quality of the clusters ([27]). Clustering does not provide you with the exact relations between terms, hence the fact that it is more often used for terminology and thesaurus building than for ontology building. Therefore, we did not convert the resulting structures to candidate DOGMA lexons yet. Once the relations between the concepts become more precise, this conversion step will be done.

Performing an automatic evaluation is another problem, and evaluation frequently implies a manual operation by an expert [6, 14], or by the researchers themselves [21]. In [27], an automatic evaluation is performed including a comparison with existing thesauri like WordNet and Roget. In a future stage, the results of the knowledge discovery process reported here should be given to knowledge engineers to have them determine the usefulness of the results.

## 6 CONCLUSIONS

Although it is still too early for solid conclusions, we feel that the method presented in this paper merits further investigations, especially regarding the discovery of more precise semantic relations. The results seem to indicate that unsupervised techniques would be useful for the discovery of a seed ontology - given sufficient data. We hope that the application of the methods described will ultimately result in the automatic creation of seed DOGMA-lexons that are "precise" enough to be useful for bootstrapping the subsequent ontology learning process by means of supervised learning techniques.

## ACKNOWLEDGEMENTS

This research has been carried out in the context of the OntoBasis project (GBOU 2001 #10069) funded by the Flemish IWT (Institute for the Promotion of Innovation by Science and Technology in Flanders).

## REFERENCES


[1] N. Aussenac-Gilles, B. Biébow, and S. Szulman, eds. *EKAW'00 Workshop on Ontologies and Texts*, volume http://CEUR-WS.org/Vol-51/. CEUR, 2000.

[2] N. Aussenac-Gilles and A. Maedche, eds. *ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*, volume http://www.inria.fr/acacia/OLT2002, 2002.

[3] S. Bechhofer (ed.), 'Ontology language standardisation efforts', OntoWeb Deliverable #D4, UMIST - IMG, Manchester, (2002).

[4] Matthew Berland and Eugene Charniak, 'Finding parts in very large corpora', in *Proceedings ACL-99*, (1999).

[5] T. Berners-Lee, *Weaving the Web*, Harper, 1999.

[6] Didier Bourigault and Christian Jacquemin, 'Term extraction + term clustering: An integrated platform for computer-aided terminology', in *Proceedings EACL-99*, (1999).

[7] C. Brewster, F. Ciravegna, and Y. Wilks, 'User centred ontology learning for knowledge management', in *Natural Language Processing and Information Systems, 6th International Conference on Applications of Natural Language to Information Systems (NLDB 2002) - Revised Papers*, eds., B. Andersson, M. Bergholtz, and P. Johannesson, volume 2553 of *LNCS*, pp. 203 – 207. Springer Verlag, (2002).

[8] Sabine Buchholz, 'Memory-based grammatical relation finding', in *Proceedings of the Joint SIGDAT Conference EMNLP/VLC*, (2002).

[9] Sabine Buchholz, Jorn Veenstra, and Walter Daelemans, 'Cascaded grammatical relation assignment', in *Proceedings of EMNLP/VLC-99*. PrintPartners Ipskamp, (1999).

[10] Sharon A. Caraballo and Eugene Charniak, 'Determining the specificity of nouns from text', in *Proceedings SIGDAT-99*, (1999).

[11] Walter Daelemans, Sabine Buchholz, and Jorn Veenstra, 'Memory-based shallow parsing', in *Proceedings of CoNLL-99*, (1999).

[12] J. De Bo, P. Spyns, and R. Meersman, 'Creating a "DOGMAtic" multilingual ontology infrastructure to support a semantic portal', in *On the Move to Meaningful Internet Systems 2003: OTM 2003 Workshops*, eds., R. Meersman and Z. Tari et al. (eds.), number 2889 in LNCS, pp. 253 – 266. Springer Verlag, (2003).

[13] A. Dingli, F. Ciravegna, David Guthrie, and Yorick Wilks, 'Mining web sites using adaptive information extraction', in *Proceedings of the 10th Conference of the EACL*, (2003).

[14] David Faure and Claire Nédellec, 'Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium', in *Proceedings EKAW-99*, (1999).

[15] P. Gamallo, M. Gonzalez, A. Agustini, G. Lopes, and V. de Lima, 'Mapping syntactic dependencies onto semantic relations', in *ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*, eds., N. Aussenac-Gilles and A. Maedche, volume http://www.inria.fr/acacia/OLT2002, (2002).

[16] Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes, 'Selection restrictions acquisition from corpora', in *Proceedings EPIA-01*. Springer-Verlag, (2001).

[17] Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes, 'Using co-composition for acquiring syntactic and semantic subcategorisation', in *Proceedings of the Workshop SIGLEX-02 (ACL-02)*, (2002).

[18] A. Gangemi, R. Navigli, and P. Velardi, 'The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet', in *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA and ODBASE*, eds., R. Meersman, Z. Tari, and D. Schmidt et al. (eds.), number 2888 in LNCS, pp. 820 – 838, Berlin Heidelberg, (2003). Springer Verlag.

[19] A. Gómez-Pérez and D. Manzano-Macho (eds.), 'A survey of ontology learning methods and techniques', OntoWeb Deliverable #D1.5, Universidad Politécnica de Madrid, (2003).

[20] Gregory Grefenstette, *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers, 1994.

[21] Ralph Grishman and John Sterling, 'Generalizing automatically generated selectional patterns', in *Proceedings of COLING-94*, (1994).

[22] N. Guarino and P. Giaretta, 'Ontologies and knowledge bases: Towards a terminological clarification', in *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, ed., N. Mars, pp. 25 – 32, Amsterdam, (1995). IOS Press.

[23] N. Guarino and C. Welty, 'Evaluating ontological decisions with ontoclean', *Communications of the ACM*, **45**(2), 61 – 65, (2002).

[24] B. Humphreys and D. Lindberg, 'The unified medical language system project: : a distributed experiment in improving access to biomedical information', in *Proceedings of the 7th World Congress on Medical Informatics (MEDINFO92)*, ed., K.C. Lun, pp. 1496–1500, (1992).

[25] H. Karanikas, M. Spiliopolou, and B. Theodoulidis, 'Parmenides system architecture and technical specification', Parmenides Deliverable #D22, UMIST, Manchester, (2003).

[26] H. Karanikas and B. Theodoulidis, 'Knowledge discovery in text and text mining software', Technical report, UMIST - CRIM, Manchester, (2002).

[27] Dekang Lin, 'Automatic retrieval and clustering of similar words', in *Proceedings of COLING-ACL-98*, (1998).

[28] A. Maedche, S. Staab, C. Nédellec, and E. Hovy, eds. *IJCAI'01 Workshop on Ontology Learning*, volume http://CEUR-WS.org/Vol-38/. CEUR, 2001.

[29] Alexander Maedche and Steffen Staab, 'Ontology learning for the semantic web', *IEEE Intelligent Systems*, **16**, (2001).

[30] R. Navigli, P. Velardi, and A. Gangemi, 'Ontology learning and its application to automated terminology translation', *IEEE Intelligent Systems*, **18**(1), 22 – 31, (2002).

[31] A. Nazarenko, P. Zweigenbaum, J. Bouaud, and B. Habert, 'Corpus-based identification and refinement of semantic classes', in *Proceeding of the AMIA Annual Fall Symposium - JAMIA Supplement*, ed., R. Masys, pp. 585–589. AMIA, (1997).

[32] P. Parpola, 'Managing terminology using statistical analyses, ontologies and a graphical ka tool', in *EKAW'00 Workshop on Ontologies and Texts*, eds., N. Aussenac-Gilles, B. Biébow, and S. Szulman, volume http://CEUR-WS.org/Vol-51/. CEUR, (2000).

[33] S. Peeters and S. Kaufner, 'State of the art in crosslingual information access for medical information', Technical report, CSLI, (2001).

[34] H. Pinto, A. Gómez-Pérez, and J.P. Martins, 'Some issues on ontology integration', in *Proceedings of the IJCAI'99 Workshop on Ontology and Problem-solving methods: lesson learned and future trends*, eds., R. Benjamins and A. Gómez-Pérez, pp. 7.1–7.11. CEUR, (1999).

[35] James Pustejovsky, *The Generative Lexicon*, MIT Press, 1995.

[36] M.-L. Reinberger, P. Spyns, W. Daelemans, and R. Meersman, 'Mining for lexons: Applying unsupervised learning methods to create ontology bases', in *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA and ODBASE*, eds., R. Meersman, Z. Tari, and D. Schmidt et al. (eds.), number 2888 in LNCS, pp. 803 – 819, Berlin Heidelberg, (2003). Springer Verlag.

[37] Marie-Laure Reinberger and Walter Daelemans, 'Is shallow parsing useful for the unsupervised learning of semantic clusters?', in *Proceedings CICLing03*. Springer-Verlag, (2003).

[38] Marie-Laure Reinberger, Bart Decadt, and Walter Daelemans. On the relevance of performing shallow parsing before clustering. Computational Linguistics in the Netherlands 2002 (CLIN02), Groningen, The Netherlands, 2002.

[39] F. Rinaldi, K. Kaljurand, J. Dowdall, and M. Hess, 'Breaking the deadlock', in *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA and ODBASE*, eds., R. Meersman, Z. Tari, and D. Schmidt et al. (eds.), number 2888 in LNCS, pp. 876 – 888, Berlin Heidelberg, (2003). Springer Verlag.

[40] S. Staab, A. Maedche, C. Nédellec, and P. Wiemer-Hastings, eds. *Proceedings of the Workshop on Ontology Learning*, volume http://CEUR-WS.org/Vol-31/. CEUR, 2000.

[41] M. Ushold, 'Where are the semantics in the semantic web?', *AI Magazine*, **24**(3), 25 – 36, (2003).

[42] R. Volz, S. Handschuh, S. Staab, L. Stojanovic, and N. Stojanovic, 'Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the semantic web', *Web Semantics: Science, Services and Agents on the World Wide Web*, **1**, 187 – 206, (2004).