Modelling syntactic context in automatic term extraction

Roberto Basili, Maria Teresa Pazienza, Fabio Massimo Zanzotto
Dipartimento di Informatica, Sistemi e Produzione,
Universita' di Roma Tor Vergata (ITALY)
{basili,pazienza,zanzotto}@info.uniroma2.it

Abstract

Terms are key components of terminological knowledge bases (TKBs). These are valuable but very expensive resources for a wide range of applications devoted to knowledge access and management. A large number of approaches to corpus-driven term knowledge base acquisition and extension have been proposed. Syntactic constraints are widely used to characterize terms in source texts. However, contextual information is generally exploited for semi-automatic acquisition of semantic relations among terms and a rich notion of semantic context is usually adopted.
The aim of our research is to investigate whether syntactic context (i.e. structural information on local term contexts) can be used for determining "termhood" of given term candidates. A weakly supervised model is here proposed where predictive rules are built over the grammatical representation of the contexts available from limited terminological resources. Extensive experimental evidence derived from the analysis of a large legal corpus and a controlled terminology suggests the viability of this automatic method over unrestricted texts.

1 Introduction

Terminological knowledge bases (TKBs) (Meyer *et al.* 92) are valuable resources for a wide range of applications devoted to knowledge access and management. These are repositories in which the domain concepts (i.e. terms) are connected via specific relations (e.g. hyperonymy, meronymy). Since their manual development is very expensive, several models for corpus-based acquisition of TKBs have been defined (referred as CTKBs in (Condamines & Rebeyrolle 98)). Terms are key components of TKBs, so that acquisition methods able to select terms out from document collections play a critical role. Symbolic, statistical, or hybrid methods have been proposed to address this issue (see (Jacquemin 97) for a survey).

Both manual and automatic terminology extraction processes rely on a **domain model** that consists of two sources of information on the specific domain (as also pointed out in (Soininem *et al.* 99)):

- a *domain corpus* as source of evidence for new candidate terms
- an *available terminological knowledge base* as representation of the previous selected knowledge in the field.

In the hand-coding of terminological resources the domain model is heavily used. Implicitly, the model represents an extensional definition of "*termhood*" in the domain. Syntagmatic constraints are in fact embodied by the different available terms, as systematic constituency rules for terminological expressions (*constraining patterns*). Terms like *fluid mechanics*, *conservation of energy*, *bread-and-butter equation* suggest constraining patterns like *Noun Noun*, *Noun Preposition Noun*, as typical grammatical structures. These patterns can be matched in the corpus when new candidate terms are searched. Furthermore, grammatical information related to observable use contexts of existing terminology (*contextual information*) is also available from the corpus evidence: here external syntagmatic constraints (i.e. syntactic relations between terms and other contextual material, like verbs) can be derived as general rules of usage for terms in the domain. For instance, the following fragments taken from scientific prose:

- *The bread-and-butter equation of fluid mechanics governs the conservation of energy of everything from flows to jets and turbulence.*
- *The generalized airfoil equation governs the pressure across an airfoil oscillating in a wind tunnel.*

show that both *bread-and-butter equation* and *generalized airfoil equation* appear in similar grammatical structures: the same relations are

established with a given verb *govern*. In this case, lexical and grammatical information provide contextual hints useful to characterize "*termhood*" of the target candidates.

The above information is helpful for deriving an operational *term* definition, i.e. a domain-specific intentional notion, based on *endogenous* and *exogenous* constraints. Constraining patterns are *endogenous* as they refer to restrictions over structure of the term itself. The evidence suggested by the contextual information is instead *exogenous* as it refers to restrictions (on the term) depending on its (typical) contexts. Although term behaviour in the target domain can be characterized at different levels (e.g. syntactically, semantically, distributionally), any such modelling can operate independently over endogenous as well exogenous information.

Several proposed approaches to term selection emphasize *endogenous* properties by almost neglecting *exogenous* information. The role of *exogenous* (i.e. contextual) information has been more frequently exploited in the derivation of semantic relations among terms (Condamines & Rebeyrolle 98; Davidson *et al.* 98; Morin 99; Maynard & Ananiadou 00b). However, the task of positioning an item in a semantics network is rather different from the assessment of its termhood. This latter activity deals with the relevance of a given concept in the target domain. A stronger attention to contextual information for term extraction is paid in (Frantzi & Ananiadou 97; Maynard & Ananiadou 00a): window-based contexts are used to determine termhood of candidates and position them within an *is-a* network. This work makes use of extensive and well-grounded semantic networks: the portability to "poorer" domains/languages is then limited.

The aim of our research is to define a weakly supervised "termhood" model suitably combining *endogenous* and *exogenous* syntactic information, as syntactic approaches have been suggested to outperform window-based models in the slightly different task of defining word similarity (Grefenstette 93). The available resources (i.e. the domain model) are used to study the *exogenous* behaviour of positive examples. Firstly a suitable set of selective features is derived (Section 2). Then a similarity metrics is imposed in the space of grammatical features aiming to rank lists of candidate terms (Section 2.2). The *exogenous* information made available in this way is expected to improve the termhood evidence over traditionally adopted distributional models (as in (Daille 94)). The model has been tested on the Italian language in the legal domain where limited semantics resources are available. Results derived on the above-mentioned domain and on the available controlled terminology will be discussed in Section 3.

2 Making use of exogenous syntactic information

As the task of terminology extraction makes extensive use of knowledge and resources, an automatic procedure is expected to rely on all the available information. A model of *termhood* should thus exploit *endogenous* and *exogenous* grammatical information. In particular, representing the *exogenous* syntactic behaviour requires at least the following questions to be properly answered: (a) which units of information are to be represented? (b) which properties characterize them? (c) what is the role of the pre-existing terminology in determining the target representation? (d) how to use the represented items to support new term detection? or, in other words, which predictive function can be built upon these units and properties? (e) and, finally, how the corpus enters this process?

The above issues that will be adressed in the following sections are basic principles characterizing a classification problem: information units are categorial representations, like feature vectors in machine learning methods; the source terminology provide positive examples to seed the categorial representations while the corpus is the source of training information (and testing as well). As a result, *termhood* is modelled as a class membership function embodying a classification inference.

2.1 Representing term prototypes in a syntactic feature space

Since feature vectors are to be used as formalism, the target of our representation problem is the definition of a suitable space where accepted terms and new candidates lie. For this reason, these objects will be hereafter referred as *term samples*.

Points in the space should represent the

available exogenous information as observable in the corpus. Exogenous syntactic relations are always referred to the *heads* of terms. Moreover, as (Kister 93) points out, singleton terms (i.e. one-word terms) are often elliptic occurrences of complex terms. For example, in the following article (extracted from the Italian Civil Code):

Art. 2781 *Concorso di* ***privilegi speciali*** *con crediti pignoratizi*

Qualora con crediti assistiti da ***privilegio speciale*** concorra un credito garantito con pegno (2784 e seguenti) e uno dei ***privilegi*** debba essere preferito rispetto al pegno, tale ***privilegio*** prevale su quegli altri che devono essere posposti al pegno, anche se anteriori di grado (att. 234).[1]

the second and the third occurrence of *privilegio/privilegi* are elliptic occurrences of the term *privilegio speciale* (*special privilege*). As previously underlined, the syntactic head of the term is used as a referent for the whole structure. These occurrences can thus be all used to increase available evidence about the syntactic properties of the term *privilegio speciale*.

Finally, the term sense is usually determined by its head. In a given domain several terms may share the same head. In the controlled terminology associated to the Italian Civil Code *amministratore unico*, *amministratore delegato*, *amministratore della societá*, *amministratore giudiziario* are all terms headed by *amministratore* (*administrator*). Since the semantics is determined by the same head, their syntactic relations in the corpus are expected to be similar. The representation of heads instead of whole terms has thus the obvious benefit of increasing the observable evidence, to capture anaphoric phenomena, without any loss of "semantic" information. Accordingly, points in the target syntactic space, i.e. the *term samples*, will be the *heads* of available terminology (training instances) and, similarly, the *heads* of new candidates (test instances).

[1]a rough translation: **Art. 2781** *Competition of special privilege supported by deposit.* In the case of credit supported by deposit in addition to credits supported by special privilege: if one of the privileges is preferred to the deposit, the latter privilege takes priority over the credits supported by deposit, even if the credit supported by deposit have a higher grade.

2.2 Syntactic features and the discriminating function

The position of a term sample in the syntactic feature space must be determined by its observed syntactic behaviour. Features, retained as independent properties, should express individual syntactic relations between the term and some word in the surrounding text. Each feature is thus made of a couple (T, h) where T is the relation type (e.g. Subj), and h is the related word (i.e. the head of the syntactic relation). Axes of the space are thus (T, h) couples. For example, a potential feature for *administrator* in *The head administrator is ...* is $F = (V\_Subj, to\_be)$. These properties capture patterns of use on a lexical and grammatical basis. Details on the detection of terms and their relations in the corpus are discussed in Section 2.3.

The value of each feature is its observed frequency. This (cumulative) value is thus derived from all the occurrences of simple heads throughout the available texts. More precisely, a term sample t is represented as a vector $\tau(t)$ in R^{+n} given by:

$$\tau(t) = (f_1, f_2, ..., f_n) \tag{1}$$

where f_i is the value related to the attribute $F_i = (T_i, h_i)$ and n is the total number of features.

In the above space, a variety of metrics can be established, all of them being equivalent for ranking purposes. In fact, the focus is not on the kind of metrics adopted but on the analysis of the feature space proposed. The similarity measure adopted in our experiments (i.e. a cosine measure used for instance also in (Salton 91) for estimating the document distance) is:

$$sim(\tau_i, \tau_j) = \frac{\tau_i \cdot \tau_j}{|\tau_i||\tau_j|} \tag{2}$$

where $\tau_i \cdot \tau_j$ is the scalar product and $||$ is the norm.

The similarity measure defined in the syntactic feature space can be used to rank lists of candidate term heads. The role of the controlled terminology T is to provide evidence of the behaviour of correct examples: some term samples are thus available as positive instances of the *termhood* property for training. No explicit information is available on negative examples.

Therefore, only the distance between feature vectors of candidates and vectors representing

true terms can be used as ranking function. This function defines naturally a synthetic representation of the termhood property. Now, as independent vectors for different heads are prone to sparse data, a unique term sample is derived by cumulating evidence from all the positive examples. A *centroid* $\tau(T)$ is obtained able to represent the *termhood* portion of the syntactic space. It is determined by summing up vectors representing the controlled (training) terms, as follows:

$$\tau(T) = \sum_{t \in T} \tau(t) \tag{3}$$

Accordingly, the *exogenous weighting* score $exw(t)$ of a generic term sample t is obtained by imposing the similarity measure, i.e.:

$$exw(t) = sim(\tau(t), \tau(T)) \tag{4}$$

The $exw(t)$ alone can be used for ranking, as opposed to the frequency $f(t)$. However, the *exogenous* weighting factor is strongly biased by the frequency: more the analysed term t is frequent, more the syntactic information gathered is valuable. Furthermore, even if the ranking proposed by the *exogenous* information is precious, it does not outperform over the term frequency (as discussed in sec. 3).

The benefits of the exogenous information are better exploitable if used in combination with the term frequency. Two combined weighting functions are explored. Firstly, a score $exf(t)$ is derived by multiplication, as follows:

$$exf(t) = exw(t)f(t) \tag{5}$$

where $exw(t)$ is used as bias of the frequency $f(t)$. Alternatively, the frequency $f(t)$ can be firstly imposed and than ranking on the survival candidates can be driven by $exw(t)$. More precisely, a cascade score $cw(t)$ can be defined for a candidate t as:

$$cw(t) = \begin{cases} 0 & \text{if } f(t) < \alpha \\ exw(t) & \text{otherwise} \end{cases} \tag{6}$$

First threshold is imposed on frequency (e.g. the first M terms are selected), and then the exogenous syntactic information, the $exw(t)$, is used as re-ranking score.

In the experimental section (see sec. 3), both approaches will be investigated: the ranking obtained by the $exf(t)$ will be compared with the one obtained using only the frequency information $f(t)$; and, the re-ranking made on the first M more frequent terms will be contrastively analysed with respect to the previous ranking.

2.3 Extracting Endogenous and Exogenous information

Both the selection of candidate terms and the extraction of their contextual syntactic information must rely on robust methods for corpus processing. Since the proposed method to assess termhood should be applicable for different domains, the corpus processors should be shallowly related to the domain/sub-domain or, at least, (semi)automatically tuneable.

Term candidate recognition is the matching of complex surface structures able to represent terminological concepts. Endogenous properties derived from the pre-existing terminology are helpful to select among different kind of linguistic expressions. The candidate term extractor is based on shallow parsing techniques. It makes use of a cascade of modules: (a) a yellow page look-up phase and a named entities matcher able to detct basic and complex named entities; (b) a morphologic analyser that determines (possibly ambiguous) syntactic categories and morphological interpretations for each word; (c) a rule-based part-of-speech tagger; (d) a syntactic parser based on modularisation and lexicalisation (Basili *et al.* 00) that, after chunking recognizes syntactic relations among chunks.

The last step is a form of lexicalised recognition, where sentence structure (i.e., clause(s) determined by the underlying verb argument structures) provides important constraints on the relation matching mechanisms. It relies on a lexicon of verb subcategorization frames (details are in (Basili *et al.* 00)).

The output of the syntactic processor is an extended dependency graph (XDGs) whose nodes are chunks and edges are syntactic dependencies among chunks, called *inter chunk dependencies*, (*icd*s). For each sentence s of the input text, the analyser produces an

$$xgd(s) = (C, L) \tag{7}$$

where C is the set of constituents (i.e. chunks) and L is the set of the *icd*s. Each *icd* in L is associated to a plausibility score ranging in

the $[0, 1]$ interval. It represents the persistent ambiguity: value 1 characterizes unambiguous links. For instance, the extracted chunks for the sentence title of the article *Art. 2781* is:

[1/C_Nom *Concorso*] [2/C_Prep *di privilegi*] [3/C_Adj *speciali*] [4/C_Prep *con crediti*] [5/C_Adj *pignoratizi*]

while the inter chunk dependencies are depicted in Fig. 1.

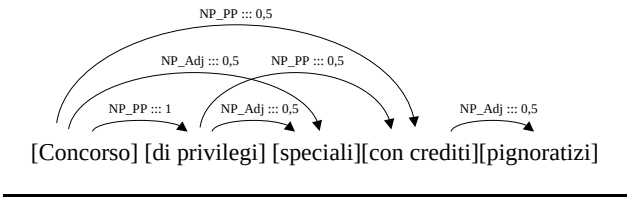


Figure 1: Sample sentence

The selection of candidate terms builds upon the grammatical information represented in extended dependency graphs. The candidates are chosen among all the *partial phrases* that satisfy imposed grammar constraints (Basili *et al.* 01). For each matched candidate term its head is determined. These heads will be then modelled as *term samples* in the syntactic feature space presented in Section 2.1.

Given the analysis of the syntactic parser, the extraction of the exogenous syntactic behaviour of a candidate term comes straightforward: as described in the following, local (i.e. sentence-level) information is collected to capture the term global exogenous behaviour (i.e. the vector $\tau(t)$ of def. 1). The ambiguity handling is worth a specific attention. In this case, the parser preserves all the ambiguities that cannot be resolved within a redundant dependency graph. Ambiguity is modelled by plausibility scores assigned to syntactic relations.

The local syntactic behaviour of the term, at the sentence level, is represented in the related *xdg*. We are here interested in the relations in which a given term head t is captured as modifier since the syntactic relations in which it plays the role of head are endogenous. Therefore, the local syntactic behaviour is described by the subset L_t of the set of relations. Given, then, $L_t = \{(H_1, t, T_1, Plaus_1), ..., (H_n, t, T_n, Plaus_n)\}$ the local feature vector $\tau_l(t)$ is obtained putting the $Plaus_i$ value in the related dimension $F = (T_i, H_i)$. Therefore, the global feature vector $\tau(t)$ is obtained summing up all the contributions of the single samples $\tau_l(t)$, i.e.:

$$\tau(t) = \sum_{\tau_l(t) \in C(t)} \tau_l(t) \qquad (8)$$

where $C(t)$ are all the samples collected in the corpus C for the target term head t.

3 Experimental Evaluation

Extensive experiments have been carried out in order to investigate the effectiveness of the *exogenous* term properties encoded in the syntactic feature space. The ranking proposed by the three weighting functions based on the *exogenous* information is analysed against the ranking produced by the frequency. The more promising results are obtained using the cascade score $cw(t)$ where the *exogenous* properties are used after the threshold imposed on the frequency (see fig. 4).

The test-bed consists of the Italian Civil Code, playing the role of the corpus C, and a related terminology T, playing the role of both the pre-existing and the testing terminology. Sizes of the two variable are in tab. 1.

	size
Italian Civil Code	250,000 words
Terminology	1,000 terms

Table 1: Test-bed sizes

Before discussing the results (sec. 3.2), we will introduce hereafter to the testing methodology (see sec. 3.1).

3.1 Methodologies and Measures

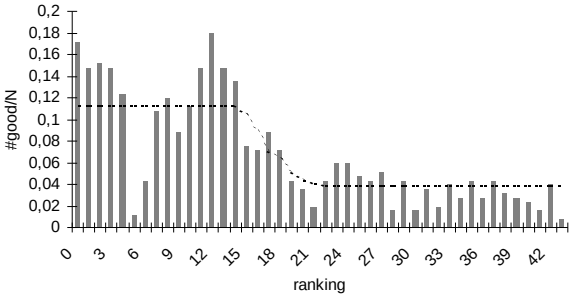
Since the methodology is based on pre-existing knowledge (i.e. the pre-existing terminology), a n-fold cross validation is required to state the correctness of the obtained results. The terminology T then will play the two roles: pre-existing terminology (the one used by terminologists in their work) and testing terminology. These two sublists are obtained by selecting randomly terms from T; more precisely, the terminology T is randomly divided in n portions and the evaluation is carried out in n steps. For each step, a portion plays the role of testing material while the remaining $n - 1$ are used as pre-existing terminology.

On the other hand, the plots have to show whether or not the proposed weighting factor is able to select terms. As the weighting factor is used to rank candidate term heads, a method to investigate the distribution may be the histogram: the ranked list is divided in bins of homogeneous size N and, for each bin, the count of good observations that fall into is determined. The counts are normalized by dividing by the total number of observation N, i.e. the ratio $r = \#good/N$ is represented (where $\#good$ is the number of the elements that have to be detected by the use of the selecting function). Since this representation shows where the good elements are in the list, the ideal selecting function should by a step function: good elements are ranked in the first positions.

It is worth noticing that, since a n-fold cross validation is used, each point in the plots is the mean of n different measurements.

3.2 Results

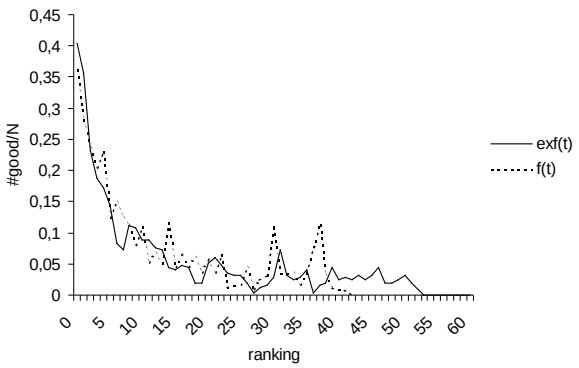
The effect of the *exogenous* properties (the score $exw(t)$) in ranking candidate terms is shown in the plotting of Fig. 2 where is also reported the bin size N and the number of folds $\#folds$ used in the cross validation (this latter information is reported on all the figures). The plotting is very similar to a step function (i.e. the dashed line) even if it doesn't reach the zero level. However, this ranking factor alone is not sufficient to outperform with respect to the term frequency (cf. Fig 2 with the dashed line in Fig 3).



N=50
#folds=5

Figure 2: The *exogenous* properties as ranking factor

The we can use the *exogenous* properties factor to improve the results obtained by the term frequency (the score $exf(t)$). In Fig. 3, the ranking driven by the $exf(t)$ is compared with the one based on the pure $f(t)$: the benefits are visible for the first seven bins. After that, the *exogenous* information becomes too sparse and it is not able to contrast frequency. However, the curve based on the *exogenous* information is more monotonic with respect to the frequency ranking. This suggests that the *exogenous* information helps in capturing the termhood (even if this alone is not sufficient).



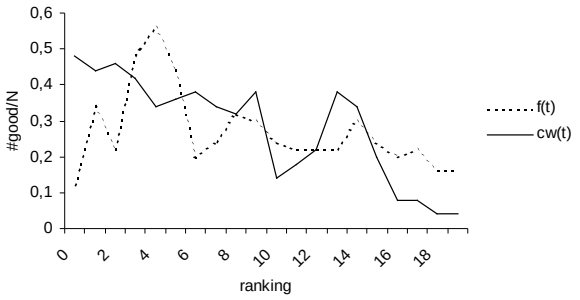
N=50
#folds=5

Figure 3: *Exogenous* properties combined with frequency with respect to simple frequency

Looking closer to the obtained lists and comparing the ranking based on the *exogenous* information with respect to the ranking based only on the term frequency, we discover some interesting issues. For instance:

- the typical temporal expression lose positions: both *giorno* (day) and *anno* (year) lose 10 positions.
- some proposed head of good terms gains positions and reach the top 50; these are *privilegio* (privilege), *separazione* (separation), and *donazione* (donation).

An more interesting plot is the one that describes how the ranking based on the syntactic information perform on the M more frequent terms, i.e. the $cw(t)$ score. The plot is depicted in Fig. 4 where the first 200 more frequent terms have been re-ranked using the endogenous weighting factor $exw(t)$. The ranking obtained by this latter is contrastively compared with the pre-existing sorting based on the frequency $f(t)$ (the dashed line). In the plotting $N = 10$ has been taken. What



N=10
#folds=5

Figure 4: The effect on the 200 more frequent term sample

emerges is that the measure based on the syntactic feature space enables correct terms to be better ranked. In fact, good term samples are promoted to higher positions. This results from: the better performance that this method has for the first three sections and the drastically reduced value reckoned for the last four. It is worth noticing that in the re-ranking the coverage is not augmented.

4 Conclusion

Our approach shows that the role of *exogenous* properties of terms is important in the identification of the termhood, especially when combined with the term frequency (Daille 94). The approach differs from what is proposed in (Maynard & Ananiadou 00a) since it uses more structured contextual information. However, it does not makes use of semantics information nor on semantics network. In fact, it is based on term lists typical results of a terminology extraction process. Thus, it can be proposed as a part of the iterative process since it makes use of the available produced material (i.e. the term lists). As the terms are selected from the domain corpus, they can be used to refine the *exogenous* model of the termhood. Lastly, even if the proposed method makes use of domain information in the parsing process (i.e. verb subcategorization lexicons), it exploits unsupervisely learnt material (as described in (Basili *et al.* 99)), therefore it is portable throughout the domains.

References

(Basili *et al.* 99) Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. Lexicalizing a shallow parser. In *Proc. of the Traitement Automatique de la Langue Naturelle, TALN99*, Cargese, France, 1999.

(Basili *et al.* 00) Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. Customizable modular lexicalized parsing. In *Proc. of the 6th International Workshop on Parsing Technology, IWPT2000*, Trento, Italy, 2000.

(Basili *et al.* 01) Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. A contrastive approach to term extraction. In *Proc. of the 4th Conference on Terminology and Artificial Intelligence, TIA2001*, Nancy, France, 2001.

(Condamines & Rebeyrolle 98) Anne Condamines and Josette Rebeyrolle. Ctkb: A corpub-based approach to terminological knowledge base. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors, *Proceedings of the First Workshop on Computational Terminology COMPUTERM'98,held jointly with COLING-ACL'98*, Montreal, Quebec, Canada, 1998.

(Daille 94) Beatrice Daille. *Approche mixte pour l'extraction de terminologie: statistque lexicale et filtres linguistiques.* Unpublished PhD thesis, C2V, TALANA, Universitè Paris VII, 1994.

(Davidson *et al.* 98) L. Davidson, J. Kavanagh, K. Mackintosh, I. Meyer, and D. Skuce. Semi-automatic extraction of knowledge-rich contexts from corpora. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors, *Proceedings of the First Workshop on Computational Terminology COMPUTERM'98, held jointly with COLING-ACL'98*, Montreal, Quebec, Canada, 1998.

(Frantzi & Ananiadou 97) K.T. Frantzi and Sophia Ananiadou. Automatic term recognition using contextual cues. In *Procedings of 3rd DELOS workshop*, Zurich, Switzerland, 1997.

(Grefenstette 93) Gregory Grefenstette. Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches. In *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*, Columus, OH, USA, 1993.

(Jacquemin 97) Christian Jacquemin. *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'Habilitation Diriger des Recherches en informatique fondamentale.* Université de Nantes, Nantes, France, 1997.

(Kister 93) Laurence Kister. *Groupes nominaux complexes et anaphores: possibilité de reprise pronominale dans "N1 de (dét.) N2 ".* Unpublished PhD thesis, Sciences du Language, Univesité de Nancy, 1993.

(Maynard & Ananiadou 00a) Diana Maynard and Sophia Ananiadou. Term extraction using a similarity-based approach. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors, *Recent Advances in Computational Terminology*, 2000.

(Maynard & Ananiadou 00b) Diana Maynard and Sophia Ananiadou. Terminological acquaintance: the importance of contextual information in terminology. In *Proc. of NLP2000 Workshop on Computational Terminology for Medical and Biological Applications*, Patras, Greece, 2000.

(Meyer *et al.* 92) Ingrid Meyer, Douglas Skuce, Lynne Bowker, and Karen Eck. Towards a new generation of terminological resources: an experiment in building a terminological knowledge base. In *Proceedings of COLING-92*, Nantes, France, 1992.

(Morin 99) Emmanuel Morin. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques.* Unpublished PhD thesis, Univesité de Nantes, Faculté des Sciences et de Techniques, 1999.

(Salton 91) G. Salton. Development in automatic text retrieval. *Science*, 253:974–980, 1991.

(Soininem *et al.* 99) Pirjo Soininem, Atro Voutilainen, and Pasi Tapanainen. An experiment in automatic term extraction. In Peter Sandrini, editor, *TKE 99: Terminology and Knowledge Engineering*, Vienna, Austria, 1999. TermNet.