How to evaluate necessary cooperative systems of terminology building?

Thierry Hamon[*], **Olivier Hû**[*†]

[*]LIPN – UMR CNRS 7030
Université Paris-Nord
Avenue J.B. Clément, 93430 Villetaneuse, FRANCE
{*firstname.lastname*}@lipn.univ-paris13.fr

[†]UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - 60206 Compiegne Cedex
olivier.hu@utc.fr

Abstract

Terminology building cannot be considered as a full automated process but rather as a cooperative task between terminological tools and terminologists. Identifying terms in a technical domain is a matter of word usage and expert agreement. We point out the problem of the evaluation of such tools: their quality and their contribution to the terminology building is difficult to estimate and cannot be fully evaluated with usual precision and recall measures. We aim at evaluating more globally their technical aspects and their usability. We give a non-exhaustive list of the features of such evaluation. Then, we apply them on four terminological systems.

1. Introduction

Applications based on technical textual data (translation, information retrieval, controlled indexing, document consulting and navigation, technical authoring) require continuously specialized knowledge, e.g. terms of the domain, relations between terms, semantic classes. Traditionally, specialized vocabulary is gathered in terminological resources: thesauri, specialized dictionaries, etc.

As particular phenomena pointed out thanks to the increasing volume of data, terminological resources have to be built according to the specific needs of each application and working corpora (Bourigault and Slodzian, 1999). Moreover, limits of the reuse of existing terminologies have been pointed out (Abbas and Picard, 1999).

Terminology building process cannot be fully automated. The identification of the terms of a technical domain is a matter of word usage and expert agreement. The detection of term relations requires a thorough knowledge of the underlying concepts and an user's control.

There also exist various types of terminologies (Srinivasan, 1992). It is well-known that various experts give different results depending on their various backgrounds (Szpakowicz et al., 1996). From this point of view, terminology building is a cooperative and interactive task between the system which proposes terms or relations, and the user which accepts, rejects or modifies them according to the working corpora and the application.

In that respect, evaluation of terminological systems is problematic. Quality of the systems cannot be simply evaluated with precision and recall measures. Software engineering aspects (technical aspects and functionalities) and usability of the interface must be considered as well. They have to be adapted to the computational terminology. Therefore, it is necessary to propose a wider view of the tools including not only the quality of the results but also the performances and the features of the system, as the ergonomic quality of the interface.

We aim at defining a global evaluation of the terminology building systems. We propose evaluate these systems according to the both following axes:

- Terminological aspects of systems.
- Ergonomic features of the user interface and the interactions.

Terminology building process and terminological tools used to illustrate this evaluation are described in the section 2.. After a brief review of the methods and problems of evaluating terminological systems (section 3.), we define the criteria for the evaluation of terminology building systems (section 4.).

2. Terminology building: a two step process

Terminological tools have been developed to help the terminologists to identify, classify and organize terms according to the relations between them. We assume that the terminology building is a two step process :

1. **Extraction step** highlights the noun phrases that could be used as terms by the experts of the domain. Term extraction tools provide a list of candidate terms which can be represented as a syntactic network.

   Several approaches allow identifying noun phrases : surface grammatical analysis using words which delimit relevant noun phrases (Bourigault, 1992; Heid et al., 1996), linguistic filtering using syntactic patterns and statistical measures (Daille, 1995; Frantzi et al., 1997). The former method has been implemented in the terminology extraction software LEXTER (Bourigault, 1992) which extracts candidate terms, *i.e.* nouns, adjectives and noun phrases, while the later is applied in ACABIT (Daille, 1995) which associates several statistical measures with each term : frequency, loglike ratio, Shannon diversity, distance between term elements.

2. **Structuration step** aims at building a terminological network out of this list of terms. Semantic relations between terms and term classes enrich the syntactic network. The resulting terminological network, terms as well as relations, must be controlled by experts.

   Terminology structuration approaches are multiples. The identification of relations is generally based on boundary words and lexico-syntactic patterns (Hearst, 1992; Morin, 1998; Garcia, 1998) or inference rules triggered by lexical informations (Jacquemin, 1996; Hamon and Nazarenko, 2001). The evaluation features we propose are illustrated on two systems using this later approach: a unification-based partial parser FASTER (Jacquemin, 1996) which analyses raw technical texts while meta-rules detect morpho-syntactic variants of controlled terms or extracted candidate terms, and a rule-based system, SynoTerm (Hamon and Nazarenko, 2001), which infers synonymy relations between complex terms by using semantic informations extracted from various types of resource (general dictionary, thesaurus).

   Approaches on term classification are generally based on distributional analysis. Terms gathering uses statistical measures of their contexts (Assadi, 1997; Myaeng and Li, 1992), or symbolic clustering (Habert et al., 1996).

This architecture is generally adopted for building monolingual and bilingual terminological resources (Abbas and Picard, 1999; Dagan and Church, 1994; Davidson et al., 1998).

3. How to evaluate terminological tools?

Terminological tools are designed to help terminologists in the terminology building task. Automatically built data must be then validated, modified and completed. Various parameters are taken into account during the terminology building: domain, application context (Bourigault and Habert, 1998), terminologist practice. Regarding this, a corpus could lead to build several terminologies. In that respect, the evaluation of terminological tools is problematic.

3.1. Limits of the precision and recall measures

Most of the terminological tools are evaluated as information retrieval systems using the precision and recall measures. However, such evaluation faces several problems. In the context of the semantic tagging, Resnik and Yarowsky (1997) point out that such measure cannot differentiate some types of errors. For instance, reducing ambiguity is considered as similar to wrong semantic tagging. They propose a new evaluation score which increases or decreases error weight according to the ambiguity reduction.

In previous experiments of terminological relation extraction (Hamon and Nazarenko, 2001), we noticed that precision does not reflect specificity of terminological systems, *i.e.* the terminologist aid. Indeed, results are precious for the terminologists, even if the precision is quite low and he needs to interprete them. The evaluation of such system has to focus on the robustness of the approach and the contribution of the results. Similarly to Resnik and Yarowsky (1997), we have proposed a new evaluation measure of terminological systems, based on the precision and minoring frequent errors (Hamon and Nazarenko, 2001).

Moreover, in practice, information relevance depends on the type of validation *i.e.* strict or not. For instance, a relation different from the searched relation type would be rejected if the results are strictly validated, while it would useful for the terminologist. This feature has to be explicited in order to insure the quality of the evaluation and the comparison with other tools.

Exhaustivity of the terminology is important for the terminologist (Gouadec, 1990). In that context, the corpus coverage and the recall are favored in the terminological systems at the expense of the precision. However, evaluating the recall is also problematic. The use of a golden standard is incompatible with our terminology building methodology based on the application and the corpus. This golden standard should be built for the same application and from the same corpus: such data are not commonly available. In this way, the evaluation is difficult. All the information (relations, semantic classes, contexte) proposed by several systems allows the terminologist to select a term or a relation.

Usability is also crucial in the evaluation of terminological systems. Previous experiments point out that numerous errors can be easily and quickly corrected if results are structured (Hamon et al., 1998). In that respect, an evaluation based on the precision cannot take into account the validation cost, usability of the system or more technical parameters as the use of results provided by other terminological systems.

3.2. Various parameters to evaluate

As the precision measure does not reflect all the terminology building process, we consider that the evaluation have to take into account the help for the terminologist and the suitability for the application and the end-user.

Previous works evaluating terminological systems mainly focus on the term extraction step (L'Homme et al., 1996; Bourigault and Habert, 1998; El-Hadi and Jouis, 1998), while the evaluation of systems structuring the terms are very few (Mayfield and Nicholas, 1992; Assadi, 1998) or partial (Morin, 1998).

The recall of a tool acquiring hyperonymy relation based on lexico-syntactical patterns (Morin, 1998) is not fully considered. The evaluation is based on a local recall and precision, and then on the average of these measures. The average precision 79% does not reflect disparity in the patterns whose precision varies between 33% and 100%. It appears that the usefulness of a system cannot be just based on the precision. In our experiments, we conclude similarly: one of the rules has a very low precision while the relations are generally not manually identify.

L'Homme *et al.* (1996) characterize systems of term extraction with various parameters to take into account before and after the extraction step. They consider these pre-extraction criteria and result analysis in evaluation grid.

While El-Hadi and Jouis (1998) suggest to confront re-

sults of the term extraction with a reference terminology, Bourigault and Habert (1998) argue that such evaluation is unrealistic. A golden standard would ignore the application, the terminologist analysis of the results, and the fact that the terminological systems are designed to extract candidate terms and not terms.

From a technical point of view, several types of errors lead the systems to propose unrelevant candidate terms or relations. Some erroneous information can be easily identified. Machine translation systems could be evaluated according to a hierarchy of attributes (King, 1997). The quality of the translation deals with the intelligibility and the accuracy of the translation while the evaluation of the system would consider external criteria as the result understanding and post-processing.

Comparing several systems is also problematic: various formalisms and strategies are used to extract the candidate terms. The competitive evaluation of two term extraction tools is representative of these problems (Bourigault and Habert, 1998). It is necessary to chose a common representation of the candidate terms taking into account the specificity of each tool [1]. Moreover, the problems of sentence and term component analysis lead to only evaluate the maximal candidate terms identified by each system. The evaluation focus on the syntactic analysis of the candidate terms, not on the extraction step.

4. Towards a better evaluation of terminological cooperative systems

To tackle the problem of the terminological system evaluation, we argue that it is necessary to take into account all the features of these tools. We attempt to take advantage of an general method to help the system evaluation (Hû and Trigano, 2000). This evaluation is an adaptative questionnaire providing a hierarchical structure of criteria. We first consider technical and descriptive aspects as requirements and performances, then ergonimic aspects as usability and the interaction quality of the system.

4.1. Evaluating terminological aspects

Each terminological system has some specific but uncomparable features. According to the various parameters described at the section 3.2., the evaluation should be global but also adapted to the system. We aim at proposing an approach of the evaluation of these systems taking into account their specificities and the fact they are designed to assit the terminologists. To illustrate the evaluation, we consider four terminological tools used to build structured terminology : LEXTER, FASTER, SynoTerm, ACABIT. The table 7 will summarize the criteria applied on these tools.

We first extend the term extraction system evaluation proposed by (L'Homme et al., 1996). We also take advantage of the general criteria designed for NLP-systems (Spark Jones and Galliers, 1996). The set of specific criteria described below are classified according six meta-criteria : system purpose, strategy, parameters, functionalities, quality of the results, and exchange formats. A hierarchical structure of the meta-criteria and criteria allows to select only relevant ones according to the analyzed system.

4.1.1. Tool purpose

This set of criteria aims at describing for each system, pre-evaluation features defined in (L'Homme et al., 1996), in order to select the way of evaluation. We distinguish, on one hand, term extractors where the evaluation will focus on the type of identified elements (noun phrases, verb phrases, collocations, etc.). On the other hand, the systems for structuring terminology will be evaluated according to features as type of acquired knowledge: relation and type of relation, term similarity and classification.

Moreover, in this part of evaluation, we are interested in describing more general features as language of the processed documents, maximal corpus size, and required preliminary processing (POS tagging, parsing, etc.). These criteria are summarized in the table 1.

Criteria	Description
Term extraction	Identified units (noun phrases, verb phrases, collocation, etc.)
Term structuration	Acquisition of relations (type of relation), term similarity, classification
Miscellaneous	Language, maximal corpus size, required preliminary processing

Table 1: Tool purpose criteria

The four terminological systems are evaluated according to this meta-criteria. LEXTER and ACABIT evaluation leads to put them in the term extraction criteria. They identify mainly noun phrases. We consider FASTER and SynoTerm as terminology structuration system which acquire, respectively, morpho-syntactic variants and synonymy relations between terms. FASTER could be also considered as term extractor. However, in our analysis, we are interested in the morpho-syntactic variants of terms it proposes, candidate terms being firstly identified with LEXTER.

Relating to the miscellaneous criteria, while LEXTER and SynoTerm require French corpora, FASTER and ACABIT can be applied on French or English corpora. Maximal corpus size is fixed by a limit value for LEXTER (300 000 words) and ACABIT, and according to the memory space for SynoTerm. We are not aware of corpus limit for FASTER. Except FASTER, preliminary processing is required for LEXTER, ACABIT, and SynoTerm. The formers need a corpus with part of speech tagging. SynoTerm identifies synonymy relations from a corpus firstly analyzed by a term extractor like LEXTER.

4.1.2. Strategy

This axe analyses processing approach implemented in the system. We distinguish: (1) Linguistic approach based on transformation rules, lexico-syntactical patterns and word boundaries, and (2) statistical approach, which can be endogeneous or based on resources. In the later case, various kind of knowledge sources can be used: general language dictionary, thesaurus, manually or automatically built specialized data. The table 2 summarize these criteria.

[1] A similar problem have been encountered for the POS tagger evaluation, GRACE (Adda et al., 1999)

Criteria	Description
Linguistic approach	Transformation rules, lexico-syntactic patterns, word boundaries
Statistical approach	endogenous, resource-based
Machine learning	
Resources	general language dictionary, thesaurus, specialized resources, automatically built data

Table 2: Data extraction strategy criteria

The analysis of the term extractors shows that they are both based on the both approaches. LEXTER uses linguistic approach to identify noun phrases, and statistical one to disambiguate endogenously preposition attachment. ACABIT carries out a two step term extraction firstly using a linguistic approach based on syntactic patterns, then computing statistical measures of relevance for each candidate term. No lexical resource is required by the both term extractors.

The acquisition of relations with FASTER and SynoTerm are both based on transformation rules. SynoTerm requires lexical resources: general language dictionary, thesaurus, etc. FASTER in its basic use requires a controlled term list.

4.1.3. Parameters

We define several criteria to describe how the system could be tuned according to the application and the terminologists requirements. In the later case, a criterion will focus on the type of knowledge which will be acquired: prior or during to the processing. These criteria are summarized in the table 3.

Criteria	Description
Tuning	application, terminologist requirements
Acquired knowledge definition	prior or during to the processing

Table 3: Parameters criteria

All the terminological systems indirectly consider the application and the terminologist requirements by proposing candidate terms and relations. Acquired knowledge have to be defined before processing.

4.1.4. Functionalities

During the validation of extracted relations (Hamon and Nazarenko, 2001), we notice the significance of some functionalities as taking into account previous terminologist judgments in order to adapt the results and their computed relevance. Moreover, some rejected relations can suggest to the terminologist, modification or enrichment of input knowledge. So, the integration of new input knowledge during validation is an important advantage and functionality.

The transparency of the processing, as computing way, input konwledge, have to be described and available for the validation step in order to understand some frequent errors. While these errors are not weighted similarly to rare errors, their negative effect is therefore reduced (King, 1997). The table 2 summarize the functionality criteria.

Criteria	Description
Terminologist knowledge	correction, addition
Processing transparency	computing way, input knowledge

Table 4: Functionalities criteria

Only information about input knowledge and computing way is provided by FASTER and SynoTerm. All the terminological systems do not propose other functionalities.

4.1.5. Quality of the results

Through these criteria, we aim at describing the relevance of the results and the type of required validation. The former concerns precision and recall but also other evaluation measures such as F-measure or minoring-error precision (Hamon and Nazarenko, 2001). This later feature will describe the context of application and will lead to compare these measures from different viewpoints.

As we notice in the section 3.1., the validation process can be limited by the system purpose (strict validation, accepting or not the results) or integrate supplementary proposed information (large validation, allowing modification of the results).

We argue that the precision does not allow to distinguish various types of errors: ambiguity, wrong stemming, partial or erroneous data provided by other terminological systems. Information about the errors from the system are important to describe its performance and its robustness. These criteria are summarized in the table 5.

Criteria	Description
Type of validation	strict, large
Measure	precision, recall, F-measure, minoring-error precision
Type of errors	ambiguity, wrong stemming, partial or erroneous data

Table 5: Results quality criteria

As the evaluation of the results will be corpus-dependant, we chose here not to provide any measure. However, experimentally, we notice that the four systems are error-robust. It seems that only SynoTerm propose a large validation, allowing modification of the type of the proposed relations.

4.1.6. Exchange formats

Terminological systems extract knowledge useful for several aspects in the terminology building task: term extraction, relation acquisition, term classification, etc. Such data has to be gathered and shared between systems. For instance, SynoTerm requires terms extracted by LEXTEr to propose relations. Moreover, we argue that each tool should take into account results of other tools to increase quality of the processing.

From the process point of view, the disparity between the systems leads to define specific data exchange formats. A criterion of the evaluation of terminological systems should take into account such feature. In that respect, standard exchange formats as those proposed by TEI (Ide and Véronis, 1995) or GENETER (Le Meur, 1998) would be preferred to the specific ones for which the re-use and the consistency are not maintained. The table 6 summarize these criteria.

Criteria	Description
Standard formats	TEI, GENETER
Specific formats	

Table 6: Exchange formats criteria

The four terminological systems we analyze, use specific input and output formats. ACABIT requires Brill tagger output format, while the input format of FASTER and SynoTerm is LEXTER output format.

4.1.7. Discussion

The list of above criteria is not exhaustive. We aim rather at proposing criteria emerging from the study of several terminological systems and from a deeper analysis of the terminology building task and the end-user needs. This list is heterogeneous: some criteria are evaluated according to booleans while others require a numerical or textual value. Moreover, we argue that the weight of each criteria depends on the evaluation and experimental conditions.

4.2. Evaluating usability

The general method to help the system evaluation described in (Hû and Trigano, 2000) defines an adaptative questionnaire which provides a hierarchical structure of criteria about several themes. In the context of the terminological system evaluation, we select three generic themes which can be applied to any software: 1 the usability, evaluating how the interface is usable and the interaction quality (Bastien and Scapin, 2001), 2 the technical quality, which concerns the robustness and technical portability of the system (Meyer, 1997), and 3 the general feeling, evaluating the difference between the quality of the system and the user's satisfaction (Hû et al., 1999). These meta-criteria and the terminological feature evaluation defined at the section 4.1. are gathered in a single hierarchical structure of criteria.

In the following, we mainly focus on the application of the meta-criteria on the terminological systems. We carry out a preliminary study of LEXTER, ACABIT, FASTER, and SynoTerm. Like for the quality of results, evaluating the general feeling requires the definition of a user profile. In that respect, this meta-criterion will not be detailed here. While our observations cannot be considered as, strictly speaking, an evaluation, they show the flaws and the qualities of these systems. We consider the usability of these system which have proved their usefulness, with a critical mind. We aim at making their use easier for a terminologist without any technical abilities and training.

4.2.1. Preamble

Firstly, we point out two problems which appear when analyzing the terminological software :

Limits of laboratory prototypes The four systems are designed for research and experiments, while a larger and commercial use would require a software engineering and usability study. We aim at making the authors aware of basic usability problems. The integration of these aspects will lead to insure a technical and ergonomic quality. However, such aspects need specfic development.

Gap between required and effective end-user abilities The terminological tools we analyze, are designed for the specific needs of the users. However, while the terminology building task is well-known of the terminologists, the installation and the use of such systems require engineering abilities. As for any software, it is crucial to distinguish the task for which it is designed and its handling. This bias is generally not taken into account. The developers often consider their technical abilities as shared and easy to learn, while these abilities have to be deeply acquired by the terminologist before the use of a tool.

4.2.2. Usability and interaction of the system

All the evaluated terminological systems are run with more or less complex command line in a terminal. As SynoTerm provides also a graphical interface, we consider it separately.

While the term extraction with ACABIT is carried out by a single command, both LEXTER and FASTER propose several menu items. We note various problems with these items:

- Their label is ambiguous or not explicit.
- The vocabulary in the item label refers to previous research work of the designer. No explanation is available in a package.
- The file management is too restrictive: file name must be known in advance and once the system is launched, the directories cannot be examined.

Moreover, we notice two problems common to all the systems concerning term extraction and relation acquisition:

- The running steps, the progress and the remaining time of the process cannot be known.
- The process end implies the system exit. So, the user needs to carry out term extraction or relation acquisition on several sets of data, he has to run the system several times.

As regarding to SynoTerm, it proposes a standard interface which makes easier file management and validation of the results. A description of the menu items is provided in the package. However, this interface has several flaws:

- The layout and the label of some objects (buttons, text widgets) have ergonomic problem.

		Terminological systems			
		LEXTER	ACABIT	FASTER	SynoTerm
Terminological meta-criteria	Tool purpose	Term extraction	Term extraction	Term structuration	Term structuration
	Strategy	Ling/Stat	Ling/Stat	Ling	Ling
	Parameters	-	-	-	-
	Functionalities	-	-	+	+
	Result quality				
	Exchange formats	-	-	-	-
Ergonomic meta-criteria	Usability	-	-	-	-/+
	Technical quality	-	-	-	-
	General feeling				

Table 7: Summary of the evaluation of the terminological systems

- The use of scrollbar is sometimes inopportune or useless.
- No information is given about the file outputs.

In text mode, SynoTerm proposes the functionalities of the interface as options. But the validation cannot be carried out in this mode. In that regard, the both graphic and text mode are complementary.

4.2.3. Technical quality

The four terminological systems we evaluated are designed for the UNIX operating systems. While we notice the use of the configuration tools (`autoconf` and `automake`) for SynoTerm, they are provided as compiled or not in a package with shell scripts. So, as most of the prototypes, the setup but also the use of the tool require the user to master operating system functionalities.

Moreover, a file with explanations summarizing the process are generally not provided, except ACABIT. As regards to the data, only ACABIT provides in description file of the input/output data format. We argue that regarding these technical constraints, a terminologist without knowledge about the UNIX environment could not be able to run these tools.

5. Conclusion

We are interested in the problem of the terminological system evaluation. As the terminology building process requires terminologist control and specific domain knowlegde, terminological systems have to be considered in a cooperative way. In that respect and through several experiments, it appears that recall and precision measures are not adapted to the evaluation of such system.

These problems lead us to develop an global evaluation taking into account the terminologist task. Therefore, we propose a set of criteria describing the features of a system: specific terminological aspects and general criteria. The evaluation help method we proposed structures these criteria hierarchically.

Our aim is not to describe an exhaustive list of criteria for the terminological system evaluation, but rather to emphasize this problem. While the remarks about their usability could seem without interest, they are good examples of some ergonomic flaws leading to important problem about their distribution and their use. The analysis of four terminological tools (see table 7) shows that all these tools present similar flaws: a lack of a interactive management, problems with the vocabulary, the structuration of the dialog and the guidelines for the user.

We argue that these usability aspects must be taken into account in the development of effective terminological system and to make easier their use for the terminologists.

6. Acknowledgements

This work is the result of observations carried out during several experiments for which various terminological systems have been used. We would like to thank Didier Bourigault (CNRS), Béatrice Daille (IRIN) and Christian Jacquemin (Limsi) who make their systems available for us.

7. References

Yasmina Abbas and Marie-Luce Picard. 1999. Exemple de pratique terminographique en entreprise. *Terminologies Nouvelles*, (19):124–131.

Gilles Adda, Joseph Mariani, Patrick Paroubek, Martin Rajman, and Josette Lecomte. 1999. Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 1999)*, pages 15–24, Cargèse, France, juillet.

Houssem Assadi. 1997. Knowledge acquisition from texts: Using an automatic clustering method based on noun-modifier relationship. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic (ACL'97), Student Session*, pages 504–506, Madrid, Spain.

Houssem Assadi. 1998. *Construction d'ontologies à partir de textes techniques – Application aux systèmes documentaires*. Thèse de doctorat en informatique, Université de Paris 6, Paris, France.

Christian Bastien and Dominique Scapin. 2001. Évaluation des systèmes d'information et critères ergonomiques. In Ch. Kolski, editor, *Interaction homme-machine pour les SI - Environnements évolués et évaluation de l'IHM*, volume 2 of *Série Informatique et SI*, Paris. Hermes.

Didier Bourigault and Benoît Habert. 1998. Evaluation of terminology extractors: Principles and experimentation. In *Proceedings of the First International Language Resources and Evaluation (LREC'98)*, pages 299–305, Grenade. ELRA.

Didier Bourigault and Monique Slodzian. 1999. Pour une terminologie textuelle. *Terminologies nouvelles*, (19):29–32.

Didier Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, pages 977–981, Nantes, France.

Ido Dagan and Ken Church. 1994. *Termight*: Identifying and translating technical terminology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP'94)*, pages 34–40, Institut for Computational Linguistics. University of Stuttgart, Germany.

Béatrice Daille. 1995. Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *T.A.L.*, 36(1-2):101–118.

Laura Davidson, Judy Kavanagh, Kristen Mackintosh, Ingrid Meyer, and Douglas Skuce. 1998. Semi-automatic extraction of knowledge-rich contexts from corpora. In *Proceedings of Computerm'98 (First Workshop on Computational Terminology)*, pages 50–56, Coling-ACL'98, Université de Montréal, Montréal, Quebec, Canada.

Widad Mustafa El-Hadi and Christophe Jouis. 1998. Terminology extraction and acquisition from textual data : Criteria for evaluating tools and methods. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98) – Poster Session*, pages 1175–1178, Granada, Spain.

Katerina T. Frantzi, Sophia Ananiadou, and Junichi Tsujii. 1997. Automatic term recognition using contextuel cues. In *Proceedings of the Second Workshop on Multilingality in software Industry: The AI Contribution (MULSAIC'97) - Workshop WL1, IJCAI'9*, Nagoya, Japan, August.

Daniela Garcia. 1998. *Analyse automatique des textes pour l'organisation causale des actions. Réalisation du système informatique* COATIS. Thèse de doctorat nouveau régime en informatique, Université de Paris-Sorbone (Paris IV), Paris, France.

Daniel Gouadec. 1990. *Constitution des données*. Afnor Gestion, France.

Benoît Habert, Elie Naulleau, and Adeline Nazarenko. 1996. Symbolic word clustering for medium-size corpora. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, volume 1, pages 490–495, Copenhagen, Danmark, August.

Thierry Hamon and Adeline Nazarenko. 2001. Detection of synonymy links between terms: experiment and results. In *Recent Advances in Computational Terminology*. John Benjamins. À paraî tre.

Thierry Hamon, Adeline Nazarenko, and Cécile Gros. 1998. A step towards the detection of semantic variants of terms in technical documents. In *Proceedings of 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 498–504, Université de Montréal, Montréal, Quebec, Canada.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of 14th International Conference on Computational Linguistics (COLING'92)*, pages 539–545, Nantes, France, August.

Ulrich Heid, Susanne Jauss, and Katja Krüger. 1996. Term extraction with standard tools for corpus exploration – experience from german. In INDEKS-Verlag, editor, *Proceedings of TKE'96: Terminology and Knowledge Engineering*, pages 139–150, Vienna, Austria, August.

Olivier Hû and Philippe Trigano. 2000. A tool for evaluating using dynamic navigation in a set of questions. In J. Vanderdonckt & Ch. Farenc, editor, *TFWWG'2000, International Workshop on Tools for Working with Guidelines*, Biarritz, octobre. Springer-Verlag.

Olivier Hû, Philippe Trigano, and Stephane Crozat. 1999. Considerating subjectivity in software evaluation - application for teachware evaluation. In J. Vanderdonckt & A. Puerta, editor, *CADUI II, Computer Aided Design of User Interfaces*, pages 331–336, Louvain. Kluwer Academic Publisher.

Nancy Ide and Jean Véronis. 1995. Encoding dictionaries. *Computers and the Humanities*, 29:167–179.

Christian Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In S. Wermter, E. Riloff, and G. Scheler, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438, Springer.

Margaret King. 1997. On the notion of validity and the evaluation of mt systems. In Harold Somers, editor, *Terminology, LSP and Translation*, volume 18, pages 191–203. John Benjamins.

André Le Meur. 1998. GENETER: A generic format for the distribution and reuse of heterogeneous multilingual data. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98) – Poster Session*, pages 1165–1168, Granada, Spain.

Marie-Claude L'Homme, Loubna Benali, Claudine Bertrand, and Patricia Lauduique. 1996. Definition of an evaluation grid for term-extraction software. *Terminology*, 3(2):291–312.

James Mayfield and Charles Nicholas. 1992. SNITCH: Augmenting hypertexts documents with a semantic net. In *Proceedings of the Conference on Information and Knowledge Management (CIKM-92)*, pages 146–152, Baltimore, USA.

Bertrand Meyer. 1997. *Object Oriented Software Construction*. Prentice Hall. 2de édition.

Emmanuel Morin. 1998. Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes. In *Actes de la Conférence TALN 1998*, pages 172–181, Paris, France.

Sung H. Myaeng and Ming Li. 1992. Building term clusters by acquiring lexical semantics from a corpus. In *Proceedings of the Conference on Information and Knowledge Management (CIKM-92)*, pages 130–137, Baltimore, USA.

Philip Resnik and David Yarowsky. 1997. A perspective

on word sense disambiguation methods and their evaluation. In *Proceedings of the Lexicon Special Interest Group (SIGLEX'97)*, Washington, DC.

Karen Spark Jones and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems*. Lecture Notes in Artificial Intelligence. Springer.

Padmini Srinivasan. 1992. Thesaurus construction. In William B. Frakes and Ricardo Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, chapter 9. Prentice Hall, New Jersey.

Stan Szpakowicz, Stan Matwin, and Ken Barker. 1996. WordNet-based word sense disambiguation that works for short texts. Technical Report TR-96-03, Department of Computer Science, University of Ottawa, Ontario, Canada.