The ARC A3 Project:
Terminology Acquisition Tools: Evaluation Method and Task

Widad Mustafa El Hadi mustafa@univ-lille3.fr

Ismaïl Timimi timimi@univ-lille3.fr

Annette Béguin beguin@univ-lille3.fr

Marcilio De Brito mdebrito@noos.fr

UFR IDIST & CERSATES (CNRS UMR 8529)
Université Charles De Gaulle, Lille 3
BP 149, F-59 653 Villeneuve D'Ascq, France

Abstract

This paper describes the work achieved in the Concerted Research Project ARC A3 supported and coordinated by the AUF[1], former Aupelf-Uref[2]. The project deals with the evaluation of term and semantic relation extraction from corpora in French. Eight participants, both from public institutions and industrial corporations were involved in this project and were responsible for producing corpora suitable for extraction tasks and elaborating a protocol in order to evaluate objectively *terminology acquisition tools*. This expression covers respectively, term extractors, classifiers and semantic relation extractors. The paper also reports on the methodology used for comparing four term extractors, one classifier and three semantic relation extractors during the 2000 evaluation campaign. There are also several by-products of this campaign: first, two corpora which can be used for NLP system development and evaluation as the AUF recommended; and then terminology products: for each corpus a list of terms characterizing the field is available. We are not giving details about the results but rather an assessment of what the evaluation of Terminology Extraction Tools is: how was it done, what were the difficulties, which are the advantages and disadvantages of the adopted protocol, what are the limits and how should we proceed for future testing.

1 The ARC A3 Program

ARC A3 is a project of the ILEC[3] group coordinated and founded by AUF. It was started in 1995 in order to promote research in the field of terminology acquisition. The ARC A3, "***Term and Semantic Relation Extraction from Corpora in French***" project aim is to test software capabilities in term and semantic relation extraction from corpora in French. Systems submitted to this evaluation are designed by French and Canadian research institutions (National Scientific Research Center and Universities) and/or private businesses. These systems have been extensively described in our previous work (cf. Béguin, *et al*., 1997, 2000; Jouis *et al*., 1997; Mustafa El Hadi *et al*., 1996a, 1996b, 1997 & 1998;). The first phase of the project has been directed towards testing the systems on one corpus[4] (trial run) and towards elaborating a workable protocol based on this experience. The first results were presented during the first conference of *JST*[5] (cf. Béguin *et al*., 1997, 2000). This article reports on the second and final evaluation campaign.

2 ARC A3 Organization

ARC A3 brings together four kinds of actors: a coordinator who plays an organizational role (schedule, quality control of corpora, data production, etc.), corpora providers; participants of the test and two scientific advisors. The action has been coordinated by the University of Lille 3. The organizing team in cooperation with the discussion group made up of representatives of each participating team and two scientific

[1] The *Association des Universités Francophones*

[2] AUPELF is the "Association des Universités Entièrement ou Partiellement de Langue Française", an NGO whose mission is to promote the dissemination of French as a scientific medium.

[3] *Ingénierie de la Langue, Linguistique-informatique et Corpus écrits.*

[4] *SPIRALE*, a periodical dealing with education and pedagogy issues. Each periodical sizes around 200 pages.

[5] *Journées Scientifiques e Techniques de Francil, Avignon, France, 1997.*

advisors are supposed to co-operate in defining a methodology for testing the systems.

2.1 Participating Systems

The systems are designed by French and Canadian research institutions. There were ten registered participants at the beginning of the project and three withdrew later for a variety of reasons. The organizers then launched another call for participation in July 1999 and three more participants joined the project (two private enterprises (Xerox and Logos) and a public institution (the University of Grenoble). Logos and the University team later dropped out for reasons unrelated to the program. When the final campaign was launched in 2000 there were eight systems remaining under evaluation.

Fig. 1. Participating Systems

Software	Affiliation
Acabit	IRIN (Nantes)
Ana	IRIN (Nantes)
Conterm	LANCI (Montréal)
Iota	CLIPS-IMAG (Grenoble)
Lexter	ERSS (Toulouse)
Seek-Java	CAMS-LALIC (Paris 4)
Loria[6]	LORIA (Nancy)
Xerox-Termfinder	XEROX (Grenoble)

2.2 Overview of the Tested Tools

Terminology plays a major role in information processing and management and in specialized communication. Its role has been enhanced by the spread of automation and by the availability of electronic corpora. These two factors have had a massive impact on many different applications: systematic terminology[7] building, natural-language interface design, lexical units management for specific use in some sub-languages and technical writing, thesaurus construction, translation and indexing as well as the recent growth of cross-language information retrieval (CLIR).

If we focus on the tools, presented in our evaluation project, from the point of view of their functions and of the purposes for which they were designed), there are three categories: "Term Extractors", "Classifying Tools", and "Semantic relations extraction tools". As we already mentioned, these systems were extensively described in our previous publications.

2.2.1 Term Extractors (TE)

We will briefly describe the basic idea underlying TE tools. Most of the extracting tools consider terms as noun phrases. Systems identify terms by using frequency, distribution and category-pattern matching (Daille *et al.* 1995; Dagan, 1996; Lauriston, 1994). All lexical units contained in a given text are analyzed and matched to patterns (typical forms of terminological units) described in rules. More term extractors are accounted for elsewhere (L'Homme, 1996; Kageura *et al.*, 1996; Dagan *et al.*, 1994). Some of the systems described by these authors are tested in the framework of our valuation project (Acabit, Lexter, and Ana).

2.2.2 Classifiers and Semantic Relation Extractors (SRE)

Terminology resources are increasingly seen as structured data i.e. as a network of terms organized by relations. Pure alphabetical lists can hardly be used except for bilingual reference tools. The variety of tools, their functions and the different possible uses offered within the framework of ARC A3 shows this need. Consequently such lists of terms are quite difficult to evaluate except by specialists in the relevant fields which makes it a rather constraining process.

Structuring terms by semantic relations or in classes is useful for the following applications: Index-making for on-line technical documentation; browsing; information access and retrieval; building thesaurus and ontologies for information systems.

Many applications and extraction methods relevant to these tools have been described in the literature. The systems tested in the AUF framework are geared towards a variety of

[6] This name is used for practical reasons since no final software name has yet been chosen

[7] A *holistic* list of terms drawn from a representative corpus characterizing and describing a field of knowledge. In order to be of any use this type of list must be subject to a structuring which is an important step towards exploiting extraction results.

applications ranging from rough semantic relation extraction, through indexing, thesaurus construction to knowledge-based system modeling (see figure 2).

Classifiers and semantic relation extractors are tested within the same framework as the one used for evaluating term extractors. The first category is characterized as classifying tools. Their role is to build classes of networks of terms linked to a major one. This category consists of statistical and/or connectionist models such as Conterm. It is the only classifier tested within the framework of this campaign.

The second category includes semantic relation extractors which focus particularly on semantic relations (Iota, Loria and Seek-Java). A complete description of all the systems which were tested (main characteristics and purposes, description as far as approaches are concerned) is documented in previous work.

3 Evaluation paradigm

Evaluation activities are a corollary of the quick development of NLP tools in general and of terminology extraction in particular. It thus became necessary to evaluate these tools on objectively based criteria in order to have a clear picture of the state-of-the-art, assess the needs in this sector and hence promote research in this specific field. Moreover, the principal aim of existing testing methods, as reported in the literature, is to come across software errors and then try to adapt them for a particular user environment.

Evaluation paradigm is basically dependant upon two major steps: (i) Creation of textual data: raw or tagged corpora and test material. A corpus-based research is part of the infrastructure for the development of advanced language processing applications; (ii) Test and comparison of systems on a similar data (Cavazza, 1993; Adda *et al*., 2000).

3.1 The ARC A3 Evaluation Approach

The approach we adopted is a black-box qualitative approach[8] The results are compared with the human performance of a task (either experts examining results or using reference lists or both). Moreover comparisons are made with other systems performing the same task. The results are finally calculated and translated in terms of traditional IR measures[9].

The conventional distinction between b*lack-box* and *glass-box* is the following: the former considers only system input-out-put relations without regard to the specific mechanisms by which the outputs were obtained while the latter examines the mechanisms linking input and output. (Sparck-Jones, 1996 p. 26; King, 1996; 1999, among many others).

The qualitative evaluation measures as described by Sparck-Jones 1996, pp. 61-122, are based on observation or interviewing and are broadly designed to obtain a more holistic, less reductive or fragmented view of the situation. It is moreover more naturalistic. This type of evaluation naturally fits an end-free style. In our case the quality of the results is evaluated by domain experts. We distinguish two types of experts: experts for the three applications tested (*systematic terminology, translation and indexing*); and experts in the two domains of corpora (biotechnology and pedagogy).

Both quantitative and qualitative approaches are goal-oriented, that is focusing on discrepancies between performance results and initial system requirements. Sparck-Jones points out how the two types of measures are deeply interwoven although different in their nature:

- Recall is a quantitative measure of system performance while
- Declared Satisfaction is a qualitative one (i.e. such a measure is really qualitative even if the result of applying it to a set of users is a percentage figure).

The qualitative approach in the evaluation process is the easiest one for end users. It means giving a value judgment on how the system globally works (Cavazza, 1993; Chaudiron, 2000). The dominant approach today is towards quantitative evaluations which are considered as more objective and reproducible than the qualitative approach (EAGLES-1 1996; ISLE 2001). The main attempt of these approaches is

[8] This approach is adopted and validated by the vast majority of participants to the test in June 1999. The organizers have slightly adapted the protocol because more participants joined the ARC after the validation of the protocol.

[9] We chose to accompany the qualitative approach (mainly based on manual evaluations) by a translation of the manual evaluations into numerical scales of values (see below for more details).

to translate the concepts of relevance and quality into numerical data. Statistical approaches such as MUC 2 and TREC 3 are frequently used for this type of evaluation. (Chaudiron, 2000).

3.1.1 The merits of a black-box evaluation

Obviously this approach has its pros and cons. But it can be justified on the following basis:

- Since most developers cannot provide us (as test organizers) with their systems, the only way was to send them the text corpora and let them provide us with the results. A glass-box evaluation would have required an examination of the systems by the organizers which would have been impossible except for Xerox's TermFinder and Logos System's Knowledge Discovery, two commercialized systems.
- Even if this approach may be criticized on account of its subjective side, end-users like it because of its usefulness when comparing two or more systems which differ in all their parameter settings. (Chaudiron 2000; Cavazza 1993).
- A black-box evaluation is more oriented towards system's end-user when compared to a glass-box evaluation. For the latter the test will involve analyzing the system's functioning by looking at its different components. Each component is evaluated separately in itself. Such an approach allows for spotting and understanding the causes of dysfunctional results. It is a long term process which requires access to the internal parts of the system and an understanding of the architecture and global strategy of the software. This is obviously a *developer oriented approach* and not an *end-user one* (Chaudiron 2000; Cavazza 1993).
- In spite of its limited scope the evaluation protocol we adopted is used in more complicated NLP tools, such as MT tools. Evaluators examine the systems' output without considering the differences between them (cf. L'Homme, 2001). Last Spring our team took part in a workshop organized by ISSCO (University of Geneva) where we and all the other participants adopted this approach.

3.2 Elements of the Evaluation Protocol of the 2000 Campaign

3.2.1 Evaluation Task

The extraction of terms, of classes and of semantic relations was necessary to test the tools performance in the three following tasks: Systematic terminology (characterizing the tested corpora); (ii) Translation; (iii) Indexing.

This means in practice: what is the relevance of terms, classes and semantic relations provided by the systems being tested? Do the terms, classes and semantic relations satisfy minimum requirements? Do we need to define a minimum level of terms, classes, semantic production? Are discrepancies meaningful? For example, it could be that most of the systems being tested are having qualitatively poor outputs, while only one or two produce worthwhile results. Within this perspective the idea was to submit the results to specialists. We distinguished for the purpose of this campaign two types of human expertise as we mentioned above.

3.2.2 Test material

Evaluation data can normally be divided into two different categories (i) representative samples of the tested corpora (ii) test material, which, in our evaluation framework, is made up of both custom-designed lists and real life lists / thesaurus.

3.2.2.1 Corpus

Two corpora were tested: Spirale[10] and INRA[11]. We have chosen a sample representing 10% of each corpus: for Spirale n° 19 was chosen. As for INRA corpus, the providers of this corpus suggested 8 articles (603, 604, 607, 609, 631, 666, 732, 740).

3.2.2.2 Reference Lists

These lists are standard human professional results which can be used as performance exemplars or norms for comparison. This type of data is considered to be a gold standard (see SensEval, Kilgarrif 1998; ISLE 2001).

For the INRA corpus the following lists have been created:

For translation two lists were processed (i) a list created by a novice translator (ii) another one by a confirmed professional translator.

[10] 423 texts, 16 mega bytes

[11] 51 texts, 2,2 mega bytes.

For indexing: six lists were created both by professional and by non professional indexers.

We are not developing these lists in this paper given the limited scope of this type of evaluation from an indexing point of view. Hence the limited interest of term extraction tools for human indexing. We will however comment on the terminology lists provided by the two corpus providers, INRA (Institut National pour la Recherche Agronomique i.e. National Institute for Agronomic Research), the *Francis list* of INIST[12]) and the translation lists.

As far as INRA corpus is concerned:

We think that our evaluation task could have given better results if the lists had been more representative of a systematic terminology activity. For the INRA corpora, for example, only 113 terms were chosen by the experts to represent their terminology. Our estimation is that, 113 terms only constitute a poor representation of an activity. It would have been a good idea to have specialists establish the lists of terms and to compare those to the systems' output. Even if this work is time consuming it makes for a better evaluation of the systems' productivity. As far as indexing is concerned the interest of these lists is quite limited and we think that a lot of time has been lost in drawing them up and even grooming them. From a general point of view the tools we have considered, especially term extraction ones, only have a limited interest for indexing contrary to other tools (semantic relation extractors) they have not been conceived for this purpose. This point of view is shared by their own designers. However, some of the semantic extraction tools are adapted for indexing among their other applications (Iota and Loria, for instance).

As for Spirale corpus:

Terminology (i) Thesaurus *Mobis*, (educational sciences section) (ii) *Francis* list (of the INIST, covering the complete volume on educational sciences section).

Three lists for indexing: - *Dictionnaire encyclopédique de l'éducation et de la formation*[13]. - CRDP list[14] de Lille. - *Bréhier list* (*PRCE* in documentation).

3.2.2.3 Unified Presentation Format

The protocol we suggested was based on the previous evaluation sessions. The layout of some results could at times make the task of evaluation difficult. In some cases, good graphic presentation (conceptual graphs, etc.) could hide a poor term extraction and hence influence the evaluation. Conversely a system which has the capacity to extract relevant terms and semantic relations but whose layout is poor can influence the evaluation process. To prevent this, participants have been asked to adopt a unified format for their presentations for 2000 evaluation campaign.

3.2.2.4. Non-unified Tagging

Given the fact that system designers have different processing possibilities, some of the systems use an independent tagger, others have an integrated one which is part and parcel of their system. The organizers decided to allow the participants their own choice in terms of tagging methods.

3.2.2.5. Evaluation Measures

Given the three tasks to be performed (indexing, systematic terminology and translation), the usual notions of recall and precision can be used to evaluate the quality of results when matched with a manually-produced reference list. Performance failure at this level can be interpreted in terms of silence and noise (see below).

3.2.2.6. Automatic Matching by EvalTerm

If the qualitative approach offers the easiest form of systems evaluation it nevertheless retains two major drawbacks: (i) it makes up for a very boring job when there are too many results (ii) judgments can easily be slanted by the subjective approach of the expert.

Our protocol being based on the qualitative black-box principle where parameters are hard to quantify we chose to apply traditional IR measures, recall and precision which normally accompany qualitative evaluations:

R = number of correct extractions / number of reference extractions.

P = number of correct extractions / number of proposed extractions[15]

Since the manual matching of lists proved to be long and complicated due to the huge size of the

[12] INIST is the National Institute of Scientific and Technical Information. The list they provided is used to index their data-base to complete this part.

[13] P. Champy et C. Etevé. Index pp 1059-1097.

[14] *Centre Régional de la Documentation Pédagogique*.

[15] Or their equivalents in terms of noise and silence: Silence = 1 – Recall, Noise = 1 – Precision

processed data and to a variety of other inconveniences, we chose to automatically calculate these measures. We then decided to duplicate the manual evaluation with its conversion into numerical scales of values.

For this purpose we developed a program which matches the results provided by the software with the reference lists[16] The program compares two lists: L_1 represents the results given by a software and list L_2 is a reference list proposed by an expert[17]. The program output consists of two files: file (a) which contains the elements of L_2 found in L_1 (the relevant terms which the software was able to find). And file (b) which contains some elements of L_2 which have not been identified and consequently were not mentioned in L_1 (the correct terms not found by the software). Through a simple subtraction we can get a file containing the noisy terms of each software.

In our automatic matching we have not included any linguistic treatment for fear of introducing new parameters which would influence the results. Right from the beginning we have noticed that over-productive systems such as Ana or Term Finder are difficult to compare with reference lists because the noise rate becomes irrelevant.

4 An Overview of the Results

4.1 Term Extraction on the two Corpora

We will now comment globally on how the term extractors performed when run on the two corpora for the three different tasks (indexing, systematic terminology and translation):

First, automatic matching concurred with human experience which notices that the systems produce many “ noisy ” terms while on the contrary there are many terms not included in the reference lists but which the experts considered as relevant for systematic terminology. Hence the interest of some of these “ noisy ” terms for enriching and updating reference lists and terminology data bases. Matching the results of the different systems has showed a great similarity between Lexter and Acabit.

As for indexing, if the systems could generally provide relevant and effective help for terminology (systematic terminology, and translation) their contribution to indexing is less obvious. Indexing supposes other mental operations than those needed for terminology construction and simply picking out candidate-descriptors is not enough to supply a reliable form of indexing.

The three core criteria of good indexing are: reliability, selectivity and exhaustiveness. The indexer must hold a balance between exhaustiveness and selectivity. Having too many terms leads to noise and too few to silence. It is on this criteria of selectivity that human processing varies.

Softwares based on term extraction offer a large number of potential candidate terms, connecting them with more or less precise criteria of relevance, mostly of a statistical nature. At this level of processing the indexer has recourse to authorized lists and thesauri i.e. he or she refers to the work of terminologists in structuring the field and attributing a label to each and every concept. The systems which we tried to assess are not yet likely to provide a very effective help to indexing since the results are over-productive in view of the needs.

5 The Classifier and the Semantic Relation Extraction (SRE) Tools

The protocol we adopted specifies the evaluation of semantic relation and class validity, coherence and comprehensiveness on all of the three tasks (i.e. semantic relations examined from the point of view of systematic terminology, translation and indexing). The classes and semantic relations extracted were subject to a comparison with the human performance of these tasks (experts and reference lists), plus a comparison with other systems performing the same task. This qualitative evaluation is measured by the traditional IR performance measures (silence, noise, recall and precision). The first thing we can remark on is that it is very difficult to fulfill the evaluation within our proposed terms of reference. We are presenting hereunder the reasons limiting the scope of our protocol when applied to SRE results.

[16]These lists can be: a) existing lists, real-life lists (thesauri or alphabetical lists, such as Francis List); b) established by the evaluators/indexers (specifically tailored for the three tasks, indexing, terminology and translation).

[17] They are many lists proposed by our experts.

5.1 An Overview of the Results of SRE on the Two Corpora

What we observed is that these tools are too different to allow a useful comparison for the following reasons:

- SRE extract different types of relations and hence are incomparable.
- This difference is linked to the different forms of semantic model implementation. Conversely some extractors are based on models that will not allow the type of relations required for the three evaluation tasks.
- SRE are designed for different functions and have different objectives or carry out different tasks.
- These differences are reflected in the type of output or results.
- Another problem came from the fact that INRA could not provide us with a structured list corresponding to the eight selected texts. Even if this list had been available, comparing it to the results would have been of limited interest only. The remaining solution was to submit the results to a field specialist.
- Difficulties in interpreting the non-labeled Semantic Relations. Fig. Two shows these differences:

Fig. 2. Synthetic comparison table for SRE

	IOTA	LORIA	SEEK-JAVA
Objectives	Building indexes of one or more levels (layers) for document retrieval	Scientific & technical watch (identifying rare or new information	a) Cognitive text organization b) Extraction of Labeled Semantic relations between terms in a thesaurus or a network of terms c)Constructing/modeling Knowledge-based systems
Functions	Information seeking systems, automatic extraction of complex indexes	Semantic relations extraction	Semantic relations extraction and representation Presentation in a conceptual graph fashion Building relational data bases
Out-put	Lists of potential candidate terms ranked by with frequency Terminology networks	non-labeled Logico-Semantic Relations between terms Classes of terms	A descriptive network/graph of terms linked with semantic relations between complex or simple terms, on the one hand and a triplet of argument-relation-argument on the other assembled in a relational data-base

- Moreover, it is difficult or even impossible to measure silence using a protocol based on IR systems performance measure.

- Without a prior knowledge of the missing possible relations one cannot account for the *silence* measure.
- To account for noise, a thorough knowledge of both the semantic model and the field of knowledge is required.
- These observation are also valid for *recall* and *precision* measures.

We can thus say for the time being that SRE cannot be assessed by the protocol since their results cannot be matched.

The field specialist[18] *gave the following account*: “It is essential to have an interface to manipulate and interpret the relations. Everything seemed somewhat inconclusive. At times the relation “fits well”, at times it does not at all. Results are not always relevant and it is difficult to trust this type of analysis on its own if one is not at the same time be conversant with the domain, since some of the relations can be wrong.

For Iota, concept extraction seems generally quite relevant**.** However one has to wonder about the relevance of a number of extracted concepts which are not at all relevant to the field. How did these non-specific concepts get extracted more easily than others ?

As for the table on Conceptual Semantic Dependence[19] it is hard to draw any conclusions from it since it offers only one semantic label for any relation.

The Iota approach is more global than the Seek-Java one since the relations are based on the whole document and not only at the level of one sentence. These two softwares are thus difficult to compare since their purpose is not the same”.

5.2 Conterm, the Classifier: an ad hoc Evaluation

Given the difficulties we listed above and the fact that it was impossible to compare Conterm with other systems performing the same task. The only possible evaluation for Conterm would have been a *progress evaluation* for this sole classifier of the campaign[20]. This problem shows again the limits of our Protocol. The Conterm lists were matched to an automatically produced

[18] Patricia Volland-Neil, from INRA-Tours

[19] The evaluator is referring to the tables accompanying the results provided by the system’s designer.

[20] The protocol is not suitable for its evaluation. After the withdrawal of another participant who had also presented a classifier, only this one remained.

untagged list of terms which corresponds to the eight texts of the INRA corpus. The most important element in its evaluation is not that we matched its results with a tagged list but that the results had been matched with indexers' and/or experts lists and that we could observe the correspondence between Conterm's output and the lists. It does not mean that Conterm is good for indexing but that the classes suggested by this tool embody conceptual attributes which are close to the logic underlying the human selection of candidate-terms suitable for indexing, namely its rich lexico-semantic network.

6 Concluding Remarks

- This evaluating action provided us with an awareness of the State-of-the-art in the field of terminology acquisition tools. It also allowed us to test evaluation paradigms, demonstrating how difficult it was to apply a single evaluation protocol to a variety of systems operating along different lines.
- The discussions among participants aiming at the creation of a testing protocol resulted in the definition of an evaluation procedure and in an assessment of their relative merits. The comparative study of the systems' out-put also enabled a better understanding of the performances of the wide range of techniques involved. As by-products of the project two corpora can be used in further evaluation campaigns and a set of material tests (real-life and constructed or specifically tailored one that can be shared during future evaluations).
- The evaluation results can be used predictively for system design, development or modification

The limits of our evaluation approach can be sketched in the following manner:

- If the adopted protocol based upon reference list can be applicable to the two tasks (translation and terminology) it is hardly applicable to indexing tasks.
- It is not adequate to account neither for the classifiers nor for the SRE.
- Several questions remain unanswered:

a) first, is it possible to fully automate evaluation procedures? Then is it possible to abandon test material, such as reference lists or other type of human-made data, which are considered as a kind of gold standard reusable for other evaluation campaigns? (see our recent experience in MT evaluation workshop, April 2001[21].

b) As far as semantic relation extraction is concerned, is it possible to automate SRE valuation procedure in the way Grefensttete (1994) does?

7 Future Directions

1. Exploiting Results: the Campaign's Side Benefits:
Full treatment of the *Spirale* corpus will allow the creation of an index of all the reviews past numbers, which fulfills the moral contract made with its Editorial Board in exchange for getting the corpus free of charge. In addition, these results can help broaden the terminological repository for the education sciences, especially in drawing up the *Francis Thesaurus* which covers all education sciences.

2. Towards Trans-Systemic Integration: The output of the systems are divergent but can in some cases be complementary. In fact the preliminary results drawn from the first evaluation in 1997 (cf. Béguin *et al*. 2000) have led us to consider the feasibility of trans-systemic integration for strengthening their automatic terms identification capabilities. The idea is to combine two or three different types of systems in order to specify various integrated production processes. Systems could

[21] "Setting a methodology for Machine Translation evaluation". The context: evaluation of a translation made by an MT System on the following source text: INRA corpus text N°604 " corpus biotechnologique sur la reproduction chez l'animal " Source language: French - Target language: English. We carried out some manual testing but with the objective of setting a rough methodology that might be irrelevant for translating huge size corpora. The tool we used was a non interactive French / English MT System with a basic French/English dictionary that does not include any specific terminology. We had two indexes (a French index and an English index of domain specific expressions, but they are not aligned). They have been provided by the INRA and considered as gold standard. We used the indexes to create a specific dictionary in order to feed the MT systems with this specific lexical data. The next step is to assess the impact of specific terminology when integrated to an MT system by comparing the results of the two translations we get: with and without specific terminology.

increasingly be seen as parts of these integrated production processes.

3. Towards User-Oriented Evaluations: in the light of the results obtained in this campaign the most suitable type of evaluation would be a user-oriented one. Other types of approaches[22] can be designed, such as adequacy evaluation[23] which can to some extent be adopted for our case but we have to define a more strict user profile.

4. Towards developing interfaces for validating the results: even if we opted for a unified presentation format for the reasons mentioned in section 3.2.2.3, we however think it is essential for future campaign organizers to have an interface to manipulate and interpret the results (validating term, relations and classes). This type of interface can dramatically facilitate the interaction with the evaluators and the end-user of these tools.

5. Designing tools for generic bi-lingual production, allowing *ad hoc* extractions through *ad hoc* interfaces.

6. Capability to share resources in the future (test material such as gold standard lists, real-life and/or constructed ones).

7. Developing automatic evaluation tools such as *Evalterm* which can be reused in similar future evaluations.

8. Hypothesis are still to be tested for semantic relations extraction: results of the various semantic extractors will be of different quality depending on the type and nature of corpora (domain and genre) chosen (cf. also Condamines *et al*. 98; Davidson *et al* 98, among many others).

8 Acknowledgements

The authors gratefully acknowledge the financial assistance provided by *the* AUF (Association des Universités Francophones) in funding the general research project within which this paper was written.

9. References

Adda, G., Lecompte, J., Mariani, J., Paroubek, P., Rajman, M. (2000). Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de partie du discours pour le français, Chibout, K., Mariani, J., Masson, N., Neel, F. éds., (2000). *Ressources et évaluation en ingénierie de la langue*, Duculot, Coll. Champs linguistiques, et Collection Universités Francophones (AUF), pp. 645-664.

Béguin, A, Jouis, C, Mustafa El Hadi , W, (1997): "Evaluation d'outils d'aide à la construction de terminologie et de relations sémantiques entre termes à partir de corpus", In *JST'97*, FRANCIL, AUPELF-UREF, Avignon, avril 1997, pp. 419-426. This article is published in Chibout *et al.* 2001 (eds.) pp 161-179.

Béguin, A., Jouis, Ch., Mustafa Elhadi, W. (2000). Evaluation d'outils d'aide à l'extraction et à la construction automatiques de termes et de relations sémantiques. In Chibout, K., Mariani, J., Masson, N., Neel, F. éds., (2000). *Ressources et évaluation en ingénierie de la langue*, Duculot, Coll. Champs linguistiques, et Collection Universités Francophones (AUF), pp 161-179.

Bourigault, D. (1993). "Analyse syntaxique locale pour le repérage de termes complexes dans un texte", *Traitement automatique des langues* 34(2), pp. 105-117.

Bourigault, D., Jacquemin Ch. &. L'Homme M-C. éds. (1998). *Computerm '98, First Workshop on Computational Terminology,* COLING-ACL'98, 15 August 1998.

Bruandet, M.F. (1989). Outline of a knowledge base model for an intelligent Information Retrieval system. In Information Processing and management, Vol 25, N° 3, 1989.

Cavazza, M., (1993). Méthodes d'évaluation des logiciels incorporant des technologies d'informatique linguistique, Paris, Rapport MRE-DIST, 1993.

Chaudiron, S. (2000). The Relevance of Quality Model for NLP Applications, in Proceedings of RIAO, pp 1568-1577, Paris 12-I4 April 2000.

Condamines, A. et Rebeyrolle, J. (1998). CTKB: A corpus-based approch to a Terminological Knowledge Base, In Bourrigault, D., Jacquemin Ch. &. L'Homme M-C, éds.(1998). *Computerm'98, First Workshop on Computational Terminology*, COLING-ACL'98, 15 August 1998, pp. 29-35.

Dagan. I. and Church, K. (1994). Termight: Identifying and Translating Technical

[22] Spark-Jones et al. 1996; King 1999, among many others, identified more three types of evaluation processes: the *progress evaluation, the adequacy evaluation and* the d*iagnostic evaluation.* The first and second types are used for comparative benchmarking.

[23] A*dequacy evaluation* aims at finding out whether a system or product is adequate to someone's needs. This type is typically done when thinking of acquiring a system.

Terminology. In: Proceedings of 4th Applied NLP Conference 1994, 34-40.

Daille, B., (1994). ACABIT: une maquette d'aide à la construction automatique de banques terminologiques monolingues ou bilingues. In Class, A., Thoiron, P, , Béjoint (eds) Lexicomatique et Dictionnairiques, pp. 123-136, Beyrouth 1996.

Daille, B. B. Habert, C. Jacquemin et J. Royauté (1995). "Empirical Observation of Term Variation and Principles for their Description", *Terminology* 3(2), pp. 197-257.

Davidson, L., Kavanagh, J., Mackintoch, K., Meyer, I., Skuce, D. (1998). Semi-Automatic Extraction of Knowledge-Rich Contents from Corpora, In Bourrigault, D., Jacquemin, Ch. &. L'Homme M-C, éds. (1998). Computerm'98, First Workshop on Computational Terminology, COLING-ACL'98, 15 August 1998, pp. 51-56.

EAGLES (1996)

http://www.issco.unige.ch/projects/eagles.

Enguehard, C. (1993). Acquisition de terminologie à partir de gros corpus. Informatique et langue naturelle ILN' 93, Nantes, pp. 373-384, décembre 1993.

Grefenstette, G. (1994). Explorations in Automatic Thesaurus Discovery, Boston: Kluwer Academic-Press.

[Gaussier, E. (1995). Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues de termes, Thèse de doctorat, Paris: Université Paris VII.

ISLE (2001) MT Evaluation Classification, Expanded Classification (2001). http://www.isi.edu/natural-language/mteval/2b-MT-classification.htm.

Jouis, C., (1993). Contribution à la Conceptualisation et à la Modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype: le système SEEK. Thèse de doctorat. EHESS. Paris. 1993.

Jouis, C., Mustafa El Hadi, W. (1997), AUPELF Project: Term and Semantic Relation Extraction Tools. Evaluation Paradigms, In Proc. of the Speech and Language Technology Club Workshop " Evaluation in Speech and Language Technology ", Univ. of Sheffied, June 17-18, Sheffield, UK, pp. 106-113.

Kageura, K. and Umino, B. (1996). Methods of Automatic Term Recognition: A Review. In: Terminology Vol. 3(2), 1996, 259-289.

Kilgarrif, A., Rosenzweig, J. (1998). English SENSVAL: Reports and Results.

King (1999) EAGLES Evaluation Working Group, report, http://www.issco.unige.ch/projects/eagles.

King M. (1996) EAGLES, Workshop, University of Geneva http://www.issco.unige.ch/projects/eagles.

Le Priol, F, Chavallet, J-P., Bruandet, M-F., Desclès, J-P. (1998). Intégration d'un système statistique (IOTA) et d'un système sémantique (SEEK) dans une chaîne de traitement permettant l'extraction de terminologies. *Actes Ingénierie des Connaissances*, Pont-à-Mousson, pages 33-40. 1998.

L'Homme, Marie-Claude, Benali, Loubna, Bertrand, Claudine and Lauduique, Claudine. (1996). Definition of an Evaluation Grid for Term-Extraction Software. In: Terminology Vol. 3(2), 1996, 291-312.

L'Homme M-C. (2001). Évaluation d'outils d'aide à la construction automatique de terminologie et de relations sémantiques entre termes à partir de corpus ARC-A3 Rapport final, Montréal, mars 2001.

Lauriston, A. (1994). Automatic Recognition of Complex Terms: Problems and the TERMINO Solution. In: Terminology 1(1), 147-170.

Manzi S. (1999) Test Material for Evaluation, Sandra Manzi, ISSCO - University of Geneva http://www.issco.unige.ch/projects/eagles/ewg99.

Mariani, J. Masson, N., Neel, F., éds. (2000). *Ressources et évaluation en ingénierie de la langue*, Duculot, Coll. Champs linguistiques, et Collection Universités Francophones (AUF), 2000, pp. 13-24, actes des 1ères Journées Francil autour du thème: L'ingénierie de la Langue: de la Recherche au produit, Avignon 15-16 avril 1997.

MUC-3, (1991). Proceedings of the Third Message Understanding Conference. Morgan Kaufmann.

MUC-4, (1992). Proceedings of the Fourth Message Understanding Conference. Morgan Kaufmann.

Mustafa El Hadi, W. & Jouis, C. (1998), Terminology Extraction and Acquisition from Textual Data: Criteria for Evaluating Tools and Method, Proceedings of the First International Conference on Language Resources and Evaluation, Grenada, Spain may 1998, pp. 11750-1178.

Mustafa El Hadi, W. & Jouis, C. (1997), "Natural language processing techniques and their use in data modeling and information retrieval". In: Proceedings of the sixth international study conference on classification research, Knowledge Organization for Information Retrieval, University College London, London, 16-18 June 1997. The Hague: FID, 157-161.

Mustafa El Hadi, W., Jouis, C. (1996a), Evaluating Natural Language Processing Systems as a Tool for Building Terminological Databases, In *Proceedings of the Fourth International ISKO Conference: Knowledge Organization and*

Change, Washington, Library of Congress, July 1996, *Advances in Knowledge Organization*, Vol. 5, INDEX Verlag, Frankfurt/Main, pp. 346-355.

Mustafa El Hadi, W., Jouis, C. (1996b), Natural Language Processing-based Systems for Terminological Construction and their Contribution to Information Retrieval. *In Proceedings of the Fourth International Congress on Terminology and Knowledge Engineering (TKE'96),* Vienna, INDEX Verlag, Frankfurt/Main. pp. 118-130.

Seffah, A. Meunier, J.G. (1995). ALADIN :un atelier orienté objet pour l'analyse et la lecture de textes assistée par ordinateur. In : International Conference on statistics and texts

Sparck-Jones K., Gallier, J.R. (1996). Evaluating Natural Language Processing Systems: An Analysis and Review, Springer, Berlin.

Toussaint, Y. Namer F., Daille, B., Jacquemein, C., Royauté, J., Hathou N., (1998). Une approche linguistique et statistique pour l'analyse de l'information en corpus. In : TALN' 98, Paris, France, 1998.