[*Translating and the Computer 24*: proceedings of the…International Conference…21-22 November 2002, London (Aslib, 2002)]

Tearing out the Terms: Evaluating Terms Extractors

Véronique Sauron

TIM/ISSCO, ETI
University of Geneva, 40 Blvd du Pont d'Arve, CH-1211 Geneva 4
Veronique.Sauron@eti.unige.ch

Abstract

This paper reports on a comparative evaluation of three terminology extraction tools. The evaluation methodology is based on ISO standards and the work of the Eagles evaluation working group. Although concrete evaluation results are reported, the primary interest of the paper is in the development of a standardised methodology for the evaluation of such tools.

Introduction

The creation and maintenance of terminological resources is an integral part of translation work, whether it is carried out through an organised and on-going systematic effort or through more sporadic case-by-case research. The main stages in terminology work can be summarized as: extraction of terms from a corpus; validation of terms found; organisation of validated terms by domain and sub-domain.

Pioneer efforts early in the history of the creation of computer aids for translation built up large computerized terminology banks intended for electronic consultation, stocking the banks with manually created and organised terminology resources. Only recently have research efforts turned to trying to automate the creation of the resources themselves. A number of projects have been able to create automatic extraction tools, which, starting from a corpus in electronic form, identify candidate terms. Some projects go further, and on the basis of parallel corpora of texts and their translations propose not only candidate terms but also possible equivalents in a target language.

Such tools have only recently become commercially available, but the translation community has already begun to show a lively interest. This is scarcely surprising when one considers the lengthy and fastidious manual work that could be avoided if automatic extraction produces satisfactory results. This growing interest in itself justifies the labour of setting up evaluation models and associated criteria, thereby creating a basis which will allow interested parties to evaluate emerging systems and compare results. Putting in place an evaluation methodology applicable to terminology extraction tools is of obvious scientific interest. At a more prosaic level, the motivation for evaluation also comes from a desire to know to what extent and how such tools can be of help to the translator or to the terminologist in his daily work.

This paper describes the practical evaluation of three terminology extraction products, Xerox XTS (version 2.0), MultiTrans (version 1.1) and ExtraTerm (version 5.1). The evaluation is designed along lines suggested by ISO and specialised for language engineering systems by the EAGLES Evaluation Working Group. The evaluation model also takes into account the work of the Association for Terminology and for the Transfer of Knowledge (GTW).

THE ISO/EAGLES Evaluation Framework

The International Organisation for Standardization (ISO) has defined several standards for software evaluation which apply to the evaluation process, comprising five different steps, and to the definition of a quality model base on six high level quality characteristics. The European EAGLES initiative extended these standards to NLP systems through its reports[1], summarized in a *7-step recipe*[2]. Our evaluation constitutes a specific application of this general methodology to a particular kind of NLP tool and a particular context. The account given here focuses on certain of the more application specific steps.

Defining a task model

In its introduction, the EAGLES report states that the purpose of an evaluation "is to determine what something is worth to somebody". In fact, a particular evaluation is essentially a function of the needs of a user, who wants to know what the software being evaluated might give him in terms of support for a particular task to be accomplished. From the perspective of translation and terminology, the user above all wants to know to what extent the software can help him in the task of building up the terminology resources required to carry out the task of translation, whether he is involved in systematically searching for terminology or whether he is doing so on an ad hoc basis for a particular translation task. He will thus expect an extraction program to extract terms relevant to the domain(s) of interest from a corpus previously created as representative of that domain or domains. He will furthermore expect the software to present candidate terms to him for validation. These two tasks, identification and validation of terms, determine the basic parameters of the evaluation.

To ground the evaluation in practical consideration of user needs, as advocated by ISO and by EAGLES, in this paper we take as context a translation service working into French. The service includes both translators and terminologists. The texts they must deal with are relatively short summaries of technical claims and contain much highly technical vocabulary. The translators need in their daily work access to up to date terminology and to elements of phraseology which reflect the linguistic practices of specialists in the technical domain. The terminologists wish to exploit archive material consisting of previous translations in order to build up for the first time a fund of terminology which will help to ensure coherent use of terminology across the organisation. As representative of such a context, we have used a set of abstracts coining from the translation services of the World Intellectual Property Organisation. The texts are abstracts of patent applications, submitted under the rules and procedures of the Patent Cooperation Treaty. Our corpus contains 1'000 abstracts, 500 in English and 500 in French, distributed equally over two different technical domains, A61 (Human necessities, medical) and H04 (Communication). The texts are available in html format. For our purposes, we converted them to .rtf format, a format compatible with all three of the tools being evaluated.

---

[1] EAGLES Work Group (1999) EAGLES Evaluation Working Group, *Final Report, n°EAG-II-EWG-PR.1 (Draft),* Center for Sproktechnologi, Copenhague.

[2] King, M. (1999) The 7-step Recipe for Evaluation Language Technology. In *Proceedings of the European Evaluation of Language Systems (EELS) Conference,* Hoevelaken, April 1999.

The object being evaluated

A critical point in any evaluation is determining what exactly it is that is being evaluated. Terminology extraction tools are relatively new arrivals on the market, therefore it seems to be a good moment to evaluate their usefulness in general, basing a judgement essentially on current performance, and to compare those already marketed as aids to translation. Three tools, Xerox XTS (version 2.0), ExtraTerm (version 5.1) and MultiTrans (version 1.1) seem to correspond to our general aims. They are all commercially available, and are advertised as computerized aids for authors, translators and terminologists. It should be said straightaway however that none of these three is intended as primarily and solely a terminology extractor. Terminology extraction is just one component in a larger system, destined as aid for translation in the case of ExtraTerm and MultiTrans, or as an authoring aid in the case of Xerox XTS. The evaluation reported here will however concentrate only on the terminology extraction component of each of these larger systems. We shall now briefly describe each of the softwares to be evaluated.

Xerox XTS is a suite of terminology tools developed by the MKMS (Multilingual Knowledge Management Systems) group of Xerox. This programme consists of five modules; Xerox TermFinder, a monolingual or bilingual terminology extractor, Xerox TermOrganizer, an interface for terminology management and for creation of dictionaries, Xerox TermOnLine, a web interface for terminology search and consultation within an enterprise, Xerox TermChecker, an authoring aid checking that terminology is correctly used and finally Xerox Web@ssistant, a tool supporting comprehension across languages. The module of interest to us here is obviously TermFinder, which is based on an architecture designed for linguistic development known as XELDA (Xerox Engine for Linguistic Dependent Applications), developed by the research teams at XRCE and at Xerox PARC. XELDA's technology is based on finite state automata, and includes segmenters, lemmatizers, taggers and noun group identifiers using general and specialised dictionaries. The extractor itself can operate in monolingual or bilingual mode. The extraction process differs somewhat depending on the mode. As a first stage, the text is segmented into sentences. In bilingual mode, this phase is followed by alignment, relating segments of one language to corresponding segments of the other language. A tokenisation phase follows, during which sentences are segmented into lexical units. A morphological analyser lemmatises the lexical units, which are then tagged with morpho-syntactic information. A disambiguation phase then uses context to choose the most probable between competing syntactic categories suggested by the tagger. An identification module searches for combinations of words which correspond to "patterns", pre-defined models representing the syntactic structure of noun groups. In bilingual extraction, the program also couples source and target language terms extracted, using algorithms which reflect different probabilities of association between the words of the source and target languages (taking account, that is, that each word of a source language cannot be uniquely mapped onto a word of the target language). The candidate terms are then extracted in the form of lists within a second module called TermOrganizer.

MultiTrans is a fairly complete translation aid which includes a terminology extraction functionality: essentially, the system is a multilingual concordancy system with associated tools allowing the user to search and exploit the results of his search. The process used to create the reference corpus is quite simple, and allows the monolingual extraction of expressions to be done at the same time. These expressions (the term used by the system's creators) are what we have defined as complex lexical expressions, that is sequences of more than one word repeated in the corpus. The maximum and minimum length of expressions can be defined by the user, the limits being 2 words as minimum and 25 as maximum. Clearly,

the longer expressions can be whole phrases repeated in the corpus rather than terms in the purists' sense. MultiTrans is entirely statistics based, with advanced capacities for adjusting and treating raw frequency in order to extract sub-expressions. There is a filter based on exclusion lists which blocks the extraction of certain elements such as articles, certain verbs or prepositions. The result of the extraction is given in the form of lists for user validation.

ExtraTerm, a monolingual and bilingual terminology extractor recently appearing on the market, is one of the modules contained in version 5 of the Trados suite of translation aids, operating independently of MultiTerm. The extraction process is fairly simple: the user provides the system with a corpus of texts from which ExtraTerm uses primarily statistical methods to extract candidate terms. As with most commercial systems, it is almost impossible to find out exactly what the underlying technical basis might be. We can however make some informed guesses based on certain elements and parameters of the program. It looks as though the extraction operation is based quite simply on a calculation of repeated segments. Thus, whenever a word or a series of words appears twice or more in the corpus, it is automatically extracted. This may appear to be quite simplistic: the use of an exclusion list compensates a little for that. The list is the user's responsibility and can be used to block the extraction of, for example, prepositions, determiners, adverbs, conjunctions or other elements the user thinks it might be useful to block. Candidate terms are presented to the user in the form of lists for validation.

Developing a quality model

In any evaluation, we want to determine the quality of a product, in this case a software product. Several authors have designed models aimed at measuring the quality of software product in terms of characteristics or of attributes. [McCall, 1977] and [Boehm, 1978], for example, described a hierarchy of characteristics where each characteristic contributed to overall quality, although they differed in the number of characteristics used. In the 1990s, ISO, the International Standards Organisation, tried to bring together the different views into a quality model containing a set of characteristics to be verified in determining the quality of a piece of software[3]. Generally speaking, the standard distinguishes between internal characteristics and external characteristics. Essentially, internal characteristics pertain to the software itself, for example to the algorithms on which programming is based, to the number of lines of code or to the number of function calls. External characteristics concern the part of the software visible to a user, those characteristics with which the user enters directly into contact. In this evaluation we are only concerned with external characteristics, in other words we will concentrate on a black-box evaluation.

Functionality is certainly the most important of the quality characteristics proposed by the ISO standard, since it determines whether a software does what is required of it and whether it can be integrated satisfactorily into a specific work environment. Its specific attributes are suitability (whether the results are suitable to the task to be accomplished), accuracy (the ability to produce correct and agreed results), interoperability (whether the software can interact satisfactorily with specified other softwares) and security or the ability to block any unauthorized access to programs or data [ISO/IEC JTC1/SC7/WG6 N430: 7]. In our case we have chosen to concentrate on **suitability** defined by ISO as "*The capability of the software*

[3] ISO/IEC 9126-1 (2001), Software engineering, product quality - Part. 1: Quality model, Genève. International Organization for Standardization International Electrotechnical Commission.

product to provide an appropriate set of functions for specified tasks and user objectives" [op. cit., 7]. We also decided to have a close look at **accuracy** which ISO 9126 defines as *"The capability of the software product to provide the right or agreed results or effects"* [op. cit.,7]. In our context this characteristic is of over-whelming importance: lack of precision in the results obtained can mean that a terminology extractor serves no useful purpose whatever. Finally, we chose to work on interoperability that is to say *"The capability of the software product to interact with one or more specified systems"* [op. cit.,7]. Indeed, technology has by now penetrated deeply into the daily lives of translators and terminologists, and it is often the case that they are equipped with translators' work stations, or are in the process of acquiring such specialised environments.

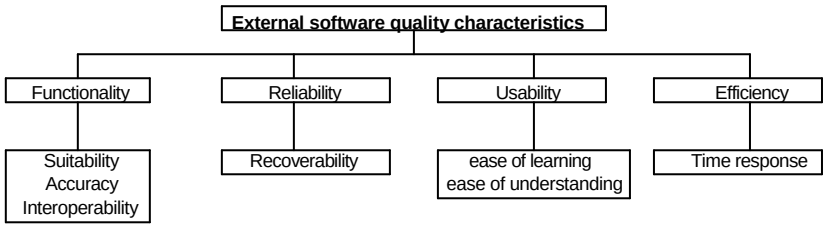
Although less important in our context than functionality, **reliability** is nonetheless an essential quality characteristic which determines the degree of confidence the user has in using the system. ISO defines reliability as "*The capability of the software product to maintain a specified level of performance when used under specified conditions."* [op. cit.,7]. The sub-characteristics of reliability in the ISO definition are maturity, fault tolerance, that is the ability to maintain a specified level of service in the face of software failure or interface violation, and recovery after failure, that is the ability to re-establish a given level of service and to restore the information directly affected by failure, taking into account the time and effort needed for recovery. It seems to us that thorough testing of the reliability of a software implies relatively complex and extensive tests in conditions which permit specific working conditions to be reproduced. This kind of evaluation falls outside the scope of the study reported here. However, one sub-characteristic seems to us of special interest, the sub-characteristic concerning recovery after failure. What makes this sub-characteristic interesting is that terminology extractors will normally be working with large, sometimes very large, amounts of data. It is worth asking what happens if problems arise during term extraction or during validation.

As well as presenting appropriate functionalities and being reliable in use, a system needs to be sufficiently simple to access and attractive in use to encourage a user to use the system, in other words, it should be usable. ISO defines **usability** as: *"The capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions"* [op. cit.,8]. Sub-characteristics are ease of understanding, i.e. the effort that the user needs to make in order to understand the logic behind the software and how it is put into practice, ease of use, or the effort the user needs to make in order to be able to use and control the software, ease of learning, or the effort the user needs to make to learn the application. From the particular viewpoint of our evaluation, this characteristic is quite important. Translators, often even more than terminologists, insist that any aids they are given must be easy to learn and to use, so that they can almost immediately start using them in the translation process, without having to spend hours they do not have to spare in training sessions. For the purposes of the evaluation, we have picked out two of the ISO sub-characteristics, ease of understanding and ease of learning. ISO defines ease of understanding as "*The capability of the software product to enable the user to understand whether the software is suitable, and how it can be used for particular tasks and conditions of use"* [op. cit.,8]. Given the working conditions of many translators, ease of learning is also a very important element when a translator's aid is to be evaluated. ISO defines ease of learning as *"The capability of the software product to enable the user to learn its application"* [op. cit.,8].

The next ISO characteristic, **efficiency** is defined as *"The capability of the software product to provide appropriate performance, relative to the amount of resources used, under stated*

conditions" [op. cit.,9]. Sub-characteristics are defined with respect to time, especially response time and processing time with respect to resources used. We picked out the sub-characteristics relating to time response which we interpreted as the overall capacity of the products being evaluated to produce satisfactory results in an acceptable amount of time.

The next two ISO quality characteristics, **maintainability and portability** have been deliberately omitted from the evaluation described here, primarily because they did not seem relevant from our perspective. The presentation below summarizes those characteristics and sub-characteristics which we have picked out as relevant for our particular evaluation. Even without entering into a discussion of all the characteristics and sub-characteristics set out in ISO 9126, the evaluation we have carried out is sufficient, we believe, to take into account those basic elements which will let the translator or terminologist form an idea of to what extent a terminology extractor may be useful to him in his work, and which may help him to decide what tool or type of tool best suits his needs.



Not all of these characteristics are of equal importance for a given evaluation. From our perspective, accuracy and suitability, and broadly functionality, have to be seen as the most important attribute in a terminology extractor at this point in the development of such tools. As the table below shows, this is reflected in the number of points attached to these attributes. We considered the other characteristics as being equally important. We therefore attached to these attributes the same number of points.

Characteristics	Sub-charac.	
Functionality	Suitability Accuracy Interoperability	20 40 10
Reliability	Recoverability	10
Usability	Learnability Understandability	5 5
Efficiency	With respect to time	10
		100

Developing attributes and metrics

As we mentioned above, ISO characteristics definitions are somewhat too general to furnish us with measurable attributes relevant to our particular task, so we have defined attributes,

based for a part on the GTW-Report[4], each one associated with a metric and a scoring scheme where scores are determined by observing the behaviour of the software with respect to the attribute. Indeed, as the EAGLES report spells out, any characteristic maintained for evaluation must be measurable. If it is not, it needs to be broken down into sub-characteristics and perhaps sub-sub-characteristics until we finally arrive at attributes that are measurable. We have furthermore specified for each metric a rating function which interprets the raw score obtained by the software as being good if the total number of points is given, average if half of the available points is given or unacceptable if no points are awarded. The scores obtained for each attribute are aggregated into a score for the quality characteristic as a whole. We shall now present the attributes and the corresponding metrics we have chosen for each quality characteristic.

As a first step, the **suitability** characteristic was broken down into four measurable attributes: the presence of all the functions described in the documentation; internal consistency and consistency with the documentation; the file formats that can be dealt with; the languages that can be dealt with. With the first criteria, we want then to verify that all the functionalities described in the documentation do in fact exist in the program, and that their behaviour is at least similar to that described in the documentation. To measure these elements, we propose the following rating scale:

Definition	**Points**	**Rating**
Presence of all the functionalities described in the documentation	5	Good
Absence of at least one of the functionalities described in the documentation	2.5	Acceptable
Absence of all the functionalities described in the documentation	0	Unacceptable

We also want to check that the terminology used to describe different functions is consistent and coherent throughout the documentation, within the program itself, and between program and documentation. For measuring internal consistency and consistency between software and documentation, in other words the use of suitable and coherent terminology, we suggest the following rating scale:

Definition	**Points**	**Rating**
Terminology used consistent	5	Good
One or more inconsistencies in the terminology used	2.5	Acceptable
Terminology badly incoherent and inconsistent	0	Unacceptable

In practical terms, the file formats and languages that can be dealt with are perhaps among the most important of all attributes. It is not unknown for an enterprise or a translation service to rush into purchase without a serious evaluation of its needs or study of what is available on the market, and to finish up with software incompatible with their local file formats or incapable of dealing with their working languages. In measuring this attribute, one could consider that if a software is compatible with the formats and languages in daily use, this is a sufficient advantage, or one could go further and suggest that the more formats and languages a system can deal with, the more advantageous this is, since it allows for easy adaptation to new work situations, without passing through conversion processes which are costly both in

4 Criteria for the evaluation of terminology management software. Association for Terminology and Knowledge Transfer.

terms of time and of money. To evaluate this attribute, we have made a distinction between formats that we might call "standard", and other more "specific" formats which appear from time to time such as *wp, asp, sgml, xml, pdf, mif.* Amongst the standard formats we have included the classic text processing formats *(rtf., doc.),* and, in view of the mass of information available on the web, *.html.* We then decided to give 4 points to the product for which standard formats are available, 1 point if specific formats are available and none if the mentioned formats are not available. The maximum score possible is 19. We therefore established the following rating scale.

Definition	**Points**	**Rating**
Maximum points scored (19 points)_	5	Good
Between 9 and 18 points	2.5	Acceptable
Less than 8 points	0	Unacceptable

As far as languages are concerned, our requirements are fairly simple, since our test case is a service translating from English into French. It is extraordinarily unlikely that a terminology extraction program on the European market would not be able to deal with these two languages. We have therefore refined our requirements a little. Not only do we ask that the software be able to treat the language in question, we also require that it treat it correctly. The measures that relate to overall performance (in particular the f-measure) that we shall discuss in more detail in the section on accuracy will deal with this question. We determined that the f-measure results for both languages should be equal for the product to claim a correct treatment of the two selected languages. If the difference in order is superior to 10%, then justifiably the product should be considered to deliver poor language results.

Definition	**Points**	**Rating**
f-measure ENG ≈ f-measure FR	5	Good
f-measure ENG ≈ f-measure FR < 10%	2.5	Acceptable
f-measure ENG ≈ f-measure FR > 10%	0	Unacceptable

Accuracy: We require of a terminology extractor exactly that it should extract the terminology relevant to a domain from a corpus of texts representative of that domain. A priori, we require all and only the relevant terms to be extracted. To use conventional terminology, there should be no noise and no silence. To soften this a little, too much noise will mean that the user spends too much time eliminating invalid candidate terms, and too much silence will mean that the user is forced to search manually for terms the software has missed. Both of these imply considerable loss of time, which has repercussions on the efficiency and effectiveness of the translation process as a whole. The simplest way to evaluate the performance of a terminology extraction tool is to compare the list of terms extracted by the program with a list of terms produced by a human terminologist [L'homme et al, 1996: 302]. This procedure gives "gold standard" status to the list produced by the human.

Recall and precision [Sparck Jones and Galliers, 1996:21] are metrics from information and document retrieval traditionally used to compare a set of reference answers to the answers produced by the program being evaluated. Both metrics presuppose the possibility of good and bad answers among the system results, the bad answers being those which do not correspond to the answers in the reference set based on human judgement. (See also [L'homme et al, 1996: 303]). Habert defines precision as the proportion of relevant answers when compared to the total number of answers given, and recall as the proportion of relevant

answers given compared to the number of possible relevant answers. [Habert, 1997:11]. Simultaneous application of these two metrics seems to be an excellent way of judging the performance of terminology extraction tools. Recall will allow us to measure the software's ability to extract all those terms present in the corpus which the user would have considered to be relevant terms. Precision will let us measure the ability of the system to extract **only** those words and noun groups that the user would consider to be terms. By combining these two metrics, we can measure the completeness (recall) of the results produced by the program and the purity (precision) of these same results: in other words, we can verify whether all and only the relevant terms are extracted. Precision and recall are usually represented by values going from 0 to 1. The closer the value is to 1, the better is the performance of the system being tested.

As Van Rijsbergen [Van Rijsbergen, 1979:129] remarks, ideally an extraction should produce the maximum score for both precision and recall[5]. This situation can be represented by a contingency table:

Terms	Valid	Invalid
Extracted	N1	N3
Not Extracted	N2	N4

In our case, recall is the total number of valid terms extracted by the program being evaluated divided by the total number of valid terms present in the corpus, i.e. $\frac{N1}{N1 + N2}$. Precision is the number of valid terms extracted divided by the total number of terms extracted, i.e. $\frac{N1}{N1 + N3}$. From these precision and recall measures we can derive noise and silence. Noise is the proportion of non-valid terms compared to the total number of terms extracted. It is clear that too much noise will have negative consequences on the work, incumbent on the user, of sorting valid terms from non-valid terms. Conversely, silence is when a valid term is not extracted by the program. [L'homme et al, 1996: 303]. Mathematically, the rate of noise is defined as the difference between 1 and the rate of precision, and the rate of silence is the difference between 1 and the rate of recall. Thus:

$$\text{Noise} : 1 - \frac{N1}{N1 + N3} \qquad \text{Silence} : 1 - \frac{N1}{N1 + N2}$$

Precision and recall are known to be asymmetric measures. We therefore decided to establish a composite measure, the f-measure[6] which will enable us to establish an average accuracy measure. There again the f-measure is represented by values going from 0 to 1. The closer the value is to 1, the better is the performance of the system being tested. On this basis , the acceptability threshold was fixed to 0.50. This gives us the following rating scale.

Definition	**Points**	**Rating**
f-measure between 0.80 and 1	50	Good
f-measure between 0.50 and 0.80	25	Acceptable
f-measure less than 0.50	0	Unacceptable

[5] In practice , the ideal is rarely, if eve r, attained.
[6] More detailed description of the f-measure can be found in [Van Rijsbergen, 1979:129].

Interoperability: Any terminology extraction tool must be compatible with the translation technology tools already available or announced for the near future. In concrete terms, a terminology extractor should be interoperable with systems that allow for the creation and management of terminology bases or which use terminology bases. Thus, it ought to be possible to store the results of the extraction in a format that can be used by the other kinds of applications found in translators' workbench systems, typically translation memory systems and terminology management systems. Two formats seem obvious candidates: the MARTIF format, designed as a terminology exchange format, or .txt, a format compatible with a large number of translation technology applications and also with text processing systems. This leads to the rating scale below:

Definition	Points	Rating
Export format .txt and MARTIF	10	Good
.txt or MARTIF	5	Acceptable
Neither MARTIF nor .txt.	0	Unacceptable

Recoverability: The programs to be evaluated should allow the user to recover all or most of the data he has been working on if there is a failure. We took this to mean that the system should include automatic saving of partial results at regular intervals. This gives the following rating scale:

Definition	Points	Rating
Automatic save function	10	Good
No automatic save	0	Unacceptable

Ease of learning: The ISO text makes it clear that this characteristic is a function of the documentation supplied and the first impressions formed by the user of the software. A first question is to ask what documentation is available. Here, we have chosen to take into account the existence of written documentation, interactive help, tutorial, and of a user group and discussion forum. Even though such groups are rarely created through an initiative of the manufacturer, their existence clearly helps with both learning, understanding and using a piece of software. These considerations lead to the scoring scheme:

Definition	Points	Rating
Existence of written documentation, interactive help, tutorial, user group and discussion forum	2.5	Good
Existence of written documentation, interactive help, tutorial	1.25	Acceptable
No documentation available	0	Unacceptable

The existence of help tools is obviously primordial, but these tools must also be of good quality if they are to be usable. In order to judge completeness and clarity of the documentation provided with the terminology extraction programs, we checked whether each functionality used in extraction and validation of terms had an entry in the on-line help or in the documentation available. If all are represented, the score is 100%. It diminishes in proportion to the number of commands for which no aid is given. These considerations lead to the scoring scheme.

Definition	Points	Rating
100% 50% Less than 50%	2.5 1.25 0	Good Acceptable Unacceptable

Ease of understanding: The ISO document makes it clear that this sub-characteristic concerns those attributes of the software which aim at making it attractive to the user. We need to find out, then, if the general presentation of the software is attractive, if the terms used in the program in menus or in the description of commands are appropriate, if the messages produces by the program are easy to read and if clear warnings and confirmations are presented, for example in order to avoid deleting or destruction of data. It has to be admitted that it is difficult to evaluate this sub-characteristic objectively. We have deliberately chosen therefore to group all the different elements into one, which we have called "user-friendliness", and have asked for a general appreciation, simply mentioning where necessary any details that ask for comment. We therefore established the following rating scale:

Definition	Points	Rating
User friendly Not very user friendly Not user friendly	5 2.5 0	Good Acceptable Unacceptable

Efficiency: A simple way of assessing the efficiency of a program would be to calculate whether it takes more or less time to use a terminology extractor over a corpus and validate the results, or to search the corpus manually for terms and record the results. In the majority of cases, using a terminology extraction tool implies extraction of terms by the program and validation of the candidate terms by a human. It is important then that the human validation process be taken into account when assessing effectiveness. The form in which the results are presented, as well as the supplementary information such as information on context attached to each candidate will affect how long the validation process takes. It goes without saying that these sub-characteristics of efficiency are affected not only by the usability considerations discussed earlier but also by the accuracy of the results, as discussed under functionality. Therefore, we decided to measure the efficiency of the products evaluated on the basis of the number of points obtained by each product under the accuracy and the usability section. These considerations lead to the following scoring scheme.

Definition	Points	Rating
Maximum points scored (60) Between 30 and 55 Less than 30	10 5 0	Good Acceptable Unacceptable

Results and Comments

Since we have adopted a comparative approach, the presentation and analysis of the results will take the form of tables containing the data collected during the execution of the evaluation, together with graphics summarizing the behaviour of the three products, organized in such a way that comparison between them is facilitated. On this basis, the advantages and disadvantages of each can be identified.

Functionality/Suitability

It is perhaps worth reminding the reader that we decided to split the characteristic of functionality into four attributes: Presence of all the functions described in the documentation, Internal consistency and consistency with the documentation, File formats supported, Languages supported. In order to assess the presence or absence of all the functions described in the documentation, we simply took the documentation provided for each of the three tools and looked to see if all the functions described in the documentation were present in the software. For Xerox XTS and ExtraTerm this was the case. MultiTrans presented one case where documentation and software did not correspond. The documentation states that at the end of the process of creating a project during which terminology extraction is carried out, the user should click on the Statistics button if he wishes to see or to print information on his new project. We were unable to find any button labelled Statistics, and we were not able to use the Print command. We have counted this as one functionality missing from the program.

	Points	**Rating**
ExtraTerm	5	Good
MultiTrans	2.5	Acceptable
Xerox	5	Good

For the attribute relating to internal consistency and consistency with the documentation, we looked at terminology use in the programs and in the documentation. Xerox XTS and Extra Term were very coherent. The problem described in the last section with MultiTrans reappear also here. In fact, once the process of project creation has been finished, a dialogue box opens, reporting that the files have been successfully imported and that the user can either display statistics on the TransCorpora files by clicking on the Details button, or immediately open the TransCorpora file in TransCorpora Search by clicking on the Open button. As we saw above, the User's Guide talks about a Statistics button in this context. There is thus inconsistency, since the dialogue displayed by the program talks of a Details button (correctly) where the documentation talks (incorrectly) of a Statistics button.

	Points	**Rating**
ExtraTerm	5	Good
MultiTrans	2.5	Acceptable
Xerox	5	Good

In the earlier discussion concerning file formats supported, we pointed out that the more formats a software can deal with, the more easily can it adapt to different work situations. We checked the file formats supported by each tool. ExtraTerm is an easy winner here: it supports a wide range of formats. At the other end of the scale, Xerox XTS which can only deal with *.rtf, .txt* and *sgml* is the weakest. MultiTrans does not, numerically, support many more formats than does Xerox XTS, but the formats are those we have considered to be standard formats, and therefore more rewarded in the scoring. It should be noted that none of the three could support *.pdf* files at the time of the evaluation.

	Points	**Rating**
ExtraTerm	5	Good
MultiTrans	2.5	Average
Xerox	1.25	Unacceptable

Concerning the languages supported, the reader will remember that rather than treat this attribute as a simple factual attribute whose value could be determined by checking the documentation, we interpreted it as meaning that the software should process the languages in question properly, and therefore decided to measure the attribute through the f-measure. The languages which interest us are English and French. We obtained the f-measure scores for each corpus and for each language. We then compared the f-measure scores for each language, taking the difference between the English and French f-measures for each corpus, and expressing the result as a percentage. If there are differences in performance across the two languages for any of the three tools, the difference is very small. However, it is interesting to note that none of the three systems produced exactly the same f-measure for each language and each corpus.

	Points	**Rating**
ExtraTerm	2.5	Average
MultiTrans	2.5	Average
Xerox	2.5	Average

Accuracy

The results presented here are based on comparing the lists of candidate terms extracted by the three tools and lists of terms manually extracted. In order to reduce subjectivity, a certain number of rules were followed during the manual extraction process. We considered uniterms, including acronyms, as well as complex words to be valid terminological resources. We also allowed expressions which were not strictly terms but which did present translation difficulties. We put no maximum length on the terms to be extracted, and specified no particular syntactic patterns. Below are some examples (not translationally equivalent) of what we considered to be valid elements for extraction for both languages:

English	***French***
electrode connector	amplification de crête de Raman
fluid administration catheter	analyse par parsage
scanning retinal laser	angle sous-tendu predetermine
ultrasonically driven pump	tannate de qualité pharmaceutique
optical signal handling devices	dispositif interface
wireless bi-directional interfaces	procédé de communication optique en espace libre

It should also be mentioned that when checking the results of the programs, we considered that if a complex term was only partially extracted, we would still count it as valid extraction since we consider that the user can complete it manually by looking at its context. The table below shows the total number of extracted terms for each program.

	A61 Eng	**A61 Fr**	**H04 Eng**	**H04 Fr**
ExtraTerm	1471	1676	1575	1822
MultiTrans	1293	1917	1293	2409
Xerox	3657	3860	4036	4325

The table below summarizes the results for each system. K stands for the number of terms manually extracted, P for precision, R for recall and f-m for the f-measure.

	A61 Eng				A61 Fr				H04Eng				H04 Fr			
	E	P	R	f-m	K	P	R	f-m	K	P	R	f-m	K	P	R	f-m
ExtraTerm	1935	0.24	0.19	0.21	1886	0.27	0.23	0.25	1987	0.26	0.21	0.23	1994	0.27	0.24	0.26
MultiTrans	1935	0.37	0.25	0.30	1886	0.30	0.29	0.29	1987	0.30	0.27	0.28	1994	0.26	0.32	0.29
Xerox	1935	0.42	0.79	0.55	1886	0.40	0.80	0.54	1987	0.39	0.79	0.52	1994	0.38	0.81	0.51

The table below summarizes the results in terms of noise and silence for each system. N stands for noise, S for silence.

	A61Eng		A61 Fr		H04 Eng		H04 Fr	
	N	S	N	S	N	S	N	S
ExtraTerm	81%	81%	73%	77%	74%	79%	73%	76%
MultiTrans	75%	75%	70%	71%	70%	73%	74%	68%
Xerox	58%	21%	60%	20%	61%	21%	62%	19%

As a first remark, it can be said that the three tools evaluated differed considerably in their scores for this attribute, both in terms of the number of candidate terms extracted and in their validity. As a general rule, Xerox XTS extracted about 2.5 times as many candidates as ExtraTerm, and about 3 times as many as MultiTrans. If we compare extraction across the two corpora, MultiTrans and ExtraTerm extract more or less the same number of candidates from each corpus. Xerox XTS extracts 400 more candidates from one corpus than from the other. Comparing the lists of automatically extracted candidates with the reference lists shows that the global results prove to be quite disappointing. Xerox XTS is a clear winner in terms of recall. Precision is correspondingly good. The other two tools score well neither for precision nor for recall. On the whole, the recall and precision results do not differ greatly across the two different corpora, except in the case of the precision scores for MultiTrans and Xerox XTS. This may be explained by the fact that the second corpus contains many cases where reference numbers occur inside a complex nominal group which would be a good candidate term; these numbers may well perturb the extraction process. However, ExtraTerm seems not to be affected, and behaves similarly across both corpora, being quite unable to extract complex terms. As we noticed in a previous section the scores differ from one language to another although in a small proportion. ExtraTerm gets scores better for extraction from English texts, both in terms of recall and of precision. MultiTrans gets better scores for recall in English than in French, but for precision does better in French. On that particular point, it should be noticed that MultiTrans extracted numerous verbs and adjectives in French. Xerox showed similar results in French and in English. In terms of noise and silence, again Xerox is a clear winner, being less noisy and silent than the two other tools which show similar bad results. This leads us to the conclusion that ExtraTerm and MultiTrans tend to extract a huge amount of non valid terms and at the same time leave valid terms not extracted.

	Points	Rating
ExtraTerm	0	Unacceptable
MultiTrans	0	Unacceptable
Xerox	25	Acceptable

Interoperability

We retained only one attribute as relevant to the assessment of interoperability. None of the three tools tested offers exportation in MARTIF format, but all allow exportation as .txt files. Xerox XTS adds the extra possibility of presenting the information in columns, which is quite

interesting for the user who wants to export not only the terms found but also their context. Given that ExtraTerm is part of the Trados suite of translators' aids, ExtraTerm's list of terms can imported in .mtw format, the format of MultiTerm, the terminology management tool of the suite.

	Points	**Rating**
ExtraTerm	2.5	Acceptable
MultiTrans	2.5	Acceptable
Xerox	2.5	Acceptable

Recoverability

The reader will remember that we defined recoverability as the inclusion within the tool of a an automatic save feature, operating during the validation process and preventing the loss of previous work in case of system failure. None of the three systems offers such a feature: the user may all or part of his work.

	Points	**Rating**
ExtraTerm	0	Unacceptable
MultiTrans	0	Unacceptable
Xerox	0	Unacceptable

Usability

Usability concerns both ease of understanding and ease of learning. Apart from attributes whose values depend on purely factual matters, such as the existence of various kinds of documentation, judgements of usability tend to be unavoidably subjective, and therefore may differ according to who carries out the assessment, according to that person's experience, ease with computing and expectations. All three tools provide basic necessary documentation. MultiTrans adds an extra source of information in the form of a tutorial which allows the user to get his hands on to the software quickly. This seems to us a strong advantage in view of translators' demands reported earlier. Trados is the only manufacturer for whose products a users' group and discussion forum exists. Although the initiative for the creation of these groups came from outside the manufacturer's company, their existence is an undeniable advantage, since they allow users to help themselves by sharing information and help about program use. Xerox scores less well than the other two tools, and has neither tutorial nor user group.

	Points	**Rating**
ExtraTerm	2.5	Good
MultiTrans	1.25	Acceptable
Xerox	1.25	Acceptable

In order to assess completeness of the documentation, we simply referred to the documentation systematically as we carried out the processes of extraction and validation for our two corpora. All the commands we used during these two processes had corresponding entries in the on-line help and in the written documentation for all three systems. We noticed no lack of clarity. Indeed, we felt that for all three programs the quality of the documentation was good.

	Points	Rating
ExtraTerm	2.5	Good
MultiTrans	2.5	Good
Xerox	2.5	Good

As far as user friendliness is concerned, we thought MultiTrans to be the least user friendly of our three tools. It would be only fair to point out that the extraction carried out by MultiTrans is only a part of the more general process of creating and indexing a corpus of texts. Thus it is perhaps a little unfair to complain that no specific interface is provided for managing the candidates extracted. However, a positive aspect is that candidates can be transferred into the terminology base by a simple click. Xerox offers a more complex interface, with many possibilities for displaying the data. The only negative point is that it is not possible, during validation, to eliminate a maximum of invalid candidates in a minimum of time, which is something of a handicap given the amount of noise generated by the program. The same remark applies to ExtraTerm, which has a pleasant interface but makes no sufficient use of keyboard short cuts to speed up the validation process.

	Points	Rating
ExtraTerm	2.5	Acceptable
MultiTrans	1.25	Unacceptable
Xerox	2.5	Acceptable

Efficiency

In assessing efficiency, we have made use of the results obtained for functionality and for ease of use.

	Points	Rating
ExtraTerm	0	Unacceptable
MultiTrans	0	Unacceptable
Xerox	5	Acceptable

The final step of our evaluation is to assemble the results. The table below summarizes the global results for each system. The maximum points available for each characteristics is shown at the end of the table.

	Functionality	Reliability	Usability	Efficiency	
ExtraTerm	17.5	0	7.5	0	25
MultiTrans	10	0	5	0	25
Xerox XTS	38.75	0	6.25	5	50
	70	**10**	**10**	**10**	**100**

The most obvious comment to be made on that table is that the performance of the three extraction tools are globally disappointing. Except for Xerox which shows almost acceptable results (50 out of 100), the tools evaluated do not seem to represent a real gain in productivity for translators and terminology. Yet, the bad results in accuracy explain the poor results of both MultiTrans and Extraterm. Considering this characteristics as being of the utmost importance, we are forced to the conclusion that MultiTrans and Extraterm are inadequate in our specific context. If we have to buy a product, even though it is not very satisfactory, Xerox will be the product to buy.

Conclusion and further work

The results make it clear that terminology extractors could not yet be considered of real help to translators and terminologists, mainly due to poor accuracy results. This study was limited to monolingual terminology extraction mostly because at the time of the evaluation not all the tools were able to work on a bilingual basis. It would be interesting to expand our evaluation to bilingual extraction. This will probably be undertaken in the next phase of our work. Another part of this work would probably be to refine some of the attributes and metrics used for efficiency. In that context, it would be interesting to carry out an *in extenso* experiment involving terminologists or translators effectively measuring the time spent on a manual extraction in comparison with an automatic one. However, we think that the attributes and the metrics developed in our evaluation are be of potential use for the evaluation of other extraction tools .

Bibliography

Ahmad K. and Rogers, M. (2001): «Corpus Linguistics and Terminology Extraction» in *Handbook of Terminology Management,* Vol. 2, Application-Oriented Terminology Management, John Benjamins Publishing Company, Amsterdam/Philadelphia.

Blank, I. (1998): «Terminology extraction from parallel technical texts». To appear in J. Véronis (ed.): Parallel text processing. Dordrecht: Kluwer Academic Publishers..

Compagnion, H. (1996): Les correcteurs orthographiques: caractéristiques, mesures et méthodes. Version électronique. http://www.osil.ch/eval/

Daille, B. (1994): *Approche mixte pour l'extraction automatique de terminologie: statistique lexicale et filtres linguistiques,* Thèse de Doctorat en Informatique Fondamentale. Université Paris VII.

Dias, G.; Guillore, S.; Bassano, J.C.; Pereira Lopez, J.G. (2000): «Extraction automatique d'unités lexicales complexes : un enjeu fondamental pour la recherché documentaire», TAL, Vol 41 N°2, p.447-472.

EAGLES Work Group (1999): EAGLES Evaluation Working Group, Final Report, n° EAG-II-EWG-PR.l (Draft), Center for Sproktechnologi, Copenhague.

Estopà, R.; Vivaldi, J.; Cabré, M.T. (1998): Sistemes d'extracció automàtica de (candidats a) termes: Estat de la qüestió, Instituto Pompeu Fabra, Institut Universitari de Lingüistica Aplicada, Barcelona.

Gamper, J.; Stock, O. (1999): «Corpus-based Terminology», in Terminology, vol. 5(2), John Benjamins Publishing Co, pp. 147-159.

GTW-Report (1996): Criteria for the Evaluation of Terminology Management software, Association for Terminology and Knowledge Transfer, Bolzano.

ISO/IEC 9126, First Edition (1991), Information technology - Software Product Evaluation - Quality Characteristics and guidelines for their use, Genève. International Organization for Standardization International Electrotechnical Commission.

ISO/IEC 14598-1 (1998), Information Technology - Evaluation of Software Products – Part 1: General guide, Genève. International Organization for Standardization International Electrotechnical Commission.

ISO/IEC 9126-1 (2001), Software engineering, product quality - Part. 1: Quality model, Genève. International Organization for Standardization International Electrotechnical Commission.

Jacquemin, C., (2001): *Spotting and Discovering Terms through Natural Language Processing,* The MIT Press, Cambridge, Massachusetts, London, England.

Justeson, J.S; Katz, S.M. (1995): ^Technical terminology: some linguistic properties and an algorithm for identification in text», in Natural Language Engineering, 1(1) Cambridge University press, pp. 9-27.

King, M. (1999): The 7-step Recipe for Evaluation Language Technology. In Proceedings of the European Evaluation of Language Systems (EELS) Conference, Hoevelaken, April 1999.

L'Homme, M.C.; Benali, L.; Bertrand, C. and Laudique P. (1996): «Definition of an evaluation grid for term-extraction software», in Terminology, Vol.3 (2), John Benjamins Publishing Co, pp. 291-312.

Pearson, J. (1998): *Terms in context, Studies in corpus linguistics,* John Benjamins Publishing Company, Amsterdam, Philadelphia.

Sparck Jones, K.; Galliers, J.R. (1993): Evaluating Natural language Processing Systems, Technical Report n°291, Computer Laboratory, University of Cambridge, England.

Van Rijsbergen C. J. (1979): Information Retrieval, 2nd edition, London, Butterworths.