Terminological Enrichment for non-Interactive MT Evaluation

Marianne Dabbadie, Widad Mustafa El Hadi, Ismail Timimi
marianne.dabbadie@lingpro.org
mustafa@univ-lille3.fr
timimi@univ-lille3.fr

Abstract

In a previous study (Dabbadie, Mustafa, Timimi, 2001) we set a methodology for non interactive machine translation evaluation on big corpora, assuming that the goal of the translation was a simple understanding of the original message. The source text, in French, provided by INRA (Institut National pour la Recherche Agronomique i.e. National Institute for Agronomic Research) deals with biotechnology and animal reproduction. It has been translated into English by REVERSO. The output of the system (i.e. the result of the assembling of several components), as opposed to its individual modules or specific components (i.e. analysis, generation, grammar, lexicon, core, etc.), has been evaluated.

In the present study we will recall the methodology and results obtained in the case of simple translation by REVERSO with no terminological enrichment and compare them to the results obtained after terminological enrichment. The aim of this study is to evaluate the impact of specific terminology when integrated to an MT System and after having run the system with a basic bilingual dictionary.

Key-words

Black-box evaluation, Lexical Fidelity, Syntactic Fidelity, Non interactive MT evaluation, terminological enrichment,user needs

1 Problem Overview

In a previous study (Dabbadie, Mustafa, Timimi, 2001) we set a methodology for non interactive machine translation evaluation on big corpora, assuming that the goal of the translation was a simple understanding of the original message. The source text, in French, provided by INRA (Institut National pour la Recherche Agronomique i.e. National Institute for Agronomic Research) deals with biotechnology and animal reproduction. It has been translated into English by REVERSO. The output of the system (i.e. the result of the assembling of several components), as opposed to its individual modules or specific components (i.e. analysis, generation, grammar, lexicon, core, etc.), has been evaluated.

In the present study we will recall the methodology and results obtained in the case of simple translation by the MT System with no terminological enrichment and compare them to the results obtained after terminological enrichment. The aim of this study is to evaluate the impact of specific terminology when integrated to an MT System and after having run the system with a basic bilingual dictionary. We have used the indexes provided by the INRA to create a specific dictionary in order to evaluate the impact of specific terminology when integrated to an MT System and after having run the system with a basic bilingual dictionary. These results give us comparative data to evaluate the impact of the addition of a domain specific dictionary to an MT system and in particular, the influence of specific terminology, over the total quality of the translated output. The result of this work constitutes the second and major part of the study. Moreover, the evaluation will concentrate on translation quality and its fidelity to the source text. The evaluation is not comparative (i.e. horizontal), which means that we tested a specific MT system, not necessarily representative of other MT systems that can be found on the market. The system tested is REVERSO.

We did carry out some manual testing but with the objective of setting a rough methodology that may reveal in most cases non relevant translations on big corpora. This evaluation has been done manually on a small corpus but the methodology designed for this test is supposedly applicable to a larger corpus provided that the test is automated.

To carry out this work in rational conditions there was a need for:

(a) linguistic resources

(b) a set of procedures for screening the text through

(c) an MT System for output display

According to the ISLE classification, the declarative evaluation on an MT system aims at measuring the ability of the system "to handle texts representative of an actual end-user". Moreover, it generally tests "for the functionality attributes of intelligibility (how fluent or understandable it

appears to be) and fidelity (the accurateness and completeness of the information conveyed)".

These criteria (i.e. intelligibility and fidelity) precisely fall within the scope of the present work. Therefore we will measure syntactic and lexical fidelity of the target text. The two separate scores thus obtained will give the total score for the intelligibility of the translated text. We will then analyse comparative data of the results obtained before and after terminological enrichment of the MT system by way of specific terminology addition.

2 Types of Analysis and Metrics

Within the framework of our previous study, we created a set of metrics to evaluate MT System syntactic and lexical correction rates. Considering that this is also a manual study on a small corpus we decided to provide an exhaustive error analysis of non parallel data.

MT softwares can be classified according to whether they are based on resources of a linguistic or statistical nature. These systems normally share the following sets of features:

(i) Segmentation, a step which is usually considered as part of preprocessing operations on a text. It consists of two sets of operations:

(a) Dividing the text into separate sentences (paying special attention to the identification of typographical symbols and abbreviations, ..);

(b) Dividing the sentences into words (paying special attention to the processing of blanks, hyphens and so on);

(ii) morphological analysis (part-of-speech tagging);

(iii) syntactic analysis, taking into consideration word-category disambiguation, identification of noun-phrases and their functions;

(iv) unit extraction: category patterns; search and retrieval strategies for pattern extraction (domain specific terms and named entities);

(v) lexical analysis.

We are detailing the various types of analysis in the following sections, adopting a black-box evaluation methodology.

2.1. Syntactic Analysis

We chose to count the number of *NPs* (noun phrases) and *VPs* (verb phrases) in source text and target texts, a first indication being given by non parallel data. *NP* is used in this paper to refer to both lexical NPs and non-lexical NPs (cf Dabbadie, Mustafa, Timimi, 2001). Obviously, a translation made by a non interactive MT System that does not include any domain specific dictionary most of the time tends to provide a word to word translation. Therefore, on big corpora a sensitive difference in terms of quantity of NPs and VPs in source and target texts may then possibly reveal a wrong translation. A threshold could be fixed in an automated procedure including the use of a previously tested and reliable bilingual syntactic parser that would generate an output file providing NPs and VPs count. The use of finer grained criteria such as a count of adjectives or prepositional phrases could also be envisaged. Any overlap of this threshold might then be considered as an indication that MT system may have failed to analyze source syntactic structure and that therefore, these figures require further analysis. A methodology including a test tool that would implement source and target transfer rules might probably prove even more accurate. For the purposes of this study we used the LATL[1] bilingual syntactic parser[2] with a manual check and correction of errors. The metrics used to measure correction rate are detailed in the following subsection.

2.2. Syntactic Fidelity

To obtain a success rate we worked out the following rates:

1-(Number of target NP – source NPs) / Number of source NPs

And

1-(Number of target VP – source VPs) / Number of source VPs

Total Correction rate : (NP correction rate + VP correction rate) / 2.

2.3. Lexical Analysis

Checking lexical correctness includes the following subtasks:

?? Polysemous words resolution: this is to check whether the system suggests the right target equivalent for a sense unit;

?? Segmentation problems;

?? Fluency problems (non idiomatic expressions – A detailed analysis is provided below in 3.2. but no numeric data will be given because we assume that MT goal in our study is limited to information).

---

[1] Laboratoire d'Analyse et de Traitement du Langage, University of Geneva.

[2] Syntactic analysis is one of the major components of a translation-oriented NLP which first applications began with MT. Analyses within the framework of an MT task can be seen as many sub-tasks which sum up the different relevant linguistic levels: morphological analysis, syntactic analysis (identifying noun and verb phrases and their functions) and finally, semantic analysis. Each of these sub-tasks can be in turn broken into smaller tasks: we can distinguish a) segmentation (identifying the word frontiers); b) lemmatization; c) tagging (identifying morpho-syntactic categories of each form), Abeille *et ali*. 2000.

?? Domain specific terminology or lexical- noun phrases (NPs).

Let us assume that to one source meaning should correspond one target meaning (which is not linked to the number of words actually present in the text). A count of "meaning units" which can either be single words or collocations with several levels of granularity has been done on the corpus. The lexical evaluation has been done manually for the purposes of this study.

2.4. Lexical Fidelity

Let us assume that the intelligibility criterion includes the characteristics of the translation process, the output characteristics, the quality of the translation, and the quality of the target text as a whole. Our point of view is that the fidelity criterion tends to answer the following question : Is the text understandable ? Let us assume that to one source meaning should correspond one and only one target meaning This allows us therefore to create a bijective relation between source and target sense units and to set a metric for fidelity that can be based on a count of the number of lexical units in the source text, as a referential figure. Success rate, precision and recall measures can then be worked out on target text.

After the syntactic tagging of source text, to obtain the number of sense units in source and target texts we applied the following metrics:

N° of words in text – N° of Determiners - N° of prepositions – N° of Coordination conjunctions.

To obtain a success rate:

(N° source sense units – total N° of wrongly translated sense units) / N° of source sense units

Total number of wrongly translated sense units = number of incorrect translations + unknown words + incorrect suggestions for polysemous word resolution.

We also calculated:

Lexical precision = number of relevant target sense unit / total number of target sense units

Lexical recall = number of relevant target sense units / total number of source sense units.

In order to work out the total quality of the output translation we set a final metric that gives in fact an average of correction rate and fidelity measures: intelligibility. The intelligibility metric can therefore be viewed as the quality of the translation as a whole. It may be worked out in the following way:

Intelligibility = average of correction rate + fidelity.

3 Manual Analysis of Output Errors

3.1. Syntactic Analysis

Whereas a gap between source and target NPs was noted in 30 % of the cases before terminolgical enrichment, this rate is reduced by 5% after terminological enrichment. In most cases the gap was due to unknown words which involve a wrong part-of-speech categorization. This is explained by the fact that unknown words, whatever part-of-speech they may belong to, are tagged as noun phrases. Before terminological enrichment, there were in fact 52 unknown words in target text, which was a great source of syntactic categorization errors and lowered the general quality of the output translation. The total number of unknown words after the creation of a specific dictionary on the background of the indexes provided by the INRA, fell to 11 occurrences.

In various cases, terminological enrichment has the effect of increasing the correctness of part –of-speech categorization on Noun Phrases. No particular impact was noted however on Verb Phrases. This is due to the fact that most of the unknown words were simple domain specific nouns and not verbs.

As a matter of example, in sentence n°5, *maturation cytoplasmique* had not been translated at all and was left in French in the target text, which had the effect of catogorizing the unknown words as two single Noun Phrases. After terminological enrichment, this sequence was translated by *cytoplasmic maturation*, *cytoplasmic* being therefore categorized as an adjective. After specific dictionary creation, this phenomenon was noticed on three other similar cases.

On the other hand, terminological enrichment has no impact on wrong part-of-speech categorization between source and target text, when due to wrong identification of a supposedly identified word : in sentence n° 10 for instance, the wrong categorization already noticed before the enrichment was reproduced after enrichment: *Les conditions de capacitation in vitro <u>diffèrent</u> selon les espèces (….).* <u>diffèrent</u> was not identified as a flexion of the French verb ***<u>différer</u>*** (which means to be different from) but as the French adjective ***<u>différent</u>*** (different). As a matter of consequence, the output translation is a verbless sentence: "*The conditions of capacitation in vitro <u>different</u> according to sorts (…)".

Another typical case of wrong syntactic categorization that can be solved by the addition of specific terminology and that due to a wrong interpretation of typographical conventions, is illustrated by the following example: sentence n° 4 was originally split into two separate sentences because of the wrong interpretation of the Roman numbering convention "II". Before dictionary creation, REVERSO had identified the two following sentences: *"The ovocyte which reaches the stage métaphase". "It in these conditions is not however competent to assure a conception and a normal embryonic development"*. After automatic processing of error correction REVERSO had assimilated the Roman number *II* to the English pronoun *Il*, which had been translated by *it* and had the effect of splitting the original

source sentence into two separate target sentences. After terminological enrichment, the inclusion to the specific dictionary of the bilingual equivalence *II=2* had the result of generating the gramatically correct following output: "*The ovocyte which reaches the stage métaphase 2 in these conditions is not however competent to assure a conception and a normal embryonic development*".

4 Lexical Analysis

Lexical analysis involves the following sub-sections: Granularity Levels: general language word level; polysemous word resolution; domain specific terminology and fluency problems.

These different levels of analysis can be illustrated by the following :

Two types of problems still remain:

1) Idiomatic expression chez + nom; chez + det+ nom. This expression is sometimes translated either by *to or by *at. Both are wrong translations. There are six cases where the translation is a word to word translation

a) Title: *Fécondation in vitro chez les ovins, caprins et équins* ✍ Conception in vitro *to ovine races, caprine and équins

b) Sentence n° 2: (…) *la fécondation in vitro chez les petits ruminants et les équins.* ✍ (…) the conception in vitro *at the small ruminants and the équins.

c) Sentence n° 10: *(…) sérum de brebis en chaleur inactivé chez le bélier (…..)* ✍ (…) of serum of ewe in heat inactivated *to the ram and the billy (..)

d) Sentence n° 15 : *Chez les ovins, après lavage (….)* .✍ *To ovine races, after wash (…)

e) Sentence n° 17 : *Avec la technique utilisée chez les ovins, (..)* ✍ With the technique used *to ovine races, (…).

f) Sentence n° 18: *Chez la jument, seuls les (….)* .✍ *To the mare, only ovocytes (…)

2) Remaining translation problems, i.e not solved by introducing domain-specific terminology and which have an impact on fluency. For the purpose of this article we mean by ***Fluency*** the capacity of the system to generate correct idiomatically formed expressions. We are limiting our examples to the good formation of domain specific terminology, mostly noun phrases. We noticed that a lot of translated English noun phrases contain prepositions (normally "of") however in English, empirically, only about 3% of terminological NPs contain prepositions[3] (most generally « of ») as shown in the examples hereafter[4]: "production in vitro" > *in vitro production*; "maturation of gametes" > *gametes maturation;* "transfer of embryos", > *embryo transfer*; "nuclear maturation in vitro"> *In vitro nuclear maturation*; "Maturation (cytoplasmique) of the ovocyte > *the ovocyte (cytoplasmique) Maturation*; delay of penetration of the ovocyte > *delay of the ovocyte penetration;* " The variability of the rate" > *the variability rate;* "The temperature of incubation" > *incubation temperature;* the rate of gestation is 50 % > *the gestation rate, etc.*

Eight cases can be reported for in the following sentences:

i) Sentence n° 1*: La production in vitro* (…) ✍ * production in vitro of fertilized (…) > **"invitro production"**[5]

ii) Sentence n° 3*:* (..) *est capable de reprendre spontanément* sa *maturation nucléaire in vitro*. ✍ (…) is capable of resuming spontaneously its nuclear maturation in vitro. > " **invitro nuclear maturation"**

iii) Sentence n° 9*: Les mécanismes de capacitation*, (..) ✍ * The mechanisms of capacitation, (…). > "**capaciation mechanism**" .

iv) Sentence n° 11: *L'efficacité du procédé de capacitation* (…)✍ *The efficiency of the process of capacitation (..) > **" capaciation process**".

v) Sentence n° 13: *La variabilité du taux de fécondation enregistrée* (…) ✍ *The variability of the rate of registered{*recorded*} conception (…) > **the variability of conception rate.**

vi) Sentence n° 14: *La température d'incubation* (….) ✍ *The temperature of incubation (..) > " **incubation temperature** ".

5 Domain Specific Terminology

This category comprises domain specific terms, be they heads or modifiers or compound terms. We added to *REVERSO*, the following words, or more exactly, simple terms (heads and modifiers) which are considered as domain specific terms (cf. INRA French Index): *capacitation, chromatine, cytoplasmique, micro-injection intra-cytoplasmique* (*cytoplasmique* is domain specific expression and part of a noun phrase acting as a modifier), *granulosa, polyspermie, transgenèse*. As a result the unknown words represent simple words or terms (heads and modifiers) that are not necessarily domain specific (cf. INA French Index): *décondensation, éjaculats, épididymaire, équins, ionophore, métaphase, oestradiol, oestrus , organelle*.

[3] This is not the case for French NPs, but since we are evaluating the English translation we chose to limit our description to English.

[5] The "expected" NPs translations are in bold character

5.1 The Impact of Adding Domain Specific Terminology

Apart from part-of-speech re-categorization already detailed in section 3.1, the most noticeable impact of the addition of domain specific terminology is on fidelity and intelligibility rates. Comparative results, before and after terminological enrichment are given in section n°6.

If we consider the translation provided by *REVERSO* after adding domain specific terminology it is obvious that we have a better understanding of the text. In other words, from the point of view of information access and extraction the user need might be satisfied. If we look at the results from the point of view of knowledge acquisition, the quality of the translation should be better and terms in English should be strictly equivalent to terms in French.

Polysemous word resolution: for polysemous words in our previous study (Dabbadie, Mustafa, Timimi, 2001) we stated that the MT System we used often suggested various equivalents but some of them were not suitable. This kind of problem still remains after terminological enrichment, because REVERSO still suggests several equivalents for polysemous words whatever the specific terminology added to the system6.

We will not propose further analysis for fluency problems. Regarding this point readers can refer to our previous study carried out on the same corpus (cf Dabbadie, Mustafa, Timimi, 2001).

6 Results - Numeric Data:

6.1 Syntactic Metrics[7]:

Source NPs	Target NPs	Source VPs	Target VPs
142	184/178	38	40

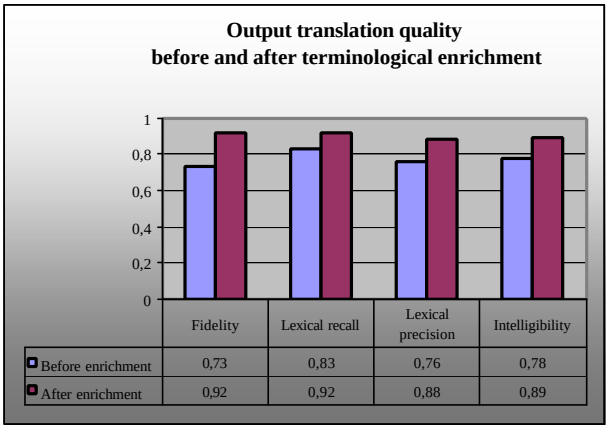
6.2 Correction Rate

NPs correction rate	VPs correction rate	MT System correction rate
0.70/ 75	0.95	0.83/0,85

6.3 Lexical Metrics[8]:

Number of words in source text	Number of words in the target text	Total unknown words	Polysemous word resolution Suggested
544	562	51/11	8 proposals
Correct/ suitable suggestions	**Number of incorrect translations**	**Number of source sense units**	**Number of target sense units**
1	21	302	322/317

The intelligibility figure, reveals that the translation is understandable in 89 % of the cases. It is important to note that an 11% rise in the intelligibility figure is due to terminological enrichment. The 19% rise in fidelity figure reveals the previous poor results (0,73% before terminological enrichment) mainly relied on terminological issues rather than on a wrong processing of syntactic units by REVERSO.



In our previous study, we pointed out the fact that the results obtained, together with the manual analysis of syntactic and lexical data led to think that unknown words are generally a great source of semantic errors and wrong syntactic categorization

The results obtained thanks to the addition of specific terminology, tend to confirm the intuitive impression that Machine Translation output can be optimized in a large part by the injection of domain specific terminology into a determined client application. Moreover, it tends to sustain the hypothesis that a system based on a terminology structured by domains, such as the dictionaries organized through a thesaurus or semantic network like structure, could very probably increase even more lexical relevance and therefore MT lexical fidelity figures and intelligibility.

---

[6] Here are a few examples quoted in the previous study that apply to polysemous word resolution: "For example: "Milieu" is translated by environment which is acceptable as a translation but the tool suggests another word *middle which is unsuitable in the context of sentence n° 7.

« La co-culture du complexe ovocyte-cumulus avec des cellules de la granulosa permet d'améliorer l'aptitude au développement des oeufs FIV dans un milieu supplémenté en FSH (…), »

"The co-culture of the complex ovocyte-cumulus with cells of the granulosa allows to improve capacity in the development of eggs FIV in an environment supplemented in FSH, (…)".

[7] Colums including two different figures refer to before and after terminological enrichment.

[8] idem

7 Further work

These two studies have led us to show how the implementation of numeric data may serve as a test tool to evaluate an MT system's syntactic and lexical fidelity on large corpora. Apart from giving a statistical overview of the general quality of the output of an MT system, these studies have also led us to prove that the addition of specific terminology to an MT environment tends to improve dramatically the quality of the output translation.

In these two studies, we chose to count the number of *NPs* (noun phrases) and *VPs* (verb phrases) in source text and target texts, a first indication being given by non parallel data. As already stated in this article, the use of finer grained criteria such as adjectives or prepositional phrases count could also be envisaged. Any overlap of this threshold might then be considered as an indication that MT system may have failed to analyze source syntactic structure and that therefore, these figures require further analysis. But this methodology is still imprecise and limited to a first indication of MT system's analysis failure, when a gap is observed on non parrallel data. The use of this methodology also implies that the test is carried out on relatively syntactically isomorphic languages such as French and English. A methodology including a test tool that would implement source and target transfer rules might probably prove more accurate and also apply to non isomorphic languages.

8 Perspectives on MT and NLP Software Evaluation

Although it is important to create tools in order to evaluate the output of an MT system, it is a generally well admitted fact that evaluation also applies to the determination of user satisfaction and needs and that Natural language applications testing is not limited to the quality of the output.

Coaching applications are conceived with the aim of observing the human machine interface processes in order to work out the logic of the usage made of a given application.

K-Now is a coaching application that has been conceived by KnowMore, a French startup, thanks to a programme co-financed by France Telecom and the French ANVAR (National Agency for the development of Research). It observes the human/machine interface in order to work out the usage made of the existing applications in a given environment. This application relies on the system's graphical interface (browser, virtual machine, operating system etc.) and requires no intrusion into the observed applications. It saves screen after screen in order to measure Information System usage along with user's level of appropriation of the system and requires no specific interface with the observed system. K-Now has a statistical diagnosis interface that produces trend graphics that give an overview of a system's use, extended to a company's intranet.

It would be interesting to consider the use made of a Natural Language Information Solution System on a large scale, through the analysis provided by a coaching application on an intranet environment. This kind of analysis does apply to MT. Although it is not relevant to test the output of an MT System in itself, it is important to be able to evaluate the use made of an MT system installed on an intranet, in order to evaluate the needs or reluctance of a given group of users to use the system and would give a statistical overview of the satisfaction rate of a group of users of a given Machine Translation system. Satisfaction diagnosis may for example be used in order to determine whether it is worth for a company to finance specific terminology customization in an extended technical domain. This kind of test could also be carried out in a presales environment on a preliminary test phase in order to determine whether a system installed on temporary licence terms matches the future user's needs and expectations.

9 Related work

Lewis, (2001: 207) presents an interesting approach to MT in order to produce more accurate, "more human" automatic translations[9]. Whilst specific products are discussed, the author believes that the methodology could be successfully implemented with different sets of tools. As the author pointed out, translation software "buyers will always prefer the "look and feel" of human translations", (Lewis, 2001: 207). The approach presented provides a way of increasing the "human look and feel" of automatically generated documents. Although in our case we only added domain specific terms, it would have been possible to add idiomatic expressions such as *chez* + noun; *chez* + det + noun to increase the accuracy and the "look and feel" of human translations. This expressions are very frequent in the text we submited as input to *REVERSO*. We think that introducing highly frequent idiomatic expressions is as

[9] The work takes a practical look at ways of combining language engineering tools to produce more accurate, automatic translations. The tools involved, as the author explains, are machine translation software, a translation memory application and alignment software, small tools or utilities written to perform simple yet very important tasks. The MT program discussed is the author's own Dutch-English translation software, which has been rewritten in Java. The translation memory software used is Trados Translator's Workbench and WinAlign. All the utilities were written in Java by the author. However, the paper is concerned with presenting an approach or methodology which could conceivably be implemented with a totally different set of tools Lewis, (2001: 207).

important as adding domain specific terminology. The only problem would be the selection criterion. In this current case (INRA text 603) we identified *chez* as a highly frequent sequence. In another text it would be another expression. We can conclude that the addition of this type of expressions to an MT software should be the rule if we desire to produce more accurate translations.

10 References

Abeille, A. Blache, Ph. (2000). Grammaires et analyseurs syntaxiques. In : Pierrel, J.-M. éds. (2000). *Ingénierie des langues*, Traité 1C2 - Section informatique et systèmes d'information, p. 51-76.

EAGLES (1999). EAGLES Reports (Expert Advisory Group on Language Engineering Standards)http://www.issco.unige.ch/projects/eagles/ewg99.

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

ISO (1999). Standard ISO/IEC 9126-1 Information Technology – Software Engineering – Quality characteristics and sub-characteristics. Software Quality Characteristics and Metrics - Part 1

ISO (1999). Standard ISO/IEC 9126-2 Information Technology – Software Engineering – Software products Quality : External Metrics - Part 2

ISSCO (2001) Machine Translation Evaluation : An Invitation to Get Your Hands Dirty!, ISSCO, University of Geneva, Workshop organized by M. King (ISSCO) & F. Reed, (Mitre Corporation), April 19-24 2001.

Justeson, J.S., Katz, S.M. (1995). "Technical Terminology: Some linguistic properties and an algorithm for identification in text", Natural Language Engineering 1(1), pp. 9-27.

Kilgarrif, A. (1998). "SENSVAL: An Exercise in Evaluating Word Sense Disambiguation Programs", in Proceedings. LREC, Granada, May 1998, pp. 581-588.

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

King, M. (1999b). "ISO Standards as a Point of Departure for EAGLES Work in EELS Conference (European Evaluation of Language Systems), 12-13 April 1999.

Lewis, H. (2001). Combining Tools to Improve Automatic Translation. In Proceedings of MT Summit VIII, Santiago de Compestalla, Spain, 18-22- September 2001, pp 207- 209.

Mustafa El Hadi, W., Timimi, I., Dabbadie, M. (2001). Setting a Methodology for Machine Translation Evaluation. In: Machine Translation Summit VIII, ISLE/EMTA, Santiago de Compestela, Spain, 18-23 October 2001, pp. 49-54.

Mustafa El Hadi, W. (1998). "Automatic Term Recognition & Extraction Tools: Examining the New Interfaces and their Effective Communication Role in LSP Discourse". In Mustafa El Hadi, Maniez, J. & Politt, S. éds Structures & Relations in Knowledge Organization, Proceedings of the Fifth International ISKO Conference, Lille, 25-29 Août 1998, pp. 204-212.

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

Véronis, J., Langlais, Ph. (2000). ARCADE: évaluation de systèmes d'alignement de textes multilingues. 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).