In Vitro Evaluation of a Program for Machine-Aided Indexing

Christian Jacquemin
LIMSI-CNRS & University Paris 11
BP 133
91403 ORSAY Cedex
FRANCE
Phone: (33) 1 69 85 80 22,
Fax: (33) 1 69 85 80 88
Christian.Jacquemin@limsi.fr

Béatrice Daille
IRIN - Université de Nantes
2, rue de la Houssinière
BP 92208
44322 Nantes Cedex 3, FRANCE
Phone: (33) 2 51 12 58 54,
Fax: (33) 2 51 12 58 12
daille@irin.univ-nantes.fr

Jean Royauté and Xavier Polanco
CNRS-INIST
Unité Recherche et Innovation
54514 Vandoeuvre-lès-Nancy
Phone: (33) 3 83 50 46 00,
Fax: (33) 3 83 50 46 50
{royaute,polanco}@inist.fr

* We would like in particular to thank all the experts, Laurence Arnaud, Bénédicte Dimbert and Dominique Vachez, who accepted to participate in this evaluation and have produced work of a very high quality. Laurence Arnaud and Bénédicte Dimbert died tragically in June 2000 in a traffic accident. They never saw the results of their work. We salute them posthumously and bitterly regret their untimely deaths. We thank Yannick Toussaint for heading the ILLIAD project, without which we could not have made this evaluation, Etienne Fleuret, former Director of INIST’s Database Technical
Department, and Vivianne Berthelier, head of the Biological Sciences and Agronomy section of that department, for allowing us to carry out this evaluation. From 1995 to 1998, ILIAD was sponsored by the Programme Sémantique and Mémoire Externe of GIS Science de la Cognition. We also thank the reviewers for their critics and suggestions, and Martine Hurault-Plantet (LIMSI) for her comments on a preliminary draft.
Abstract

This article presents the human evaluation of ILIAD, a program for Machine-Aided Indexing (MAI). It consists of two language engineering modules and is designed to assist expert librarians in computer-aided indexing and document analysis. Our aim is the expert evaluation of automatic multi-word term indexing. Evaluation is performed by documentary engineers. Cataloging and indexing are their principal tasks. They also have a good scientific knowledge of the domain to which the indexed documents belong.

We first present the ILIAD program and the two systems submitted to this evaluation, the methodology (protocol) adopted, the differences between the protocol and the implementation, and the results of these evaluations. Human evaluation is divided into three parts: firstly the evaluation of controlled indexing, then free indexing and finally term variant extraction performed during controlled indexing. Finally, we analyze the relevance of this evaluation by calculating the agreement frequency and the Kappa coefficient and propose some future developments.

Keywords:

human evaluation, machine-aided indexing, controlled indexing, free indexing, multi-word terms, key-phrase indexing.
1 Introduction

The purpose of this work is to compare the output of a program for automatic key-phrase indexing (multi-word units) with the indexes that would be chosen for human indexing. The program is used in two different modes: controlled indexing where indexes are occurrences or variants of terms from a controlled vocabulary and free indexing where indexes are chosen without consideration for an authority list. The indexes extracted automatically by the program are evaluated by expert librarians from Institut de l’Information Scientifique et Technique du Centre National de Recherche Scientifique (INIST-CNRS), a French center for the management of scientific and technical information management. Each librarian involved in the evaluation has a good scientific knowledge of the domain to which the indexed documents belong.

An early issue in automatic indexing was the designing of techniques that could index and retrieve documents with a quality at least as good as human indexing. In (Salton, 1969) the results of a search engine on the automatically indexed MEDLARS collection are compared with human indexing. Even in such early work, automatic procedures are shown to outperform human indexing. Simultaneously, various standards of evaluation are established that will serve as guidelines for subsequent work on the design and the validation of computerized systems for information access.

A third of a century later, accurate comparisons between human expertise in automatic indexing and the output of indexing programs are still an issue because of the recent introduction of sophisticated natural language processing (NLP) techniques in information retrieval (IR). We are concerned with the evaluation of the efficiency of such techniques through a detailed comparison of human indexing and phrase indexing using NLP components. Our purpose is to provide experts with the automatic indexing produced by ILIAD (whether free or controlled indexing) and to receive feedback from expert librarians on the relevance of the indexes. The evaluation presented in this paper is different from usual IR evaluations such as the ad-hoc track at TREC (Robertson
& Hancock-Beaulieu, 1992) because it is not geared towards the construction of a fully automatic IR system but towards machine-aided indexing (MAI), and more specifically, the assistance of an expert in a multi-word indexing task.

**Expert-Based Evaluation**

For the majority of activities relating to the automatic processing of language and information access, evaluation procedures are set up to compare and validate the techniques proposed by the programs for natural language information retrieval (Harman, 1995; Rubio, Gallardo, Castro, & Tejada, 1998). The standard evaluation paradigm involves automatically comparing the results given using a program with reference output defined either by specifications made beforehand or in the form of manually verified data.

Depending on the domain of activity, it is not always desirable to carry out a fully automatic evaluation. There are two major cases in which it is preferable to involve a human expert in the evaluation procedure:

- **Small ratio reference data/search space** — When the output can take a wide variety of forms, the search space is large and cannot be completely associated with a reference for the evaluation: it is not feasible to build a set of references since it would be too large or costly to construct. For example, the automatic extraction of alternative constructions from a large corpus (Lapata, 1999) cannot be evaluated using a set of references since the set of combinations of verbal predicates and arguments is enormous and only a few of them will ever appear in the corpus being studied. For this reason, the author will call in human adjudicators to verify the validity of the constructions extracted.

- **Over-complex decision** — If the decision as to the validity of the output requires a detailed analysis of the results and a high level of expertise, the evaluation cannot be automated since
the verification program will be at least as complex as the original program. We are in this second case. The aim of these experiments is to validate the output from a program for MAI focused on multi-word indexes. Due to the high degree of variability of multi-word terms is not possible to construct an evaluation program which would take into account all the criteria considered by an expert to decide on the quality of an index.

The main theme of this article is the implementation of a procedure for human evaluation of the indexes automatically extracted by the ILIAD program for MAI through linguistic engineering techniques (Toussaint et al., 1998).¹

Section 2 presents some background knowledge on the use of Natural Language systems and thesauri in information access. After briefly describing the data used for evaluation in Section 3, we specify the details and architecture of the program in Section 4 and the implementation of the evaluation in Section 5. We give the results of the evaluation in Section 6 and discuss their repercussions on future work combining MAI, NLP, and data mining in Section 7.

2 Natural Language and Concept-Based Information Retrieval

The use of multi-word terms or key-phrases for indexing and accessing documents allows the user to focus on accurate subjects and topics. However, the efficiency of multi-word term indexing remains a controversial issue in the IR community. Fagan (1987) claims that syntactically extracted key-phrases yield better results than statistical ones. However the use of phrases resulting from shallow parsing can turn out to be disappointing in usual IR tasks such as the ad hoc task at

¹This indexing program has been set up at INIST. INIST is a service unit of the CNRS (UPS 076), and is an integrated center for Scientific and Technical Information management (on-line data bases, supply of copies of articles, production of CD-ROMs, translations, etc.). INIST manages collections of documents covering the majority of international research. It produces two on-line bibliographic databases (multi-lingual and multi-disciplinary): PASCAL, for Science, Technology, and Medicine; FRANCIS, for the Humanities and Social Sciences.
TREC (Strzalkowski, 1995, 1999). One of the main reasons for disappointing results is the lack of relationships between queries and documents in cases of multi-word indexes because of the high degree of term variation.

The use of elaborate NLP techniques in IR is better suited for intelligent information access than document retrieval through natural language queries. In her introduction to the collective book on Natural Language Information Retrieval, Sparck Jones (1999) questions whether NLP can be used to improve information access:

*It is not clear, [either,] that NLP is required for some tasks that are closely related to ordinary retrieval.*

Sparck Jones concedes that there are three domains in which IR could benefit from natural language procedures: information extraction, automatic abstracting, and question-answering systems. This list is probably too restrictive, but it does outline the fact that it is necessary to address high level tasks in information access to take advantage of NLP techniques. For instance, collection browsing, a type of search more refined than query-based document retrieval, is reported by Gutwin, Paynter, Witten, Nevill-Manning, and Frank (1998) to be more effective using multi-word term indexing and searching than using single word access.

The use of key-phrases for indexing documents is a fully automatic means of information access (level 5 of MAI according to Hodge’s classification (1994)) and a computer-based support to intelligent activities (level 4 of the same classification). In this article, we evaluate the contribution of fully automatically generated key-phrases to assist expert librarians in their indexing tasks. In (Gutwin et al., 1998; Strzalkowski, 1995, 1999), the extraction of key-phrases is done by a fully automatic means without reference to an external source of conceptual knowledge such as a thesaurus or an authority
list. We call this \textit{free indexing} because of the lack of control on the extracted indexes. Conversely, \textit{controlled indexing} establishes conceptual links between documents and items in a list, a thesaurus, or a knowledge organization system. The links can be made automatically (as is the case in our study) or an expert librarian can use a MAI system as an aid for building and enriching an index list (Humphrey, 1994). Due to their specificities, we provide two different evaluation schemes for free and controlled indexing.

The analysis of the use of Knowledge Bases or Knowledge Organization Systems in Abstracting and Information Services (Hodge & Milstead, 1998) shows that

1. sources of knowledge are used effectively for information access and management in these industrial environments,
2. knowledge systems are applied to various tasks, such as clerical administration, access to related documents, or access to related knowledge bases, through different means, such as browsers, indexing post-processors, or fully automatic query expansion modules.

These authors conclude their study by noting that both trends in computer-based information access—computer-assisted human indexing through Knowledge Organization Systems or fully automatic indexing—should develop and cooperate in the future.

In brief, the future thesauri for MAI should include richer semantic information for efficient topic expansion (Hodge, 2000), should merge existing heterogenous knowledge sources into a single unified framework (Harpring, 1999), and should be appropriate for (semi-)automatic indexing (Milstead, 1997). A proper use of knowledge sources for indexing requires richer environments for browsing and enriching these databases and presupposes automatic procedures that are able to use the information they contain.

As for free indexing, the proper integration of a knowledge organization system into a MAI system through controlled indexing must take variation into consideration in order
to establish relationships between indexes and related concepts.

The ILIAD system presented and evaluated in this paper focuses on controlled and free indexing through NLP procedures with a constant concern for term variation, particularly for automatic term acquisition and indexing. In order to verify the usability of automatically extracted indexes in MAI, we now present a thorough evaluation of NLP-based indexes resulting from controlled and free indexing by expert librarians.

### 3 Conceptual and Textual Data

The task evaluated in this article is the automatic indexing of a large set of documents (the AGRO-ALIM corpus from PASCAL database) with the help of a set of controlled terms extracted from the AGROVOC thesaurus. We detail in turn these two sets of data.

1. **The AGROVOC Thesaurus**

   (AGROVOC, 1998) is a multi-lingual (French, English, Spanish) thesaurus intended for the indexing of data used in agricultural information systems. The version used contains 25,964 French terms (both preferred and synonymous).

2. **The AGRO-ALIM Corpus**

   The food corpus [AGRO-ALIM] contains 2,702 bibliographic records. These records are produced from articles by researchers and faculty members, engineers and technicians from agriculture and the food industry. We deleted the following information from these records: authors, affiliations, keywords, etc. Only titles and abstracts were retained. The [AGRO-ALIM] bibliographic references span the period from 1992 to 1998 with the following distribution: 185 for 1992, 295 for 1993, 320 for 1994, 453 for 1995, 499 for 1996, 770 for 1997 and 180 for 1998. [AGRO-ALIM] contains 427,482 words and the average size of a record is 316
The ILIAD linguistic engineering program (Toussaint et al., 1998) is an intelligent workstation prototype designed to assist expert librarians in indexing documents. In addition to the modules for automatic term extraction and MAI, the ILIAD program can represent a knowledge field in the form of maps of terms or documents. The program includes different modules which organize a set of documents and associated terms: modules for the automatic linguistic analysis of documents, for the acquisition of terminological knowledge, and for data mining.

The ILIAD program is now presented in two steps. Firstly, we give the general outline of the system and, secondly, we present in more detail the modules that are the focus of the evaluation: (1) controlled indexing through FASTR and (2) free indexing through ACABIT.

4.1 Overview

Schematically the program can be divided into three parts:

1. **Tokenization and Lemmatization**

   **Word Level.** In the first part of the processing, the corpus is split into sentences and the words are isolated and morphologically analyzed using the WinBrill tagger\(^2\) and the FLEMM morphological analyzer (Dal, Hathout, & Namer, 1999).

2. **Text Indexing**

   **Term Level.** The processing of terminological data is split between two modules,

\(^2\)http://jupiter.inalf.cnrs.fr/WinBrill/
ACABIT for the acquisition of new terms (Daille, 1996), and FASTR for automatic indexing (Jacquemin & Tzoukermann, 1999)\(^3\). FASTR either takes terms from a lexical terminology and performs controlled indexing, or it takes the terms acquired by ACABIT and then performs free indexing.

3. **Descriptor Extraction**

**Document Level.** The terms associated with documents which are identified in the previous step are used to group together documents in a conceptual manner and provide users with a synthetic representation of a knowledge field. At this step, terms are said to be descriptors because they represent the conceptual content of documents.

Two tasks are evaluated: controlled indexing and free indexing.

1. **Controlled Indexing**

Preferred and synonym terms from the thesaurus [AGROVOC] are used to build an authority list. The experts give an opinion on the relevance of the indexes automatically extracted by FASTR by using this list. Since it is necessary to take into account all the possible ways of expressing particular concepts (Salton, 1992), the terms are identified by FASTR either in the linguistic form in which they appear in the thesaurus or in a form representing a possible variation of the term. In both cases, only the linguistic form present in the thesaurus is shown to the expert.

2. **Free Indexing**

The experts must give an opinion on the ability of the terms proposed to represent conceptually the content of the document. Some of the terms proposed may well be present in the thesaurus [AGROVOC], no check is made in advance.

\(^3\)http://www.limsi.fr/Individu/jacquemi/FASTR/
4.2 Detailed Presentation of ILIAD

The two main modules of the program are ACABIT, a module for term acquisition (Daille, 1996), and FASTR, a module for automatic indexing (Jacquemin & Tzoukermann, 1999; Jacquemin, 2001). FASTR takes two types of terminological data as input: terms from the authority list in the case of controlled indexing, or terms acquired by ACABIT on the corpus to be indexed in the case of free indexing. The combination of these two modules is described in Figure 1. Both modules are fed with tokenized, tagged, and lemmatized corpora in which inflected words are associated with a unique lemma and a structure of morphological features. The structure contains features concerning the gender, number and tense of inflected words. In our experiments, the [AGRO-ALIM] corpus is divided into two parts: a training corpus of 416,231 words for defining the linguistic data and a test corpus of 11,251 words.

**ILIAD program (modules for free and controlled indexing)**

![Figure 1: The terminological part of the ILIAD program](image)
4.2.1 Automatic Term Acquisition through ACABIT

Although the set of recycled terms extracted from the thesaurus [AGROVOC] to build the authority list is highly reliable, it is not sufficient for complete and accurate conceptual indexing. Specific terms belonging to the sub-domain of the corpus are missing together with generic terms, i.e. not domain specific but meaningful terms. New terms have to be discovered. The retained approach to corpus-based term selection is an automated hybrid linguistic/statistical approach developed by (Daille, 1996).

The methodology is the following: tagged and lemmatized text is first passed through a series of linguistic filters to identify syntactically “term-like” strings, these are then submitted to a statistical test, the loglike (Dunning, 1993), in order to further refine the term selection process. Statistical modeling permits a level of abstraction in which all words in a corpus which are involved in a certain linguistic relationship can be automatically extracted without having to refer to the words themselves.

The linguistic filters consist of a series of local grammar rules to detect compound nominal expressions having a length of two, i.e. two content words. (Grammatical function words such as prepositions and determiners are permitted to intervene between the two content words.) The decision to concentrate on terms of two words is motivated by both linguistic and statistical reasons, and informational relevance.

Rules (2) and (3), expressed using regular expressions, recognize determined two-word term patterns: rule (1) detects adjectival phrases; rule (2) terms of N Adj structure such as alimentation hydrique (water supply); rule (3) terms of N1 Prep N2 structure such as eau du sol (ground water).

\[
\begin{align*}
(1) & \quad \text{adjp} \quad (\text{adv} " "\text{adj})+(\text{adj}\text{pastp})" "+) \\
(2) & \quad \text{nadj} \quad \text{noun}" \text{adjp} \\
(3) & \quad \text{nprepn} \quad (\text{noun}|\text{nadj})" \text{prep}" \text{noun}
\end{align*}
\]
The module for automatic term acquisition takes into account term variations, or at least some of them. To illustrate how this is done, let us examine rule (4) which is able to recognize a nominal argument coordination of two terms sharing a N1 Prep N2 structure:

(4) nprep_ncoordn nprepn" "coo" "prep" "noun

Rule (4) identifies two terms of length two, namely *système de surface* (surface system) and *système de profondeur* (depth system) with a coordinated sequence like *système racinaire de surface et de profondeur* (surface and depth of the rooting system).

The resulting terms of length two are then sorted by statistical criteria which are designed to separate sequences which have a terminological status and from those which do not. For each sentence of our corpus, we keep only two candidate terms, the ones sharing the highest score of the statistical criterion as shown in table 1 for the sentence:

(S) *La part respective des deux portions du système racinaire de surface et de profondeur dans l'alimentation hydrique d'un arbre varie en fonction de la demande atmosphérique et de la disponibilité de l’eau du sol.*

(The respective part of the surface and depth of the rooting system in the water supply of the tree varies according to the atmospheric need and ground water availability.)

The precision of the term extraction performed by ACABIT is evaluated using a set of 300 randomly chosen sentences. Table 2 shows that 96% of the retained candidate terms are well-formed candidate terms.
<table>
<thead>
<tr>
<th>Candidate terms</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>système racinaire</td>
<td>174.819</td>
</tr>
<tr>
<td>eau sol</td>
<td>79.1731</td>
</tr>
<tr>
<td>alimentation hydrique</td>
<td>68.6754</td>
</tr>
<tr>
<td>disponibilité eau</td>
<td>63.6223</td>
</tr>
<tr>
<td>fonction demande</td>
<td>6.47268</td>
</tr>
<tr>
<td>part respective</td>
<td>3.13626</td>
</tr>
<tr>
<td>alimentation arbre</td>
<td>2.29382</td>
</tr>
<tr>
<td>système profondeur</td>
<td>1.70805</td>
</tr>
<tr>
<td>système surface</td>
<td>0.00165</td>
</tr>
</tbody>
</table>

Table 1: ACABIT’s output on sentence (S)

<table>
<thead>
<tr>
<th>CT</th>
<th>Correct CT</th>
<th>Incorrect CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>600</td>
<td>575 (96%)</td>
<td>25 (4%)</td>
</tr>
</tbody>
</table>

Table 2: Precision of candidate terms (CT) retained per sentence

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun part of compound preposition</td>
<td>12 (48%)</td>
</tr>
<tr>
<td>Wrong tagging</td>
<td>7 (28%)</td>
</tr>
<tr>
<td>Wrong adjective attachment</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>Quantifier nouns</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>2 (8%)</td>
</tr>
</tbody>
</table>
4.2.2 Controlled and Free Indexing Through FASTR

As shown in Figure 1, the role of FASTR (Jacquemin & Tzoukermann, 1999) in the ILIAD architecture is to index the documents by automatically extracting occurrences of terms and term variants. The processing of FASTR relies on three types of data:

1. morphologically analyzed corpora where words have a unique lemma and a set of morphological features such as tense, number, and gender,

2. lemmatized term lists where words have a unique lemma and a unique syntactic category. These terms result either from the pre-processing of an authority list through a part-of-speech tagger trained for noun phrases, or from the output of ACABIT, the acquisition module presented in Section 4.2.1,

3. a small metagrammar of the language being studied (≈ 100 rules).

In this section, we outline the role of the metagrammar that implements local morphosyntactic transformations representing term variations. The metarules transform terms into term variant patterns in order to retrieve term variant occurrences from the processed corpora. Table 3 illustrates the output of FASTR when indexing sentence (S) with the terms acquired by ACABIT on this corpus (free indexing). Occurrences (3.c) and (3.g) correspond to the two terms proposed as candidate terms by ACABIT on this sentence. The other indexes are occurrences or variants of terms acquired elsewhere in the corpus.

FASTR is based on a metagrammar of the language in which indexing is carried out. The output of the parser is a set of links to the terms of the authority list. Each link in Table 3 is labelled with a type in the fourth column: 0 means that the utterance is a plain term occurrence, all the other symbols denote different families of term variations.
The example detailed in Table 3 highlights two families of variants that are currently accounted for by FASTR:

- **Syntactic variants**
  
  These variants involve a modification of the syntactic structure of the terms without morphological transformations (apart from inflections). Thus, variant (3.b), called a Modification, is the insertion of an adjectival modifier after the head noun of a Noun-Adjective term: it transforms *système de surface* (surface system) into *
système racinaire de surface* (surface root system).

- **Morpho-syntactic variants**
  
  These variants involve at least one of the content words of the controlled term being morphologically transformed. Syntactic modifications can also accompany these variants. Variant (3.e) is a Noun-to-Verb variation that associates the nominal term *variation de alimentation* (supply variation) with the verb phrase *alimentation hydrique d’un arbre varie* (the water supply of a tree varies).

The extraction of variant (3.e) is performed by the following metarule:

```
Metarule NtoV( N1 PREP2 N3 )
```

Table 3: FASTR’s output on sentence (S)
Table 4: First 10 Noun-to-Verb variants of [AGRO-ALIM] retrieved by metarule NtoV

<table>
<thead>
<tr>
<th>Variant no</th>
<th>Term</th>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4.a)</td>
<td>détermination de teneur</td>
<td>teneur en chrome est également déterminée</td>
</tr>
<tr>
<td>(4.b)</td>
<td>utilisation de chromatographie</td>
<td>chromatographie liquide est utilisée</td>
</tr>
<tr>
<td>(4.c)</td>
<td>traitement de plante</td>
<td>plantes traitées</td>
</tr>
<tr>
<td>(4.d)</td>
<td>analyse de tige</td>
<td>tiges ont été analysés</td>
</tr>
<tr>
<td>(4.e)</td>
<td>élévation de teneur</td>
<td>teneur élevée</td>
</tr>
<tr>
<td>(4.f)</td>
<td>utilisation de le chambre</td>
<td>chambres de refroidissement utilisées</td>
</tr>
<tr>
<td>(4.g)</td>
<td>développement de sol</td>
<td>sols développés</td>
</tr>
<tr>
<td>(4.h)</td>
<td>influence de variété</td>
<td>variété est influencée</td>
</tr>
<tr>
<td>(4.i)</td>
<td>apport de azote</td>
<td>azote apporté</td>
</tr>
<tr>
<td>(4.j)</td>
<td>détermination de facteur</td>
<td>facteurs déterminent</td>
</tr>
</tbody>
</table>

This metarule retrieves 3,100 variants from the [AGRO-ALIM] corpus; Table 4 illustrates the first 10 variants extracted.
Table 5: Quantitative evaluation of free indexing on [AGRO-ALIM]

ACABIT selects 21,000 candidate terms from the [AGRO-ALIM] corpus. They represent 60% of the candidates extracted using linguistic rules before statistical filtering. The numbers of non-variant and variant term occurrences retrieved by FASTR from the [AGRO-ALIM] corpus through these candidates are indicated in Table 5. This table shows that term occurrences represent only 60% of the indexes. The remaining indexes are term variants composed of 17.6% syntactic variants and 22.3% morpho-syntactic variants. Results concerning precision of indexing by FASTR are detailed in Section 6.

This section has presented the ILIAD system for machine-aided indexing and detailed two components: ACABIT for term acquisition and FASTR for automatic indexing. We now turn to the human evaluation of these components for the purpose of computer-aided manual indexing. Section 5 presents the technical organization of the evaluation (protocol, data, experts, and planning), and Section 6 describes the results.

5 Evaluation Methodology

The evaluation of controlled and free indexing by expert librarians with respect to MAI can be decomposed into the following three subtasks:

1. **Controlled indexing.** The experts evaluate the relevance of indexes that consist of occurrences of controlled terms (preferred and synonym terms from the thesaurus [AGROVOC]). Since FASTR can handle term variation, term occurrences can have
a linguistic expression in the documents that differ from their expression in the 
thesaurus. However, only the canonical linguistic expression from [AGROVOC] 
is shown to the expert.

2. **Free indexing.** The experts evaluate the relevance of indexes that consist of oc-
currences of candidate terms acquired by ACABIT. Some of the candidate terms 
also belong to the thesaurus.

3. **Variant extraction.** The experts judge whether the variants extracted through con-
trolled indexing actually refer to the [AGROVOC] term to which they are linked by 
FASTR. The purpose is to measure the quality of the informative and conceptual 
content of the variants.

### 5.1 Evaluation Protocol

The role of the protocol is to separate the various stages of the evaluation and to provide 
the experts with an evaluation form containing a number of criteria to guide them in 
their responses. To fine-tune this protocol, we have taken a number of protocol criteria 
developed for automatic indexing systems (Salton & McGill, 1983) such as time, effort 
and presentation accounted for during the practical implementation (see Section 5.5). 
Criteria for evaluating terminological extraction systems (L’Homme, Benali, Bertrand, 
& Lauduique, 1996; El-Hadi & Jouis, 1998) are also accounted for (see Section 6). 
We now turn to the presentation of the data, the experts and the various phases of the 
evaluation.

### 5.2 Data

The data to be evaluated for controlled indexing is made up of indexes extracted by IL-
IAD from bibliographic records in the field of agriculture (scientific abstracts). The out-
put of free indexing is a set of occurrences of free indexes, terms acquired by ACABIT from the same set of documents. The evaluation of variants concerns morpho-syntactic and syntactic variants of preferred and synonym terms from the thesaurus [AGROVOC] (controlled indexing). The technical details concerning this data are given in Section 3.

5.3 Experts

We have opted for three experts rather than a single expert for the following reasons:

- Working in with several experts neutralizes the effect of a mental thesaurus. The experts can use the thesaurus [AGROVOC] to reach an agreement. Differences such as points of convergence are settled by discussion.

- The thesaurus [AGROVOC] represents around 20,000 terms, while the experts only use about 2,000 for the indexing task. Working in a team increases the indexing coverage compared with the thesaurus.

The abstracts for controlled and free indexing are examined by two experts (primary adjudicators). These experts must be “experienced, knowing what they are talking about” (Gouadec, 1990). The human expert is not infallible and so a second opinion is called for. In order to be able to cross-reference the results and measure the variation between one group of experts and another, a number of abstracts for each part are examined by a third expert (secondary adjudicator). The primary adjudicators are not allowed to communicate with the secondary adjudicator during the experiments.

The two primary adjudicators have a good knowledge of the [AGROVOC] thesaurus. One of the experts has a virtually encyclopedic knowledge of the thesaurus. The first expert was a biochemist and the second an agronomy engineer. A third person, a specialist in biochemistry, but indexing the food domain, was nominated for the counter-evaluation (secondary adjudicator).
5.4 Phases and Evaluation Form

The evaluation takes place in two stages:

5.4.1 Phase 1

The abstracts, the authority list built from [AGROVOC] and the proposed indexing are made available to the experts. They must rapidly come to a decision concerning the acceptability of the proposed index. They have three options: (A) the term can be kept for indexing; (B) the term poses a problem which justifies its rejection or verification; (C) the term represents noise or an artifact.

5.4.2 Phase 2

For cases (B) and (C), and for both types of indexing, the expert must answer a set of questions to identify the motive of the doubt or rejection. For case (A) only, and for free indexing only, the expert must indicate whether an index, which is considered to be a good descriptor, is present or not in the thesaurus. This phase is the direct responsibility of a researcher and is carried out as if it were an inquiry interview based on the questionnaire given in Appendix. The experts have to answer these questions on blank paper. Any linguistic questions arising from the protocol are discussed in the final evaluation meeting.

5.5 Task Description

The evaluation of controlled and free indexing is based on 50 bibliographic records to which 10 control records were added for each type of indexing. The evaluation of the different variants of the [AGROVOC] terms was based on a set of 200 variations.
Figure 2 gives an example of a record given to the experts. For the evaluation of the relevance of variants, the indexing data presented to the experts is the sentence, followed by the term, its variant, and the variation identifier. Figure 3 gives an example of a variant given to the experts: it is a verbal variant of one of the terms in [AGROVOC].

001402
Numéro : 96-0154564
Résumé : *Le séchage par atomisation d’un concentré de lait de jument obtenu par évaporation sous vide permet l’obtention d’une poudre d’excellente qualité chimique.*

<table>
<thead>
<tr>
<th>Indexes:</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Séchage par atomisation</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>(Drying by atomizing)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Lait de jument</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>(Mare’s milk)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Presentation of the candidate indexes

The main remark does not concern the differences between the protocol and its implementation, but the differences with respect to the day to day indexing tasks of the experts. The evaluation only looked at multi-word terms, that is, those made up of at least two full words. However, the single word terms are very frequent and are often sufficiently precise to index documents. The accompanying document showed that their absence was often recorded. In addition, the experts were also quite perturbed by the absence of Latin terms, which was due to a particular process carried out within the indexing program.
<table>
<thead>
<tr>
<th>Tasks</th>
<th>Provisional schedule</th>
<th>Actual schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preliminary meetings</td>
<td>8 h.</td>
<td>9 h.</td>
</tr>
<tr>
<td>Controlled indexing</td>
<td>20 h.</td>
<td>15 h.</td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free indexing</td>
<td>20 h.</td>
<td>15 h.</td>
</tr>
<tr>
<td>Counter-evaluation controlled and free indexing</td>
<td>4 h.</td>
<td>4 h.</td>
</tr>
<tr>
<td>Variants</td>
<td>4 h.</td>
<td>12 h.</td>
</tr>
<tr>
<td>Final meeting</td>
<td>3 h.</td>
<td>3 h.</td>
</tr>
<tr>
<td>Total</td>
<td>59 h.</td>
<td>63 h.</td>
</tr>
</tbody>
</table>

Table 6: Scheduling

6 Results of the Evaluation

The evaluation of the ILIAD program for MAI looked at the performance of the system in terms of its ability to propose adequate descriptors to the librarians. This performance takes into account the noise and silence of the system. We present the results for the three tasks evaluated.

Les travaux des généticiens ont permis d’augmenter le rendement en filet de cette espèce. (Work by geneticists has allowed an increase in the yield of this species.)

< mc >

< pf > Augmentation de rendement (increase in the yield)

< st > augmenter le rendement (increasing the yield)

< tr > XXX,5,NtoV

Figure 3: Presentation of a variation
6.1 Controlled Indexing Evaluation

The quantitative global results of the controlled indexing evaluation on 50 bibliographic records give rise to the histogram shown in Figure 4. This histogram shows the total number of indexes produced by FASTR from the binary terms in the authority list, the number of correct indexes is marked A, the number of uncertain indexes is marked B, and the number of bad indexes is marked C. We also show the number of indexes which appear in the abstracts and in the authority list and which were not indexed by FASTR. These enabled us to evaluate the silence.

In order to evaluate the quality of indexing by ILIAD, we have adapted the measures of recall and precision used in IR to automatic term extraction. In IR the goal is to retrieve all the documents relevant to a query, and only them. In automatic controlled indexing, the purpose is to extract all the occurrences terms and variants, and only them. Thus the precision $P$ and the recall rate $R$ can calculated by transposing the usual formulae of (Salton & McGill, 1983) from IR to term extraction.

The precision corresponds to the number of indexes marked A by the experts divided by the number $T$ of indexes produced. The recall rate corresponds to the number of indexes marked A by the experts divided by the number of correct indexes, that is, the number of indexes marked A together with the number $S$ of indexes in the authority list which were not identified by FASTR. The precision and the recall rate were calculated for each abstract. Two types of mean can be calculated. The first, called User Oriented Mean, reflects the performance that a user can expect of the system and corresponds to the arithmetic mean of the precision and recall rate divided by the number of records.
Figure 4: Controlled indexing quantitative evaluation

- **T**: Total number of candidate indexes;
- **A**: Marked A;
- **B**: Marked B;
- **C**: Marked C;
- **S**: Indexes not produced.
NUM. The User Oriented mean precision and mean recall are given by:

\[ P_{UO} = \frac{1}{NUM} \sum_{i=1}^{NUM} \frac{A_i}{T_i} \]  
(1)

\[ R_{UO} = \frac{1}{NUM} \sum_{i=1}^{NUM} \frac{A_i}{A_i + S_i} \]  
(2)

The second, called System Oriented mean, corresponds to the mean values of the precision and recall for all records. The System Oriented mean precision and mean recall are given by:

\[ P_{SO} = \frac{\sum_{i=1}^{NUM} A_i}{\sum_{i=1}^{NUM} T_i} \]  
(3)

\[ R_{SO} = \frac{\sum_{i=1}^{NUM} A_i}{\sum_{i=1}^{NUM} A_i + S_i} \]  
(4)

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Oriented</td>
<td>( P_{UO} = 0.75 )</td>
<td>( R_{UO} = 0.47 )</td>
</tr>
<tr>
<td>System Oriented</td>
<td>( P_{SO} = 0.87 )</td>
<td>( R_{SO} = 0.77 )</td>
</tr>
</tbody>
</table>

Table 7: Precision and recall for controlled indexing

According to the figures shown in Table 7, we observe a large difference between the User Oriented and System Oriented mean recall. The System Oriented mean recall gives an equal weight to each abstract while the User Oriented mean recall gives more weight to abstracts where many indexes have been produced than to those where few have been indexed. This difference shows that a number of abstracts were either not indexed or only slightly indexed.

### 6.2 Free Indexing Evaluation

The quantitative global results of the free indexing evaluation on 50 bibliographic records give rise to the histogram shown in Figure 5. This histogram shows the to-
Figure 5: Free indexing quantitative evaluation
tal number of indexes extracted by FASTR when using the binary terms acquired by
the ACABIT module. The number of correct indexes is marked A, the number of un-
certain indexes is marked B, and the number of bad indexes is marked C. We also show
in column S the number of indexes which appear in the abstracts and are to be found
in either the authority list or in the list of candidate terms judged to be interesting by
the adjudicators, but which were not indexed by FASTR. These allow us to evaluate the
silence, i.e. the non-availability of multi-word terms.

The precision $P$ is calculated in the same way for free indexing as for controlled in-
dexing. The recall rate $R$ corresponds to the number of indexes considered A by the
adjudicators divided by the number of good indexes, that is, the indexes considered A
added to the number of unidentified indexes which are contained in the authority list
and the number of indexes which are not contained in this list, but which the experts
consider should be added to it (column S). The two types of mean are calculated using
Formulae 1, 2, 3 and 4. These figures are given in Table 8.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Oriented Mean</td>
<td>$P_{UO} = 0.78$</td>
<td>$R_{UO} = 0.69$</td>
</tr>
<tr>
<td>System Oriented Mean</td>
<td>$P_{SO} = 0.80$</td>
<td>$R_{SO} = 0.66$</td>
</tr>
</tbody>
</table>

Table 8: Precision and recall for free indexing

In contrast to the controlled indexing, the figures for recall are better for the user ori-
ented mean than the system oriented mean. This is due to the fact that the system is
more stable from one abstract to another and is never silent.

6.3 Variant Evaluation

FASTR takes into account two major families of variants:
1. **Syntactic variations** bring into play a transformation of the syntax of the term without morphological modifications to the words other than by inflection.

Among these variations, three sub-families can be identified:

- Weak syntactic variations including variations in the preposition: *séchage par vide* (drying by vacuum) / *séchage sous vide* (drying under vacuum) and the insertion or not of a determinant: *lait de vache* (milk from cows) / *lait de ces vaches* (milk from these cows).

- Insertion variations allowing the insertion of an adjective or modifying adverb in the term: *teneur en lipide* (contains fat) / *teneur variable en lipide* (contains variable fat), or a more complex sequence: *réglementation de production* (legislation in production) / *réglementation dans certaines zones de production* (legislation in certain areas of production).

- Coordination variations: *transfert d’énergie* (transfer of energy), *transferts de masse et d’énergie* (transfers of mass and energy).

2. **Morpho-syntactic variations** containing at least one word of an initial term which has been subject to a morphological transformation which may be accompanied by a syntactic variation.

Four sub-families can be distinguished:

- transformation from adjective to noun: *activité alimentaire* (food activity) / *activité de l’aliment* (activity in food);

- transformation from noun to adjective: *agent de conservation* (preservative agent) / *agents conservateurs* (preserving agent);

- transformation from noun to noun: *acide gras* (fatty acid) / *acidité grasse* (fatty acidity);
• transformation from noun to verb: *contrôle de qualité* (the control of quality) / *contrôlent la qualité* (controlling the quality).

Figure 6 summarizes the adjudication of the experts on the relevance of the syntactic variations.

```
<table>
<thead>
<tr>
<th></th>
<th>WSV</th>
<th>INS</th>
<th>COO</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>27</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>B</td>
<td>22</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
```

WSV: Weak Syntactic Variations;

INS: Insertions;

COO: Coordinations;

A: Marked A

B: Marked B

C: Marked C

Figure 6: Syntactic variant evaluation

The in Figure 6 histograms show the high precision of these variations:

• The coordination variants are nearly all correct. The rejections expressed by the experts cover variations which reflect another domain from the one covered by the thesaurus. For instance, *milieu de culture* (lit. middle of culture) refers to micro-organisms while the identified coordination: *milieu et l’année de culture* (lit. middle and the year of culture) refers to natural conditions.
The weak syntactic variations are also correct in general with the exception of certain prepositional variations: the variation of the preposition *de* (of) with the preposition *par* (by): *transport par eau* (transportation by water) does not refer to *transport d’eau* (transportation of water). Similarly, for certain terms, the use of a given preposition allows us to identify unambiguously the subject of the concepts represented by the term: thus *perte de poids* (lit. loss of weight) applies to animals and humans while *perte en poids* (lit. loss in weight) only affects vegetation.

The problematic variations of insertion only concern the insertions of a sequence *L Preposition (Determinant) Noun* inside a term in the form *(Noun Prep Noun)* or *(Noun Adj)*: *traitement du poisson* (lit. with processing of the fish) and the variant *traitement par les huiles de poisson* (lit. processing by fish oil). In this example, the problem of the semantic difference between the prepositions *de* and *par* is again seen. *Sécurité alimentaire* (lit. food security) refers to the domain of provision while the variant *sécurité en consommation alimentaire* (lit. security in alimentary consumption) refers to consumption.

<table>
<thead>
<tr>
<th></th>
<th>Coor</th>
<th>Modif</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic precision</td>
<td>97.2%</td>
<td>88.7%</td>
<td>95.7%</td>
</tr>
<tr>
<td>Documentary precision</td>
<td>89.3%</td>
<td>71.4%</td>
<td>80.4%</td>
</tr>
</tbody>
</table>

Table 9: Comparison of precision for syntactic variations

Table 9 provides a comparison, for syntactic variations, of the linguistic precision, which only takes into account the linguistic correction of the variant, and the documentary precision, which covers both the linguistic correction and the validity of the variant as an index of the document.

Figure 7 summarizes the opinion of the experts on the relevance of morpho-syntactic
variations. A quick look at these histograms show that these variations are much less

\[ \begin{array}{cccc}
\text{A to N} & \text{N to A} & \text{N to N} & \text{N to V} \\
23 & 17 & 29 & 26 \\
11 & 11 & 15 & 15 \\
4 & 0 & 0 & 0 \\
8 & 0 & 0 & 0 \\
\end{array} \]

precise than the syntactic variations. The variations which are rejected almost systematically by the experts concern the derived variations accompanied by a permutation: *aliment de régime* (diet food), *régime alimentaire* (food diet). Changes in domain appear between the basic term and the variant in the same way as for the syntactic variations, and in roughly the same proportion. A large number of variations considered to be bad by the experts concerned variations which generate a term that is contained in the thesaurus. Thus *filets de poisson* (fish fillet) obtained from *filetage du poisson* (fish filleting) is refused since *filet de poisson* (fish fillet) is contained in the thesaurus.
Table 9 compares, for morpho-syntactic variations, the linguistic precision, taking into account the linguistic correction of the variant, and the documentary precision which takes into account not only the linguistic variation, but also the validity of the variant as a suitable index.

<table>
<thead>
<tr>
<th></th>
<th>A to N</th>
<th>N to A</th>
<th>N to N</th>
<th>N to V</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic Precision</td>
<td>68.5%</td>
<td>69.6%</td>
<td>92.1%</td>
<td>75.3%</td>
<td>84.6%</td>
</tr>
<tr>
<td>Documentary Precision</td>
<td>47.8%</td>
<td>64.7%</td>
<td>51.7%</td>
<td>57.7%</td>
<td>54.7%</td>
</tr>
</tbody>
</table>

Table 10: Comparison of precision for morpho-syntactic variations

7 Synthesis and Discussion

7.1 Evaluation Report

The preceding section has presented the results on the evaluation of indexing and variant extraction. These results show that, on the one hand, controlled indexing is more precise than free indexing, but that the values of precision obtained from both kinds of indexing are very close. On the other hand controlled indexing can sometimes be silent as is shown by the values of recall. This is never the case for free indexing. Their respective performances are thus complementary and show the interest of terminological extraction for MAI. The ideal system would be a hybrid system combining free and controlled indexing.

The variant evaluation gives cues to improve controlled indexing:

- **Syntactic variations**
  - Weak syntactic variations: it would be necessary to prevent some preposition swap such as of/bys.
– Insertion variations: insertions of complex sequences such as a sequence \((\text{Preposition (Determinant) Noun})\) inside a \((\text{Noun Prep Noun})\) or \((\text{Noun Adj})\) term should not be accepted.

- **Morpho-syntactic variations**

  For all families of morphological transformation:
  
  – the permutations should not be accepted;
  
  – a variant generated by a morpho-syntactic rule should only be accepted if it belongs to the controlled vocabulary.

### 7.2 Comparison Between Human Indexing and Automatic Indexing

The primary goal of the work that G. Salton carried out on the SMART system (Salton & Lesk, 1971) was the automation of the task of expert librarians in order to limit the manpower cost of information storage and retrieval. The results obtained with the SMART system at the end of the Sixties and in the Seventies were evaluated by comparing them with those obtained from collections that had been indexed manually. These evaluations showed that SMART, after the incorporation of a certain number of refinements, could do just as well if not better than human indexing at a considerably reduced cost in manpower.

In the SMART evaluations, the comparison between two indexing methods was made at a functional level (in vivo evaluation): the parallel task of looking at the two indexing contents was not carried out, but the performance of an information retrieval system using either one or the other type of indexing was evaluated. The comparison measure did not therefore concern the indexing as such (in vitro evaluation), but rather the quality of the documents retrieved as a function of the queries. To do this, the precision and recall measurements are plotted on graphs, the best system being the one which
comes closest to the theoretical curve of an ideal system with high recall and precision. Measurements combining recall and precision have been suggested, such as the average 11-point precision or the efficiency measure $E$ (Van Rijsbergen, 1975), as well as measurements for comparing sets of precision and recall values, such as the $t$-test, the sign test or the Wilcoxon test (Salton & McGill, 1983).

The conclusions drawn from the comparisons between human and automatic indexing applied to information retrieval show that automatic indexing is better than human indexing (Salton & Lesk, 1968; Salton, 1969, 1972). Salton (1969) describes the improvement of the information search as a function of refinements in the indexing function. The research procedures based on a simple automatic indexing are slightly worse than those based on human indexing (15 to 20%). However, the use of automatically generated dictionaries or a thesaurus associated with the use of document relevance indicators improved the performance dramatically and made automatic indexing superior to human indexing.

In (Salton, 1969), the sources of errors in the information retrieval from human indexes in MEDLARS are analyzed. MEDLARS is a system for accessing medical information that relies on human indexing methods and automatic search techniques. Poor matching between human indexes and query formulation often results in inadequate query/document matching. From the point of view of the indexing language, the terms are too specific (coordination of incorrect simple terms) or too generic (absence of a specific pertinent term). In terms of the formulation of the queries, the same problems arise because of the inadequacy of the terms chosen by MEDLARS. Human indexing is also too specific, too generic or lacking in terms, which prevents adequate query/document matching. Finally, the absence of relevance feedback in the MEDLARS system is also a factor degrading performance.
The comparative evaluations of SMART and MEDLARS provide information on the respective merits of automatic and human indexing. However, two characteristics of these evaluations must be borne in mind. Firstly, they were carried out in a pioneering period in the search for information when it was necessary to prove that the machine was better than the human operator. From this point of view, the human indexes were analyzed on massive documentary tasks where it was certain the performance of the machine would be good. Other tasks, such as information analysis, for which expert competence is more important, were not considered. It is not necessarily the best features of human indexing that are being compared. Moreover, techniques in human indexing have also developed in parallel with automatic techniques as a result of the availability of tools to assist with indexing. Consequently, the level required has changed and the tasks have diversified: scientific monitoring, topical analysis, cartography of a domain, etc. Finally, and this is the most important point, the trials that we carried out did not consist in analyzing human indexing from the point of view of an automated information search, but to see how automatic indexing and automatic acquisition of terminology could be useful to an expert librarian. The idea was thus to investigate the human usability of automatic indexing, not to see its effectiveness, but to analyse its real efficiency in performing the various information analysis tasks that face the expert.

7.3 Agreement Between Experts

In order to analyse the relevance of this evaluation, we performed a counter-evaluation on a sub-set of records. As a measure of the reliability of the results obtained from the expert evaluations, we now evaluate how much the experts agreed on their decisions. To evaluate the agreement of the experts on the tasks of classifying the extracted terms (results A, B and C), we use the Kappa coefficient (Siegel & Castellan, 1988).
The Kappa test measures the degree of agreement between the adjudicators who have carried out classification tasks that we gave them i.e. to categorize the quality of the terms.

The Kappa coefficient \( (K) \) compares the agreement rates obtained with the agreement rates that would have been obtained if they had answered at random. The rate combines the observed agreement frequency \( P(A) \) with the randomly obtained agreement frequency \( P(E) \) (Formula (5)).

\[
k = \frac{P(A) - P(E)}{1 - P(E)} \tag{5}
\]

The agreement frequency is the quotient of the number of identical responses supplied by the adjudicators over the total number of responses. The agreement frequency which would be due to chance is the sum of the products of the response frequencies in each category by each adjudicator (applying the hypothesis that the probabilities of responses between adjudicators are independent). For the controlled and the free indexing, 20% of the records for each indexing are submitted to another adjudicator. Tables 11 and 12 summarize the agreement and dissension rates respectively for the three possible answers on controlled and free indexing. Table 13 presents the agreement frequencies and Kappa coefficient.

Landis and Koch (1977) give a scale of agreement between the adjudicators as a function of the value of the Kappa coefficient. One considers that the adjudicators were in agreement if the agreement rate is greater than 0.279, while a value greater than 0.805 corresponds to total agreement between the evaluators. On the other hand, values of less than 0.2 put into question the classification carried out because of a lack of agreement between the evaluators due either to a bad formulation of the task or a lack of competence.
### Table 11: Agreement data for controlled indexing

<table>
<thead>
<tr>
<th>Primary Adjudicators</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>30</td>
<td>1</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>68.2%</td>
<td>2.3%</td>
<td>9.1%</td>
<td>79.5%</td>
</tr>
<tr>
<td>Secondary B</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>15.9%</td>
<td>2.3%</td>
<td>2.3%</td>
<td>20.5%</td>
</tr>
<tr>
<td>Adjudicator C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>37</td>
<td>2</td>
<td>5</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>84.1%</td>
<td>4.5%</td>
<td>11.4%</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Table 12: Agreement data for free indexing

<table>
<thead>
<tr>
<th>Primary Adjudicators</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>67</td>
<td>2</td>
<td>10</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>70.5%</td>
<td>2.1%</td>
<td>10.5%</td>
<td>83.2%</td>
</tr>
<tr>
<td>Secondary B</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>5.3%</td>
<td>1.1%</td>
<td>2.1%</td>
<td>8.4%</td>
</tr>
<tr>
<td>Adjudicator C</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1.1%</td>
<td>0%</td>
<td>7.3%</td>
<td>8.4%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>73</td>
<td>3</td>
<td>19</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>76.9%</td>
<td>3.1%</td>
<td>20%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 11: Agreement data for controlled indexing

Table 12: Agreement data for free indexing
We note that while there is a good agreement frequency, the Kappa coefficient reveals no agreement between the experts for the controlled indexing and a moderate agreement for free indexing. These 'weak' values of the Kappa coefficient could be due to the small number of indexes submitted to both groups of experts. However, careful reading of the accompanying document shows a different interpretation of the protocol by the pairs of primary adjudicators and the secondary adjudicator.

**Primary Adjudicators** These adjudicators identified the domain of the bibliographic reference and accepted (response A) the indexes belonging to this domain even when they did not strongly reflect the main theme of the article. Moreover, they systematically refused (response C) the indexes presented in a linguistic form which was considered incorrect.

**Secondary Adjudicator** This expert only accepted (response A) those indexes which characterized the main theme of the article. Those indexes which did not reflect the main subject were marked B. However, he was not overly concerned by the linguistic form of the index.

This difference in interpretation shows that, even though the protocol was precisely defined, it was interpreted differently by the different experts, particularly for responses A and B. To this problem should also be added the problem of documentary usage for which there was no consensus. Each expert had his own interpretation on whether a term was central or not for an indexing task.

<table>
<thead>
<tr>
<th></th>
<th>Agreement Frequency</th>
<th>Kappa Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled Indexing</td>
<td>80.4%</td>
<td>0.08</td>
</tr>
<tr>
<td>Free Indexing</td>
<td>78.9%</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Table 13: Agreement Frequency and Kappa Coefficient
Agreement between the experts is distinctly better for free indexing. Since this type of indexing does not refer to an authority list, the unknown terms proposed are more often than not accepted positively. Moreover, we were unable to detect from the accompanying document any differences in interpretation concerning the free indexing. The experts generally accepted the indexes if they correctly reflected the contents of the bibliographic record.

The previous section showed that the expert groups interpreted their evaluation task differently, particularly for controlled indexing, which a priori was simpler. Apart from differences in interpreting the linguistic form of an index, the primary adjudicators accepted an index if it was present in the thesaurus and it characterized satisfactorily the domain of the bibliographic record. The secondary adjudicator accepted an index if it corresponded well to the key-phrase which he would have used for indexing. This difference in interpretation is directly related to the operation of the ILIAD program: the primary adjudicators evaluated the help that the program could give them in carrying out their indexing task, while the secondary adjudicator evaluated it relative to automatic indexing. The aim of this evaluation was in fact to measure the help that the ILIAD program could be to human indexers, and so the incorrect interpretation made by the secondary adjudicator is not too serious.

Whatever precautions are taken for preparing an evaluation, there remain many possibilities for disagreement between judges. Several studies have focused on the issue of relevance and on the variety of criteria affecting users’ relevant judgment. Mizzaro’s exhaustive review of the literature on relevance (1997) cites several papers that insist on the factors that can lead to disagreements between judges. Among these factors, subjectiveness is important and depends both on intellectual and psychological characteristics of the judges such as level of expertise and openness to information. In his
introduction to a special issue of the Journal of the American Society for Information Science, Froehlich (1994) also outlines all the difficulties linked to the measure of relevance because of its grounding in user’s experience. Our results show that all the precautions taken during the elaboration of the protocol and the explanations given at the preliminary meetings with the experts were still insufficient and that human factors have led to significant disagreements.

8 Conclusion

The ILIAD project has defined a program for MAI based on NLP techniques which can be used to work with both free indexing and controlled indexing of multi-terms. Controlled indexing allows the identification of terms—either in their initial form or in a linguistic variant—from an authority list of terms. Free indexing is a way of updating the authority list by proposing candidate terms when no reference terms are available to index a document. These two kinds of indexing need a pre-syntactic processing of the data (managing problems of case and segmentation at word or sentence level): this phase, considered as a low-level process, is of prime importance in a documentary program (Grefenstette & Tapanainen, 1994). This evaluation demonstrates that the ILIAD environment offers good efficiency for the assistance of human indexing. It has also proved that controlled and free indexing are complementary.

Three main tasks would improve the efficiency of multi-word term indexing:

- Increasing the precision of the controlled indexing by constraining some variant generation rules;
- Adding consistency tests for the controlled indexing to prohibit the generation of term variants already present in the indexing dictionary;
Improving the normalizing of candidate-terms provided by the termer in order to avoid consistency tests for free indexing.

As this evaluation shows, an ideal system would be a hybrid system combining free and controlled indexing, efforts have still to be made to improve the interface of the two systems. In addition, to complete the program, other modules need to be added: one for the indexing of Latin terms and the second for the indexing of single word terms (terms containing only one word).

References


Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In J. L. Klavans & P. Resnick (Eds.), The balancing act: Combining symbolic and statistical approaches to language (pp. 49–66). MIT Press.


Milstead, J. L. (1997). Use of thesauri in the full-text environment. In P. A. Cochrane & E. H. Johnson (Eds.), Visualizing subject access for 21st century information resources; proceedings of the 34th annual clinic on library applications of data processing (pp. 28–38). Champaign, IL.


Appendix: Questionnaire to the Experts

Controlled and Free Indexing

1. INDEX IS PHYSICALLY PRESENT IN THE THESAURUS

The index physically appears in that form in the thesaurus. Numerical inflectional variations are accepted. However, the expert must specify whether the candidate index and the term in the thesaurus share:

- same sense and same domain;
- same domain but different sense: polysemy;
- different domain but same sense;
- different domain and different sense.

2. INDEX IS NOT PHYSICALLY PRESENT IN THE THESAURUS BUT ANOTHER EQUIVALENT FORM (PHYSICAL OR SEMANTIC) OF THIS INDEX IS PRESENT IN THE THESAURUS

For this index, the expert must provide the equivalent form from the thesaurus. The characterization of the link between the proposed index and its equivalent form as found in the thesaurus using the following criteria will be specified during the post-evaluation session:

- Is it a morpho-syntactic or morphological variant? If so, it is necessary to be able to link explicitly all the "occurrences” of the term in the thesaurus with the proposed term.
- Is it a partial synonym, that is, not including all the complete lexical units of the term?
- Is it a total synonym, that is, no visible link exists which could be used to link the index with the term in the thesaurus.
3. INDEX DOES NOT APPEAR IN THE THESAURUS BUT SHOULD BE THERE IN THIS FORM

What are the criteria which led the expert to decide that this term should be in the thesaurus?

- Is it a direct extension of an existing branch, that is, a term which could be the parent (son, brother, cousin, etc.) of an existing term? If so, specify the related term. What type of link is there between the new term and its parent in the thesaurus ("is a" relation, "part of" relation, etc.)? Try to characterize it.

- Does the term belong to a sub-domain which is not covered by the thesaurus?

4. INDEX IS NOT PRESENT IN THE THESAURUS BUT SHOULD BE INCLUDED IN A DIFFERENT FORM

(Use criteria from (2) and (3))

5. UNWANTED TERM

Why?

- Too general, uncharacteristic of the domain, no terminological status.

- Bad descriptor (too frequent?) but nevertheless pertinent from a terminological point of view.

- Too precise or refined to be used as an index. By which generic term which is present in the thesaurus should it be replaced?

- Results from the composition of one or more terms. Which ones?

TERM VARIANTS - CONTROLLED INDEXING

Are the variations identified by FASTR correct? Determine whether these variations correspond semantically to the term in the thesaurus.
The following criteria for rejection or acceptance should be applied:

1. Incorrect variation: is it caused by one of the following phenomena?
   - Syntax: the construction proposed is incorrect.
   - Polysemy: the sense of certain words in the variant is not the same as in the original term.

2. Correct variation but the sense is too far removed from the original term, for example, the variant is much more specific than the original term.

3. Variation correct, but is incorrectly linked with other terms than the original term. into question.