The ILIAD Project : Analysing Information using Informetrics Techniques and Natural Language Processing

Yannick Toussaint, Nicolas Capponi
INRIA Lorraine & CRIN-CNRS

Abstract

We present the ILIAD project and its current results. ILIAD aims at combining statistic and linguistic approaches in order to analyse information in large documentary databases. The resulting analysis should enable a human operator to collect the information content of a set of texts without having to read it sequentially. Our current experimentation concerns the analysis of a set of abstracts extracted from a documentary database. Our approach relies upon the recent terminology advance in both linguistics and knowledge aspects. In a first step we identify terms in texts and classify them using a statistical algorithm. The second step is a partial linguistic analysis which focusses on terms highlighted by the classification process. Finally, techniques form artificial intelligence are called upon in order to collect and organise the information that emerges from these texts.

1 Introduction

Progress in telecommunication (INTERNET, news, mail...) [Stephens, 1994], in information diffusion (CD-ROM) has made a great amount of information available. Moreover, navigation tools based on hypertext has considerably changed the way of obtaining information trying to conciliate two opposite dimensions: large amounts of texts versus competing in time. These techniques do not involve a *"comprehensive"* process of the texts and stay at the level where a word is simply a string of characters. To improve them, a more detailed analysis of the texts, the sentences and the words could be of great interest. Obviously, linguistic problems occur as well as problems in structuring this information in a knowledge structure. Natural language processing has now reached a mature stage of development which allows us to take into account the specificities of areas of information retrieval and extraction of information from texts. Tools based on these principles have to handle a huge volume of text and to identify valuable information, extract it and structure it.

Jacobs, in [Jacobs, 1994], mentions that the various US Projects such as MUC and TIPSTER shows that words and word relationships approaches are well adapted to treat a large amount of texts. It seems that approaches based on well-developed grammars or discourse models, which emphasise syntax or traditional syntax, handicap projects more than help them [Salton *et al.*, 1994].

The ILIAD Project aims at developing tools and methods that enable a human operator to collect information without reading it sequentially. Its originality lies in the combining of both a statistical and a linguistic approach in order to develop more robust tools. It relies on the established fact that the major part of information in technical or scientific texts is preferably located in noun phrases. Therefore, our strategy for analysing information relies upon the recent terminology advances in both linguistics and knowledge aspects. Linguistics will contribute to the project by identifying relationships. Informetrics will be helpful in selecting the information.

ILIAD is dedicated to the treatment of abstracts of technical or scientific texts stored in documentary databases. It is developed in two stages which also correspond to two different steps in the treatment of texts. The **first step** is a platform of experiments in which tools perform a robust analysis of the

texts : for each text, it identifies the important terms and their co-occurrence relationships with other terms of the domain. The **second step** is based on a more detailed linguistic analysis of noun phrases in order to detect predicate structures. These predicates extracted from texts will then be structured in such a way that the elementary information does not remain divided into small portions, but integrated into a knowledge base.

We will test ILIAD on a real corpus in different situations : as terminology is heavily dependent upon the domain, the main domain of experiment is agriculture. Nevertheless, we have already done some experiments (in the 1st step) on medicine. We decided to focus on the analysis of the abstract entry of document descriptors in a documentary database. The advantage, compared to full texts, is that the information is more *concentrated*, terms are more *standardised*. As a result, linguistic structures are more complex, sentences longer... We also plan to observe how ILIAD works on full texts.

Partial linguistic description means that it is easier to treat different languages, especially if the languages have the same Latin origin (French, Spanish...). The first stage is near completion for the French and English language and the current results are very promising. We plan to develop modules to treat Spanish following the same architecture. We are currently working on the second phase, more specifically on the French language.

2 The ILIAD Project

2.1 What is information analysis ?

Information analysis started with Information Retrieval techniques using keywords to index the texts. A simple list of keywords allows a boolean search but doesn't reflect the weight of each keyword and the relationships between them. Vector processing is more flexible and takes into account both the keywords describing documents and also keywords in the question. The vector of the question is compared to the vectors describing the documents and the closer to the question the answer is, the better. An even better perfomance can be reached by weighting the vector associated to a keyword by a value which reflects its frequency [Salton *et al.*, 1994].

Cesare,[Di Cesare, 1994], adopted a different definition, evaluating grey litterature using bibliometric indicators. She proposed to compare two sets of documents looking at the distribution of these documents, for example, by comparing journal or types. The analysis of the set of most representative journals, for example, then gives the *scientific* trend of the set of documents. However, this approach is too far from linguistics and terminology and we will not take it into account in our framework.

Combining keyword indexation, collocalisation or co-occurrence, tools such as NDOC or SDOC have been developed at INIST [Grivel and Francois, 1995; Grivel *et al.*, 1995]. They build clusters of terms that can be visualised on a map. SDOC builds a matrix of co-occurrences and uses an algorithm which fills the clusters. At the first step of the classification, some keywords can be clearly identified as noise. After one or two filtering operations by an expert, clusters are homogenous and can be visualised on a 2-dimensional map (see Figure 1).

The y-axis corresponds to density and two regions are taken into consideration : the A-B half of the map contains clusters that are composed of closely related terms (inside the cluster), as opposed to the C-D half in which clusters contain more general terms. The x-axis corresponds to centrality : the B-D half of the map contains clusters whose terms have more relations with other clusters (ouside the cluster) than A-C clusters have.

The B-quarter is then very interesting. Clusters are more homogeneous and have a lot of relations with other clusters. We will use this method of classification to discriminate information, important or not, in a domain. Another experiment in the context of ILC Project – and in cooperation with the Knowledge Base Team of INIST – shown that these clusters can be verbalised by an expert and are *coherent* in the domain.

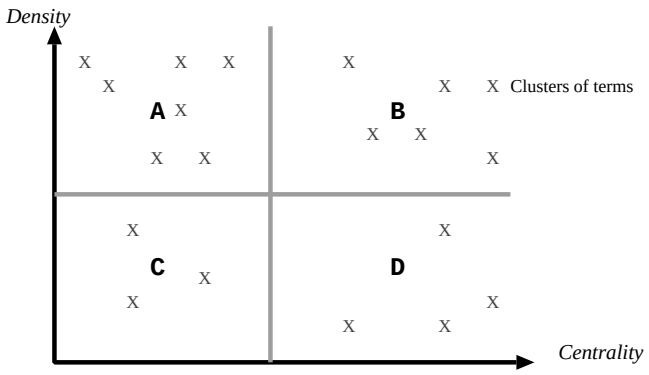

Figure 1: Visualisation of clusters using SDOC

Our definition of the information analysis process is a mixture of several approches. We define **Information Analysis** as **the step after the information retrievial process** : a user asks a documentary database, a question using, for example, a logical combination of keywords, (let us take the keyword *agriculture*). The answer given by the system is a set of document descriptors (*2069 descriptors*). Each descriptor gives the title of the document, the author(s), the keywords and the abstract... The information analysis enables this user to characterise the overall content of the set of documents i.e. to extract from the abstracts the specificity of this set of documents. To perform this analysis, our system extracts terms are from abstracts and indexes them. The clusterisating process discriminates information (filters terms which are important or not) : we focus on terms which are in the B-quarters' clusters. The linguistic analysis then labels relations between terms according to the set of documents. This information is structured and thereafter, the structured information can be considered as a partial knowledge base of the domain corresponding to the knowledge refered to in the set of documents.

2.2 ILIAD global architecture

Figure 2 gives the general architecture of ILIAD, whose specificness relies on the association of linguistics and statistical techniques. The architecture results from several considerations :

1. Sentences in scientific abstracts are complex sentences that actual syntactic parsers are not able to treat efficiently : they would generate lot of ambiguities in syntactic trees ;

2. Terminology – identifying terms and relations – seems to provide a good answer to the syntactic analysis with partial techniques. However, if the notion of a term is clear for translators[1], it is less clear when looking for information analysis.

3. At present, there is no good[2] linguistic criteria to distinguish a noun phrase which is a term of a domain from one which is not a term. Neither is there any linguistic criteria to distinguish nounphrases which carry information and those which do not.

[1] A term is introduced in the terminological base if it is a noun phrase for which the translation of each of the word of the nounphrase does not correspond to the translation of the global term.

[2] Of course this could be open to discussion. However, if this criteria exists, they involve a very detailed linguistic analysis not only of sentences but, also of texts.

4. The tools for clustering documents on the basis of their keyword description and the tools for visualising these clusters on a map highlights very efficiently important terms and relations (co-occurences) between them.

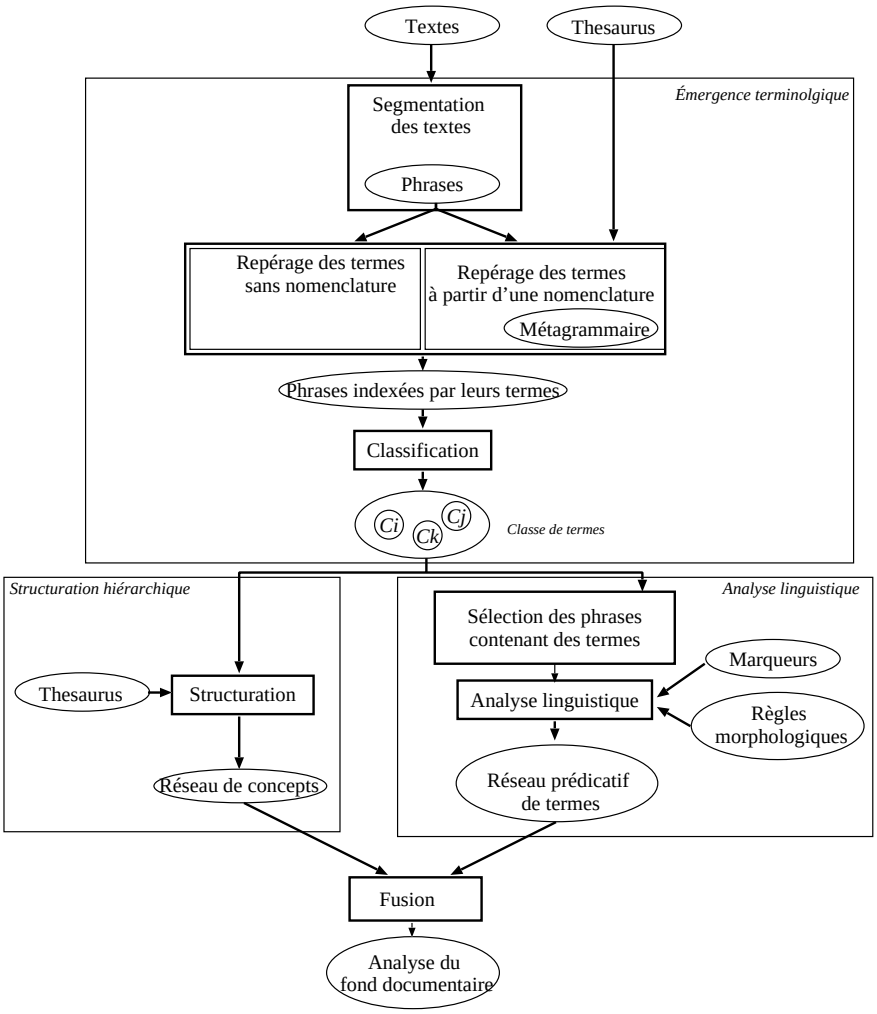

Figure 2: ILIAD global architecture

The first step produces a set of terms that are grouped into clusters on the basis of the co-occurence in the documents. Informetrics (looking at the map) then provides criteria to choose the most significant clusters.

For each of these clusters, the second step will perform a more detailed linguistic analysis to extract predicate structures. As we are looking for relations between terms inside one cluster, the analysis does not need to be as detailed as in the case of syntactic parsing.

3 Combining informetrics and linguistic tools to classify terms and texts

3.1 Searching terms in texts and indexing

The first phase of ILIAD is based on term occurence and coocurences. This first step consists of identifying terms. Some tools that we will integrate to our platform in order to identify terms already exist (FASTR, ACABIT). In order to run these tools we need to tag the texts and the thesaurus. We will compare the two different strategies that will be detailed in 3.1.2 and 3.1.3 :

- a thesaurus of the domain is used to extract a previous list of attested terms. Then, FASTR will locate these terms and their variation in texts. This method is very efficient but is limited by the coverage of the thesaurus : news terms will not be located ;
- ACABIT is used to automatically extract terms. In contrast to the first one, this method is noisy and proposes term candidates which are not of interest in the domain.

We have already integrated the first one. The second one is not yet operational.

3.1.1 Tagging and lemmatisation

In order to run FASTR and ACABIT we had to tag the thesaurus and the texts and lemmatise them. First of all, we tagged the thesaurus in order to build term rules for FASTR. We are currently working on tagging French texts whose complexity needs a more complete training. We used the Brill tagger [Brill, 1992], a statistical tagger that can be trained. We trained it on 3000 French entries of the AGROVOC Thesaurus and then tagged the thesaurus (20,000 entries) with a level of correctness of 95%. After being trained and whilst tagging a text, the Brill tagger proceeds in three steps : a lexicon gives, for all the words the tagger has been trained on, the list of tags that could be assigned. The first tag of the list is the most probable. If the tagger tries to tag an unknown word, it will then access the lexical rules. These rules have been automatically built during the training phase. They are built from a set of morphological primitive functions that try to assign a tag following the surface form of the word. Then, at the end of the process of tagging, contextual rules are applied which modify the tag previouly assigned to the current word according to the tag preceeding or following it.

The lemmatisation process is based, at present, on a partial analysis of the words. About 10 rules for nouns and 10 for adjectives allow the system to identify the lemma. However these rules do not permit disambiguisation. For example, the *-is* suffix for a noun can designate either a singular or a plural. *"acquis"* in French is either singular or plural but *"ami"* is singular, and its plural form is *"amis"*. Anyway, FASTR treats morphological ambiguities and the process is robust and does not need a dictionary. Its could be improved by integrating a dictionary.

3.1.2 Terms from a thesaurus and their variations

This approach relies on a thesaurus which gives a preliminary set of terms in the domain of agriculture. We are currently experimenting with our tools on agricultural texts using the AGROVOC thesaurus. This set of terms is completed by the term variations. Term variations are located by FASTR [Jacquemin, 1994]. For each term in a thesaurus or a lexicon, FASTR describes the syntactic structure of the term using PATR-II formalism. Then it uses a set of metarules which enables FASTR to recognise a variation of a term. For example, in French *solution de fluor* can occur also under the forms *solution fluorique* or *solution fluorée* ; *valeur moyenne* under *valeur annuelle moyenne*. In English, the term *term variation* can be found in texts as *variation of terms*. When processing texts with FASTR, we keep both the thesaurus' (initial) form of the term and the type of variation (ex: *of-inversion*). Texts are indexed by the terms extracted.

3.1.3 Automatically extracting terms from texts

This work is still in progress. We would like to complete the initial set of terms given by the thesaurus by the list that could be provided by a tool automatically extracting terms from a text. However, these tools generate noise and propose a great number of terms. The work we will perform in the project involves adapting the ACABIT [Daille, 1994] tool to our purpose and trying to reduce the noise in term extraction.

3.2 Clustering

We will not detail the process of clustering. SDOC is a tool developed at INIST. We processed 2069 abstracts. The average of recognised terms per abstract is 13.36. The system recognised a total of 4106 terms. All these words are good indexes.

4 From terms to knowledge

Different kind of knowledge acquisition environments have been developed over the past years. However, these environments usually take into account human-being factors and concentrate on the "how-to-make-the-expert-talk" about his knowledge more than how this knowledge should be structured and encoded. In the Software Engineering Process, [Plant, 1994] tried to give rules to write texts which describe the knowledge of the expert. He also tried to give a methodology to extract and code this knowledge.

In his article, [Hjorland, 1994] proposes nine principles of knowledge organisation. These principles are fairly general, for example *"Categorisations and classifications should unite related subjects and separate unrelated ones"* or *"Any given categorisation should reflect the purpose of that categorisation"*...This shows how complex the problem is, of defining what we call knowledge and how we structure it.

Behaviour, as well as linguistic entities of terms extracted from texts, is strongly linked with its conceptual description. Terminological approaches try to make the link between knowledge acquisition and terminology building [Meyer *et al.*, 1992; Condamines, 1995; Jacobs, 1994; Czap, 1993] However, a gap between information and knowledge exists. Each different level of the text (lexical, syntactic, semantic, rhetorical...) contributes to the identification of the knowledge a sequence of words carries. The statistical clustering allows us to avoid the difficult question of identifying knowledge and building ontologies. The knowledge we extract from abstracts is partial and we are only aiming to represent a certain form of knowledge which characterises the set of documents we extracted from the documentary base.

Two tasks compose the second step of ILIAD. The first one consists of extracting predicates that operate over two terms. Though, a *flat* list of predicates would be of little help for the user who needs the information analysis of a set of texts. Structuring these predicates following knowledge representation criteria will make the information understandable and enables us to represent complex predicates (predicates involving predicates).

4.1 Analysing noun phrases to identify predicates

Clusters are built following the criteria of co-occurence inside a text. We will assume that terms which are in the same cluster are close semantically. We are planning to experiment with different segmentation. We have done it on the whole abstract but segmenting sentence by sentence may give a more significant result from a linguistic point of view. The best analysis is the one which will let the most important terms of the domain emerge and which will build clusters of terms which are close semantically. Two strategies can then, make explicit the link between them.

4.1.1 Linguistic marks

Linguistic marks that we take into account are expressions such as "*effect of* something *on* something". Other expressions of causality has been studied by [Garcia, 1996]. Consequence is also characterised by a large set of marks. The structure generated by such marks are very often complex and involves predicates build with predicate structures. Most of these marks are not domain-dependent, even though if the frequency of a mark may vary from one domain to another.

4.1.2 Morphological analysis

The goal of this morphological analysis is to identify, where it exists, the existing relation between a predicate in its nominal structure and the verbal or adjectival structure. The use of a dictionary to predict, for a predicate, its argument structure will guide the analysis using LCS-like structures [Jackendoff, 1983; Jackendoff, 1987; Jackendoff, 1990].

This module will be using the ALEP Plateform [Cruickshank *et al.*, 1994] developed for the AT/6.1 and ET9 Projects. It integrates a two-level tool for word segmentation [Koskenniemi, 1983] and a morpho-syntatic parser. These two modules are linked by a type-feature dictionary.

Let us take the folowing sentence as an exemple:

L'effet de l'addition d'enzyme pectolytique au moût sur l'évolution de la fermentation du cidre a été étudiée.

We are able to detect the following predicates :

- *addition de* "term" *à* "term",
- *évolution de* "term",
- *fermentation de* "term".

Predicates that will be produced are :

- *addition(enzyme-pectolitique,moût)*,
- *fermentation(cidre)*,
- *évolution(fermentation(cidre))*.

and the relation : *effect (addition(enzyme-pectolitique,moût), fermentation(cidre))*

4.2 Structuring terms and predicates in a knowledge base

In order to structure the knowledge acquired, we use the thesaurus of the domain. The AGROVOC thesaurus contains hierarchical and non-hierarchical relations whose semantics is more or less precise. Nevertheless, it constitutes a starting point as [Liddy and Paik, 1994] suggests. In order to improve the capacity of structuring, we augmented the thesaurus with a top hierarchy, partly using UMLS categories.

We use a description logic [Brachman and Levesque, 1987; Nebel, 1991] to represent terms of the domain, which are further completed by the predicates found during the linguistic analysis. The hierarchical structure is first used to type terms and associations links resulting from the clustering algorithm. For example, in the CHROMATOGRAPHY cluster, *putrescine*, *histamine* and *biogenic amine* are recognised as kinds of *amine*, and the associations *(putrescine, histamine)* and *(histamine, biogenic amine)* can be typed as subsomption (is-a) links.

Description logics provides a logical framework for automatically classifying concepts given their structural description. This property is used to organise the predicative structures. These are represented using a set of thematic roles. For example, the phrase "*quantitative analysis of biogenic amine*

by chromatography" is represented by the predicative structure "*quantitative analysis(patient:biogenic amine, instrument:chromatography)*" where *patient* and *instrument* are two thematic roles. Given a set of predicates, it is possible to organise them along different criteria to find similarities or differences. For example, the predicative structure "*quantitative analysis(patient:amine, instrument:liquid chromatography)*" will be recognised as similar to the preceding one because of its arguments being more specific with the same thematic roles. By contrast, the predicative structure "*quantitative analysis(patient:ascorbic acid)*" shows another use of a quantitative analysis, due to the differences between *amine* and *ascorbic acid*. This reasoning can be used as well for detecting similarity beetwen different predicates. For example, the predicative structure "*determination(patient:spermine, instrument:liquid chromatography)*" can be deduced to be very close to the first and second one, because *spermine* is a kind of *amine*.

We are still working on specifying how this structuring will help in understanding the content of a set of documents, and how a user would like to interact to extract information.

5 Conclusion and perspectives

The ILIAD architecture has been defined to get different tools to cooperate. The cooperation between the linguistic and the statistic approaches is of utmost importance. It makes the process more robust and proposes an answer to questions about the selection of *important* terms.

With regards to multilinguality, ILIAD allows to process different languages : french, english (and later spanish). However, ILIAD has not been designed to process together texts in different languages. Rather, it can process separately english texts and french texts. But it could be interesting to join multilingual texts for the clustering step, using the thesaurus to match terms from different languages. The other steps are languages dependent, and therefore should always be done separately.

The first step of the project is now running (in french and english) and we are still evaluating it. The actual result is already very interesting : an expert is able to comment on the content of a cluster and to name the relation between terms inside a cluster. We are now working more specifically on the second step in order to automate the process of information analysis.

6 Acknowledgement

The ILIAD Projet is supported by the French Program of Research in Cognitive Sciences (*GIS-Sciences de la Cognition*). Partners involved in this projects are Institut National de l'Information Scientifique et Technique, the Intitut de Recherche en Informatique de Nantes, the University of Nancy II, and the Institut National de la Langue française.

References

[Brachman and Levesque, 1987] R.J. Brachman and H.J. Levesque. Expressivness and tractability in knowledge representation and reasoning. *Computational Intelligence*, 3:78–93, 1987.

[Brill, 1992] E. Brill. A simple rule-based part of speech tagger. In ANLP-ACL, editor, *Third Conference on Applied Natural Language Processing*, Avril 1992.

[Condamines, 1995] Anne Condamines. Terminology : New needs, new perspectives. *Terminology*, 2(2):218–238, 1995.

[Cruickshank *et al.*, 1994] G. Cruickshank, M. Groenendijk, and N. Simpkins. Alep2.3 user guide. Technical report, Luxembourg, CEC, 1994.

[Czap, 1993] H. Czap. Guiding principles for (re-)constructing concepts. In K.-D Schmitz, editor, *Proceedings Third International Congress on Terminology and Knowledge Engineering*, pages 16–23. Indeks Verlag, 1993.

[Daille, 1994] Béatrice Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Université de Paris VII, Computer Communication and Vision, TALANA, février 1994.

[Di Cesare, 1994] R. Di Cesare. The evaluation of grey literature impact using bibliometric indicators. the case of physical sciences. In D.I. Raitt and B. Jeapes, editors, *Proceeding of the 18th International Online Information Meeting*, pages 405–13, Oxford, UK, december 1994. Learned Inf.

[Garcia, 1996] Daniela Garcia. Coatis, un système de repérage d'expressions d'action reliées par des causaalités. Journée Scientifique "Modélisation et Capitalisation des connaissances à partir de textes", Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche, june 1996.

[Grivel and Francois, 1995] L. Grivel and C. Francois. *Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique*, pages 81–113. Jean-Max Noyer, presses universitaires de rennes edition, 1995.

[Grivel *et al.*, 1995] Luc Grivel, Peter Mutschke, and Xavier Polanco. Thematic mapping on bibliographic databases by cluster analysis : a description of the sdoc environment with solis. *Journal of Knowledge Organization*, 22(2):70–77, 1995.

[Hjorland, 1994] B. Hjorland. Nine principles of knowledge organization. In H. Albrechtsen and S Oernager, editors, *Knowledge Organization and Quality Management, Proceedings of the Third ISKO Conference*, pages 91–100. INDEKS Verlag, june 1994.

[Jackendoff, 1983] Ray Jackendoff. *Semantics and Cognition*, volume 8 of *Current Studies in Linguistics*. MIT Press, Cambridge, Mass., 1983.

[Jackendoff, 1987] Ray Jackendoff. The status of thematic relations in linguistic theory. *Linguistic Inquiry*, 18:369–411, 1987.

[Jackendoff, 1990] Ray Jackendoff. *Semantic Structure*. MIT Press, Cambridge, Mass., 1990.

[Jacobs, 1994] P.S. Jacobs. Words, words, words: lexical representation and knowledge acquisition. In *Proceedings of Language Engineering Convention*, pages 75–81. Eur. Network in Language & Speech, july 1994.

[Jacquemin, 1994] Christian Jacquemin. Fastr: A unification-based front-end to automatic indexing. In *Proceedings of Information Multimedia Information Retrieval Systems and Management*, pages 34–47, New-York, october 1994. Rockfeller University.

[Koskenniemi, 1983] K. Koskenniemi. Two-level model for morphological analysis. In *8th IJCAI conference, Karlsruhe*, 1983.

[Liddy and Paik, 1994] E. Duross Liddy and W. Paik. Automatic recognition of semantic relations in text, 1994.

[Meyer *et al.*, 1992] Ingrid Meyer, Lynne Browwker, and Karen Eck. Cogniterm : An experiment in building a terminological knowledge base. In *Fifth Euralex International Congress*, Tampere, Finlande, août 1992.

[Nebel, 1991] Bernhard Nebel. *Terminological Cycles : Semantics and Computational Properties*, chapter 11, pages 331–362. Morgan Kaufmann, 1991.

[Plant, 1994] R.T. Plant. Techniques for knowledge acquisition from text. *Journal of Computer Information Systems*, 35(1):64–70, fall 1994.

[Salton *et al.*, 1994] G. Salton, L. Allan, and C. Buckley. Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2):97–108, february 1994.

[Stephens, 1994] Charlotte S. Stephens. The nature of information technology research : a seven year analysis. *Journal of Computer System Information Systems*, 34(4):67–76, Summer 1994.