Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax*

Christian Jacquemin
Institut de Recherche en Informatique de Nantes, BP 92208
2, chemin de la Houssinière
44322 NANTES Cedex 3
FRANCE
jacquemin@irin.univ-nantes.fr

Judith L. Klavans
Center for Research on Information Access
Columbia University
535 W. 114th Street, MC 1101
New York, NY 10027, USA
klavans@cs.columbia.edu

Evelyne Tzoukermann
Bell Laboratories,
Lucent Technologies,
700 Mountain Avenue, 2D-448,
P.O. Box 636,
Murray Hill, NJ 07974, USA
evelyne@research.bell-labs.com

Abstract

A system for the automatic production of controlled index terms is presented using linguistically-motivated techniques. This includes a finite-state part of speech tagger, a derivational morphological processor for analysis and generation, and a unification-based shallow-level parser using transformational rules over syntactic patterns. The contribution of this research is the successful combination of parsing over a seed term list coupled with derivational morphology to achieve greater coverage of multi-word terms for indexing and retrieval. Final results are evaluated for precision and recall, and implications for indexing and retrieval are discussed.

1 Motivation

Terms are known to be excellent descriptors of the informational content of textual documents (Srinivasan, 1996), but they are subject to numerous linguistic variations. Terms cannot be retrieved properly with coarse text simplification techniques (e.g. stemming); their identification requires precise and efficient NLP techniques. We have developed a domain independent system for automatic term recognition from unrestricted text. The system presented in this paper takes as input a list of controlled terms and a corpus; it detects and marks occurrences of term variants within the corpus. The system takes as input a precompiled (automatically or manually) term list, and transforms it dynamically into a more complete term list by adding automatically generated variants. This method extends the limits of term extraction as currently practiced in the IR community: it takes into account multiple morphological and syntactic ways linguistic concepts are expressed within language. Our approach is a unique hybrid in allowing the use of manually produced precompiled data as input, combined with fully automatic computational methods for generating term expansions. Our results indicate that we can expand term variations at least 30% within a scientific corpus.

We would like to thank the NLP Group of Columbia University, Bell Laboratories - Lucent Technologies, and the Institut Universitaire de Technologie de Nantes for their support of the exchange visitor program for the first author. We also thank the Institut de l'Information Scientifique et Technique (INIST-CNRS) for providing us with the agricultural corpus and the associated term list, and Didier Bourigault for providing us with terms extracted from the newspaper corpus through LEXTER (Bourigault, 1993).

2 Background and Introduction

NLP techniques have been applied to extraction of information from corpora for tasks such as *free indexing* (extraction of descriptors from corpora), (Metzler and Haas, 1989; Schwarz, 1990; Sheridan and Smeaton, 1992; Strzalkowski, 1996), *term acquisition* (Smadja and McKeown, 1991; Bourigault, 1993; Justeson and Katz, 1995; Daille, 1996), or *extraction of linguistic information* e.g. support verbs (Grefenstette and Teufel, 1995), and event structure of verbs (Klavans and Chodorow, 1992).

Although useful, these approaches suffer from two weaknesses which we address. First is the issue of filtering term lists; this has been dealt with by constraints on processing and by post-processing overgenerated lists. Second is the problem of difficulties in identifying related terms across parts of speech. We address these limitations through the use of *controlled indexing*, that is, indexing with reference to previously available authoritative terms lists, such as (NLM, 1995). Our approach is fully automatic, but permits effective combination of available resources (such as thesauri) with language processing technology, i.e., morphology, part-of-speech tagging, and syntactic analysis.

Automatic controlled indexing is a more difficult task than it may seem at first glance:

- controlled indexing on single-words must account for polysemy and word disambiguation (Krovetz and Croft, 1992; Klavans, 1995).
- controlled indexing on multi-word terms must consider the numerous forms of term variations (Dunham, Pacak, and Pratt, 1978; Sparck Jones and Tait, 1984; Jacquemin, 1996).

We focus here on the multi-word task. Our system exploits a morphological processor and a transformation-based parser for the extraction of multi-word controlled indexes.

The action of the system is twofold. First, a corpus is enriched by tagging each word unambiguously, and then expanded by linking each word with all its possible derivatives. For example, for English, the word *genes* is tagged as a plural noun and morphologically connected to *genic*, *genetic*, *genome*, *genotoxic*, *genetically*, etc. Second, the term list is dynamically expanded through syntactic transformations which allow the retrieval of term variants. For example, *genic expressions*, *genes were expressed*, *expression of this gene*, etc. are extracted as variants of *gene expression*.

This system relies on a full-fledged unification formalism and thus is well adapted to a fine-grained identification of terms related in syntactically and morphologically complex ways. The same system has been effectively applied both to English and French, although this paper focuses on French (see (Jacquemin, 1994) for the case of syntactic variants in English). All evaluation experiments were performed on two corpora: a training corpus [ECI] (ECI, 1989 and 1990) used for the tuning of the metagrammar and a test corpus [AGR] (AGR, 1995) used for evaluation. [ECI] is a subset of the *European Corpus Initiative* data composed of 1.3 million words of the French newspaper "Le Monde"; [AGR] is a set of abstracts of scientific papers in the agricultural domain from *INIST/CNRS* (1.1 million words). A list of terms is associated with each corpus: the terms corresponding to [ECI] were automatically extracted by *LEXTER* (Bourigault, 1993) and the terms corresponding to [AGR] were extracted from the *AGROVOC* term list owned by *INIST/CNRS*.

The following section describes methods for grouping multi-word term variants; Section 4 presents a linguistically-motivated method for lexical analysis (inflectional analysis, part of speech tagging, and derivational analysis); Section 5 explains term expansion methods: constructions with a local parse through syntactic transformations preserving dependency relations; Section 6 illustrates the empirical tuning of linguistic rules; Section 7 presents an evaluation of the results in terms of precision and recall.

3 Variation in Multi-Word Terms: A Description of the Problem

Linguistic variation is a major concern in the studies on automatic indexing. Variations can be classified into three major categories:

- **Syntactic (Type 1)**: the content words of the original term are found in the variant but the syntactic structure of the term is modified, e.g. *technique for performing volumetric measurements* is a Type 1 variant of *measurement technique*.
- **Morpho-syntactic (Type 2)**: the content words of the original term or one of their derivatives are found in the variant. The syntactic structure of the term is also modified, e.g. *electrophoresed on a neutral polyacrylamide gel* is a Type 2 variant of *gel electrophoresis*.
- **Semantic (Type 3)**: synonyms are found in the variant; the structure may be modified, e.g. *kidney function* is a Type 3 variant of *renal function*.

This paper deals with Type 1 and Type 2 variations.

The two main approaches to multi-word term conflation in IR are *text simplification* and *structural similarity*. Text simplification refers to traditional IR algorithms such as (1) deletion of stop words, (2) normalization of single words through stemming, and (3) phrase construction through dictionary matching. (See (Lewis, Croft, and Bhandaru, 1989; Smeaton, 1992) on the exploitation of NLP techniques in IR.) These methods are generally limited. The morphological complexity of the language seems to be a decisive argument for performing rich stemming (Popovič and Willett, 1992). Since we focus on French, a language with a rich declensional inflectional and derivational morphology—we have chosen the richest and most precise morphological analysis. This is a key component in the recognition of Type 2 variants. For structural similarity, coarse dependency-based NLP methods do not account for fine structural relations involved in Type 1 variants. For instance, *properties of flour* should be linked to *flour properties*, *properties of wheat flour* but not to *properties of flour starch* (examples are from (Schwarz, 1990)). The last occurrence must be rejected because *starch* is the argument of the head

noun *properties*, whereas *flour* is the argument of the head noun *properties* in the original term. Without careful structural disambiguation over internal phrase structure, these important syntactic distinctions would be incorrectly overlooked.

4 Part of Speech Disambiguation and Morphology

First, **inflectional morphology** is performed in order to get the different analyses of word forms. Inflectional morphology is implemented with finite-state transducers on the model used for Spanish (Tzoukermann and Liberman, 1990). The theoretical principles underlying this approach are based on generative morphology (Aronoff, 1976; Selkirk, 1982). The system consists of precomputing stems, extracted from a large dictionary of French (Boyer, 1993) enhanced with newspaper corpora, a total of over 85,000 entries.

Second, a **finite-state part of speech tagger** (Tzoukermann, Radev, and Gale, 1995; Tzoukermann and Radev, 1996) performs the morpho-syntactic disambiguation of words. The tagger takes the output of inflectional morphological analysis and through a combination of linguistic and statistical techniques, outputs a unique part of speech for each word in context. Reducing the ambiguity of part of speech tags eliminates ambiguity in local parsing. Furthermore, part of speech ambiguity resolution permits construction of correct derivational links.

Third, **derivational morphology** (Tzoukermann and Jacquemin, 1997) is achieved to generate morphological variants of the disambiguated words. Derivational generation is performed on the lemmas produced by the inflectional analysis and the part of speech information. Productive stripping and concatenation rules are applied on lemmas.

The derived forms are expressed as tokens with feature structures[1]. For instance, the following set of constraints express that the noun *modernisateur* is morphologically related to the word *modernisation*[2]. The <ON> metarule removes the *-ion* suffix, and the <EUR> rule adds the nominal suffix *-eur*.

<cat> = N
<lemma> = 'modernisation'
<reference> = 52663
<derivation cat> = N
<derivation lemma> = 'modernisateur'
<derivation reference> = 52662
<derivation history> = '<ON<>EUR>'.

The morphological analysis performed in this study is detailed in (Tzoukermann, Klavans, and Jacquemin, 1997). It is more complete and linguistically more accurate than simple stemming for the following reasons:

- Allomorphy is accounted for by listing the set of its possible allomorphs for each word. Allomorphies are obtained through multiple verb stems, e.g. *fabriqu-*, *fabric-* (fabricate) or additional allomorphic rules.
- Concatenation of several suffixes is accounted for by rule ordering mechanisms. Furthermore, we have devised a method for guessing possible suffix combinations from a lexicon and a corpus. This empirical method reported in (Jacquemin, 1997) ensures that suffixes which are related within specific domains are considered.
- Derivational morphology is built with the perspective of overgeneration. The nature of the semantic links between a word and its derivational forms is not checked and all allomorphic alternants are generated. Selection of the correct links occurs during subsequent term expansion process with collocational filtering. Although *étable* (cowshed) is incorrectly related to *établir* (to establish), it is very improbable to find a context where *établir* co-occurs with one of the three words found in the three multi-word terms containing *étable*: *nettoyeur* (cleaner), *alimentation* (feeding), and *litière* (litter). Since we focus on multi-word term variants, overgeneration does not present a problem in our system.

5 Transformation-Based Term Expansion

The extraction of terms and their variants from corpora is performed by a unification-based parser. The controlled terms are transformed into grammar rules whose syntax is similar to *PATR-II*.

5.1 A Corpus-Based Method for Discovering Syntactic Transformations

We present a method for inferring transformations from a corpus in the purpose of developing a gram-

[1] In the remainder of the paper, N is Noun, A Adjective, C Coordinating conjunction, D Determiner, P Preposition, Av Adverb, Pu Punctuation, NP Noun Phrase, and AP Adjective Phrase.

[2] Each lemma has a unique numeric identifier <reference>.

mar of syntactic transformations for term variants. To discover the families of term variants, we first consider a notion of *collocation* which is less restrictive than variation. Then, we refine this notion in order to filter out genuine variants and to reject spurious ones. A Type 1 collocation of a binary term is a text window containing its content words w_1 and w_2, without consideration of the syntactic structure. With such a definition, any Type 1 variant is a Type 1 collocation. Similarly, a notion of Type 2 collocation is defined based on the co-occurence of w_1 and w_2 including their derivational relatives.

A ±5-word window is considered as sufficient for detecting collocations in English (Martin, Al, and Van Sterkenburg, 1983). We chose a window-size twice as large because French is a Romance language with longer syntactic structures due to the absence of compounding, and because we want to be sure to observe structures spanning over large textual sequences. For example, the term *perte au stockage* (storage loss) is encountered in the [AGR] corpus as: *pertes occasionnées par les insectes au sorgho stocké* (literally: loss of stored sorghum due to the insects).

A linguistic classification of the collocations which are correct variants brings up the following families of variations[3].

- **Type 1 variations** are classified according to their syntactic stucture.
  1. **Coordination:** a coordination the combination of two terms with a common head word or a common argument. Thus, *fruits et agrumes tropicaux* (literally: tropical citrus fruits or fruits) is a coordination variant of the term *fruits tropicaux* (tropical fruits).
  2. **Substitution/Modification**: a substitution is the replacement of a content word by a term; a modification is the insertion of a modifier without reference to another term. For example, *activité thermodynamique de l'eau* (thermodynamic activity of water) is a substitution variant of *activité de l'eau* (activity of water) if *activité thermodynamique* (thermodynamic activity) is a term; otherwise, it is a modification.
  3. **Compounding/Decompounding**: in French, most terms have a compound noun structure, i.e. a noun phrase structure where determiners are omitted such as *consommation d'oxygène* (oxygen consumption). The decompounding variation is the transformation of a term with a compound structure into a noun phrase structure such as *consommation de l'oxygène* (consumption of the oxygen). Compounding is the reciprocal transformation.
- **Type 2 variations** are classified according to the nature of the morphological derivation. Often semantic shifts are involved as well (Viegas, Gonzalez, and Longwell, 1996).
  1. **Noun-Noun variations**: relations such as result/agent (*fixation de l'azote* (nitrogen fixation)/*fixateurs d'azote* (nitrogen fixater)) or container/content (*réservoir d'eau* (water reservoir)/*réserve en eau* (water reserve)) are found in this family.
  2. **Noun-Verb variations**: these variations often involve semantic shifts such as process/result *fixation de l'azote*/*fixer l'azote* (to fix nitrogen).
  3. **Noun-Adjective variations**: the two ways to modify a noun, a prepositional phrase or an adjectival phrase, are generally semantically equivalent, e.g. *variation du climat* (climate variation) is a synonym of *variation climatique* (climatic variation).

A method for term variant extraction based on morphology and simple co-occurrences would be very imprecise. A manual observation of collocations shows that only 55% of the Type 1 collocations are correct Type 1 variants and that only 52% of the Type 2 collocations are correct Type 2 variants. It is therefore necessary to conceive a filtering method for rejecting fortuitous co-occurrences. The following section proposes a filtering system based on syntactic patterns.

6 Empirical Rule Tuning

6.1 Syntactic Transformations for Type 1 and Type 2 variants

The concept of a grammar of syntactic transformations is motivated by well-known observations on the behavior of collocations in context (e.g. (Harris et al., 1989).) Initial rules based on surface syntax are refined through incremental experimental tuning.

We have devised a grammar of French to serve as a basis for the creation of metarules for term variants. For example, the noun phrase expansion rule is[4]:

$$NP \rightarrow D^{?}\ AP^{*}\ N\ (AP \mid PP)^{*} \qquad (1)$$

[3] *Variations* are generic linguistic functions and *variants* are transformations of terms by these functions.

[4] We use UNIX regular expression symbols for rules and transformations.

From this rule a set of expansions can be generated:

$$NP = D^? \ (Av^? \ A)^* \ N \ (Av^? \ A \mid P \ D^? \ (Av^? \ A)^* \ N \ (Av^? \ A)^*)^* \quad (2)$$

In order to balance completeness and accuracy, expansions are limited. After the initial expansion is created for a range of structures, empirical tuning is applied to create a set of maximum coverage metarules.

We briefly illustrate this process for coordination. For this example, we restrict transformations to terms with N P N structures which represent a full 33% of the binary terms. Examples of metarules of Type 1 and Type 2 variations are given in Table 1.

6.2 Development of a Coordination Transformation for N P N Terms

The coordination types are first calculated by combining the pattern $N_1 \ P_2 \ N_3$ with possible expansions of a noun phrase with a simple paradigmatic structure $A^? \ N \ (A \mid P \ D^? \ A^? \ N \ A^?)$[5]:

$$Coord(N_1 \ P_2 \ N_3) = N_1 \ ((C \ A^? \ N \ A^? \ P) \mid (A \ C \ P) \mid (P \ D^? \ A^? \ N \ A^? \ C \ P^?)) \ N_3 \quad (3)$$

The first parenthesis (C $A^?$ N $A^?$ P) represents a coordinated head noun, the second (A C P) and third (P $D^?$ $A^?$ N $A^?$ C $P^?$) represent respectively an adjective phrase and a prepositional phrase coordinated with the prepositional phrase of the original term.

Variants were extracted on the [ECI] corpus through this transformation; the following observations and changes have been made.

First, coordination accepts a substitution which replaces the noun N_3 with a noun phrase $D^? \ A^? \ N_3$. For example, the variant *température et humidité initiale de* **l'***air* (temperature and initial humidity of **the** air) is a coordination where a determiner precedes the last noun (*air*).

Secondly, the observations of coordination variants also suggest that the coordinating conjunction can be preceded by an optional comma and followed by an optional adverb, e.g. *la production*, *et* **surtout** *la diffusion des semences* (the production, and **particularly** the distribution of the seeds).

Thirdly, variants such as *de l'humidité et* **de la** *vitesse de l'air* (literally: of humidity and **of the** speed of the air) indicate that the conjunction can be followed by an optional preposition and an optional determiner.

[5]Subscripts represent indexing.

The three preceding changes are made on the expression of (3) and the resulting transformation is given in the first line of Table 1 (changes are underlined).

Our empirical selection of valid metarules is guided by linguistic considerations and corpus observations. This mode of grammar conception has led us to the following decisions:

- reject linguistic phenomena which could not be accounted for by regular expressions such as sentential complements of nouns;
- reject noisy and inaccurate variations such as long distance dependencies (specifically within a verb phrase);
- focus on productive and safe variations which are felicitously represented in our framework.

Accounting for variants which are not considered in our framework would require the conception of a novel framework, probably in cooperation with a deeper analyzer. It is unlikely that our transformational approach with regular expressions could do much better than the results presented here. Table 2 shows some variants of *AGROVOC* terms extracted from the [AGR] corpus.

7 Evaluation

The precision and recall of the extraction of term variants are given in Table 4 where precision is the ratio of correct variants among the variants extracted and the recall is the ratio of variants retrieved among the collocates. Results were obtained through a manual inspection of 1,579 Type 1 variants, 823 Type 2 variants, 3,509 Type 1 collocates, and 2,104 Type 2 collocates extracted from the [AGR] corpus and the *AGROVOC* term list.

These results indicate a very high level of accuracy: 89.4% of the variants extracted by the system are correct ones. Errors generally correspond to a semantic discrepancy between a word and its morphologically derived form. For example, *élevée pour un sol* (literally: high for a soil) is not a correct variant of *élevage hors sol* (off-soil breeding) because *élevée* and *élevage* are morphologically related to two different senses of the verb *élever*: *élevée* derives from the meaning *to raise* whereas *élevage* derives from *to breed.* Recall is weaker than precision because only 75.2% of the possible variants are retrieved.

Improvement of Indexing through Variant Extraction

For a better understanding of the importance of term expansion, we now compare term indexing with

Table 1: Metarules of Type 1 (Coordination) and Type 2 (Noun to Verb) Variations.

Variation	Term and variant
Coord(N_1 P_2 N_3) = N_1 (((Pu? C Av? P? D? A? N A? P) | (A C Av? P) | (P D? A? N A? C Av? P?)) D? A?) N_3.	*teneur en protéine* (protein content) → *teneur en eau et en protéine* (protein and water content)
NtoV(N_1 P_2 N_3) = V_1 ($Av^?$ ($P^?$ D | P) $A^?$) N_3: <V_1 derivation reference> = <N_1 reference>.	*stabilisation de prix* (price stabilization) → *stabiliser leurs prix* (stabilize their prices)

Table 2: Examples of Variations from [AGR].

Term	Variant	Type
Échange d'ion (ion exchange)	*échange ionique* (ionic exchange)	N to A
Culture de cellules (cell culture)	*cultures primaires de cellules* (primary cell cultures)	Modif.
Propriété chimique (chemical property)	*propriétés physiques et chimiques* (chemical and physical properties)	Coor.
Gestion d'eau (water management)	*gestion de l'eau* (management of the water)	Comp.
Eau de surface (surface water)	*eau et de l'évaporation de surface* (water and of surface evaporation [incorrect variant])	Coor.
Huile de palme (palm oil)	*palmier à huile* (palm tree [yielding oil])	N to N
Initiation de bourgeon (bud initiation)	*initier des bourgeons* (initiate buds)	N to V

and without variant expansion. The [AGR] corpus has been indexed with the *AGROVOC* thesaurus in two different ways:

1. Simple indexing: Extraction of occurrences of multi-word terms without considering variation.
2. Rich indexing: Simple indexing improved with the extraction of variants of multi-word terms.

Both indexings have been manually checked. Simple indexing is almost error-free but does not cover term variants. On the contrary, rich indexing is slightly less accurate but recall is much higher. Both methods are compared by calculating the effectiveness measure (Van Rijsbergen, 1975):

$$E_\alpha = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)} \text{ with } 0 \leq \alpha \leq 1 \quad (4)$$

P and R are precision and recall and α is a parameter which is close to 1 if precision is preferred to recall. The value of E_α varies from 0 to 1; E_α is close to 0 when all the relevant conflations are made and when no incorrect one is made.

The effectiveness of rich indexing is more than three times better than effectiveness of simple indexing. Retrieved variants increase the number of indexing items by 28.8% (17.3% Type 1 variants and 11.5% Type 2 variants). Thus, term variant extraction is a significant expansion factor for identifying morphologically and syntactically related multi-word terms in a document without introducing undesirable noise.

Table 3: Evaluation of Simple vs. Rich Indexing.

	Precision	Recall	$E_{0.5}$
Simple indexing	99.7%	72.4%	16.1%
Rich indexing	97.2%	93.4%	4.7%

As for performance, the parser is fast enough for processing large amounts of textual data due to the presence of several optimization devices. On a Pentium133 with Linux, the parser processes 18,100 words/min from an initial list of 4,300 terms.

Conclusion

This paper has proposed a syntax-based approach via morphologically derived forms for the identification and extraction of multi-word term variants.

Table 4: Precision and Recall of Term Variant Extraction on [AGR]

	Type 1 variants			Type 2 variants				**Total**
	Subst.	Coord.	Comp.	AtoN	NtoA	NtoN	NtoV	
# correct	808	228	404	19	60	273	471	2263
# rejected	87	26	26	7	5	28	90	269
Precision	90.3%	90.0%	94.0%	73.1%	91.6%	93.0%	84.0%	**89.4%**
	91.2%			86.4%				
Recall	75.0%			75.6%				**75.2%**

In using a list of controlled terms coupled with a syntactic analyzer, the method is more precise than traditional text simplification methods. Iterative experimental tuning has resulted in wide-coverage linguistic description incorporating the most frequent linguistic phenomena.

Evaluations indicate that, by accounting for term variation using corpus tagging, morphological derivation, and transformation-based rules, 28.8% more can be identified than with a traditional indexer which cannot account for variation. Applications to be explored in future research involve the incorporation of the system as part of the indexing module of an IR system, to be able to accurately measure improvements in system coverage as well as areas of possible degradation. We also plan to explore analysis of semantic variants through a predicative representation of term semantics. Our results so far indicate that using computational linguistic techniques for carefully controlled term expansion will permit at least a three-fold expansion for coverage over traditional indexing, which should improve retrieval results accordingly.

References

AGR, Institut National de l'Information Scientifique et Technique, Vandœuvre, France, 1995. *Corpus de l'Agriculture*, first edition.

Aronoff, Mark. 1976. *Word Formation in Generative Grammar.* Linguistic Inquiry Monographs. MIT Press, Cambridge, MA.

Bourigault, Didier. 1993. An endogeneous corpus-based method for structural noun phrase disambiguation. In *Proceedings, 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, pages 81–86, Utrecht.

Boyer, Martin. 1993. *Dictionnaire du français.* Hydro-Quebec, GNU General Public License, Québec, Canada.

Daille, Béatrice. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* MIT Press, Cambridge, MA.

Dunham, George S., Milos G. Pacak, and Arnold W. Pratt. 1978. Automatic indexing of pathology data. *Journal of the American Society for Information Science*, 29(2):81–90.

ECI, European Corpus Initiative, 1989 and 1990. *"Le Monde" Newspaper.*

Grefenstette, Gregory and Simone Teufel. 1995. Corpus-based method for automatic identification of support verbs for nominalizations. In *Proceedings, 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95)*, pages 98–103, Dublin.

Harris, Zellig S., Michael Gottfried, Thomas Ryckman, Paul Mattick Jr, Anne Daladier, T. N. Harris, and S. Harris. 1989. *The Form of Information in Science, Analysis of Immunology Sublanguage*, volume 104 of *Boston Studies in the Philosophy of Science.* Kluwer, Boston, MA.

Jacquemin, Christian. 1994. Recycling terms into a partial parser. In *Proceedings, 4th Conference on Applied Natural Language Processing (ANLP'94)*, pages 113–118, Stuttgart.

Jacquemin, Christian. 1996. What is the tree that we see through the window: A linguistic approach to windowing and term variation. *Information Processing & Management*, 32(4):445–458.

Jacquemin, Christian. 1997. Guessing morphology from terms and corpora. In *Proceedings, 20th*

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), Philadelphia, PA.

Justeson, John S. and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Klavans, Judith L., editor. 1995. *AAAI Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity.* American Association for Artificial Intelligence, March.

Klavans, Judith L. and Martin S. Chodorow. 1992. Degrees of stativity: The lexical representation of verb aspect. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 1126–1131, Nantes, France.

Krovetz, Robert and W. Bruce Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

Lewis, David D., W. Bruce Croft, and Nehru Bhandaru. 1989. Language-oriented information retrieval. *International Journal of Intelligent Systems*, 4:285–318.

Martin, W.J.F., B.P.F. Al, and P.J.G. Van Sterkenburg. 1983. On the processing of a text corpus: From textual data to lexicographical information. In R.R.K. Hartman, editor, *Lexicography, Principles and Practice.* Academic Press, London, pages 77–87.

Metzler, Douglas P. and Stephanie W. Haas. 1989. The Constituent Object Parser: Syntactic structure matching for information retrieval. *ACM Transactions on Information Systems*, 7(3):292–316.

NLM, National Library of Medicine, Bethesda, MD, 1995. *Unified Medical Language System*, sixth experimental edition.

Popovič, Mirko and Peter Willett. 1992. The effectiveness of stemming for Natural-Language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5):384–390.

Schwarz, Christoph. 1990. Automatic syntactic analysis of free text. *Journal of the American Society for Information Science*, 41(6):408–417.

Selkirk, Elisabeth O. 1982. *The Syntax of Words.* MIT Press, Cambridge, MA.

Sheridan, Paraic and Alan F. Smeaton. 1992. The application of morpho-syntactic language processing to effective phrase matching. *Information Processing & Management*, 28(3):349–369.

Smadja, Frank and Kathleen R. McKeown. 1991. Using collocations for language generation. *Computational Intelligence*, 7(4), December.

Smeaton, Alan F. 1992. Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal*, 35(3):268–278.

Sparck Jones, Karen and Joel I. Tait. 1984. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66.

Srinivasan, Padmini. 1996. Optimal document-indexing vocabulary for Medline. *Information Processing & Management*, 32(5):503–514.

Strzalkowski, Tomek. 1996. Natural language information retrieval. *Information Processing & Management*, 31(3):397–417.

Tzoukermann, Evelyne and Christian Jacquemin. 1997. Analyse automatique de la morphologie dérivationnelle et filtrage de mots possibles. *Silexicales*, 1:251–260. Colloque Mots possibles et mots existants, SILEX, University of Lille III.

Tzoukermann, Evelyne, Judith L. Klavans, and Christian Jacquemin. 1997. Effective use of natural language processing techniques for automatic conflation of multi-word terms: the role of derivational morphology, part of speech tagging, and shallow parsing. In *Proceedings, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, Philadelphia, PA.

Tzoukermann, Evelyne and Mark Y. Liberman. 1990. A finite-state morphological processor for Spanish. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, pages 277–281, Helsinki, Finland.

Tzoukermann, Evelyne and Dragomir R. Radev. 1996. Using word class for part-of-speech disambiguation. In *SIGDAT Workshop*, pages 1–13, Copenhagen, Denmark.

Tzoukermann, Evelyne, Dragomir R. Radev, and William A. Gale. 1995. Combining linguistic knowledge and statistical learning in French part-of-speech tagging. In *EACL SIGDAT Workshop*, pages 51–57, Dublin, Ireland.

Van Rijsbergen, C. J. 1975. *Information Retrieval.* Butterworth, London.

Viegas, Evelyne, Margarita Gonzalez, and Jeff Longwell. 1996. Morpho-semantics and constructive derivational morphology: A transcategorial approach. Technical Report MCCS-96-295, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.