19. TERM EXTRACTION AND AUTOMATIC INDEXING

Abstract

This chapter presents a new domain of research and development in Natural Language Processing (NLP) that is concerned with the representation, acquisition, and recognition of terms.

Terms are pervasive in scientific and technical documents; their identification is a crucial issue for any application dealing with the analysis, understanding, generation, or translation of such documents. In particular, the ever-growing mass of specialized documentation available on-line, in industrial and governmental archives or in digital libraries, calls for advances in terminology processing for such purposes as information retrieval, cross-language querying, indexing of multimedia documents, translation aids, document routing and summarization, etc.

This chapter introduces the basic linguistic characteristics of terms. It presents the main methods in NLP for recognizing or discovering terms and their interrelationships in large corpora. It is divided into three sections: an introduction to the basic concepts of this field (Section 19.1), the description of some significant studies in term-oriented NLP (Section 19.2), and the presentation of the most promising areas for the next few years (Section 19.3).

19.1 Basic Notions

In this section, the major notions of computational term processing are introduced together with the corresponding activities being developed in this domain. Terms are presented together with data: on the one hand, rich contexts in corpora and, on the other, management data in thesauri (conceptual links, linguistic variants, and usage restrictions). Terms are divided into single-word terms and multi-word terms. The former are ambiguous and call for context-based disambiguation (see Chapter *Word-sense Disambiguation*), while the latter are prone to a wide range of variations.

19.1.1 What is a Term?

The definition of what constitutes a term proposed in computational approaches to term identification differs from the traditional notion of terms as elaborated by the Vienna School. The characterization of terms in a computational framework must take into account novel dimensions of termhood in relation with the applied purposes of terminological engineering.

The Classical View

In the traditional sense, a term is considered as the linguistic label of a concept. This dominant approach to termhood stems from the General Theory of Terminology which was elaborated by E. Wüster in the late 30s, in relation with the Vienna Circle (Felber 1984). Born in the positivist movement during the interwar period, the classical doctrine of terminology relies on a unifying view of knowledge. It assumes that knowledge is organized into domains, each domain being equivalent to a network of concepts. In one domain, each concept is (ideally) associated with one term, which is its linguistic label. Such a concept-centered approach to terminology is well-suited for normalization. It is however less adapted to a computational approach to term analysis.

Problem with the Applicability of the Classical View

First of all, the classical view assumes that experts in an area of knowledge have conceptual maps in their minds. This assumption is misleading and unproductive because experts cannot build a conceptual map from introspection. Terminologists constantly refer to textual data and analyze the lexical elements in order to acquire and validate a conceptual description.

Secondly, the semi-automatic construction and exploitation of terminological resources gives rise to a wide variety of terminological data for the same field. There are as many types of resources as types of applications: thesauri for automatic indexing and information retrieval, structured indices for hyper-documents, authority lists for computer-aided controlled writing, bilingual term lists for computer-aided translation, ontologies for industrial intelligence, structured keywords for digital libraries, etc. Failure of introspection and polymorphy of term banks suggest that the classical view of terminology is not well-adapted to empirical applications.

Terminological Engineering

In a definition of term that is better suited to corpus-based terminology, a term must be stated as the *output* of a procedure of terminological analysis. A single word, such as *cell*, or a multi-word unit, such as *blood cell* is a term because it has been decided that it would be so. The decision process can involve a community of researchers or practitioners, a normalization institution, or even a single engineer or terminologist in charge of building a terminological resource for a specific purpose.

Building a terminological resource should be viewed as the *construction* of *an instance of* the terminological structures from a corpus, not as the *discovery* of *the unique* conceptual representation of a given area. The resulting terminological database should be dually relevant:

- wrt the corpus: it should be made of stable and domain-specific lexical items; and
- wrt the application: it should contain units which are useful for the intended application in terms of economy, internal cohesion, and efficiency.

Thus, building a terminology pertains to engineering as outlined by (Sager 1990, p. 10):

> *The theories underlying applied fields of study benefit from being application driven rather than following separate paths as terminological theory has been doing in recent years. By adopting the engineering approach of identifying problems and seeking solutions, significant advances have been made (...)*

Sager's characterization of terminology construction has the merit of fitting the practical contemporary way of building terminologies through computer assistance. It also provides a common denominator for terminology processing in computational terminology, information retrieval, or information extraction.

19.1.2 Linguistic Work on Terms in Context

Recently, several linguists in terminology have focused on the notion of *rich contexts* which are involved in the detection of terms, relations between terms, or definitions and properties of terms (Condamines & Rebeyrolles 1998; Davidson *et al.* 1998; Pearson

1998). Even though these works were developed independently, they agree on the fact that tools for term acquisition should recognize and use terminological contexts.

In rich contexts relationships are established between terms "[*knowledge rich contexts are*] *contexts that illustrate domain-specific conceptual relations*" (Davidson *et al.* 1998, p. 50). In such contexts, definitions are proposed for a characterization of terms through formal or semi-formal defining expositive (Pearson 1998). The following examples are formal expositives: *Compost* ***is*** *the controlled decomposition of organic matter through biological process* (Davidson *et al.* 1998), *The periods in which data are accumulated* ***are called*** *test periods* (Pearson 1998), or *A configuration file* ***is*** *a family of source files* (Condamines & Rebeyrolles 1998). Connective verbs are frequently used in semi-formal expositives, for instance *used to*, *known as*, etc.: *Major basic protein has been localized to placental trophoblasts* ***known as*** *X cells.*

In a less rigorous manner, connective phrases such as *e.g.* or *called* provide paraphrastic equivalents or synonyms of terms in texts: *Animals* [*are*] *exposed to one of several non-genotoxic h epatocarcinogens,* ***e.g.*** *2,3,7,8-tetra-chlorodibenzo-p-dioxin, carbon tetrachloride, peroxisome proliferators and choline-devoid diet.* Thus, the hypothesis underlying the use of linguistically-rich contexts is that the expression of terminological relationships in texts is made through cue words or structures (*linguistic patterns* for Davidson *et al.* (1998), *linguistic signals* and *connective phrases* for Pearson (1998), and *markers* for Condamines & Rebeyrolles (1998)). The cue words are in bold in the examples given in this section.

Rich contexts can be used in an incremental way to acquire terms and their relationships. A first list of cue words is used to identify pairs of terms from corpora. Then the analysis of the occurrences of these terms is used to acquire other relations and a second list of patterns and associated cue words from the corpus. The lexico-syntactic patterns are progressively refined through an analysis of the occurrences in terms of recall and precision. Finally patterns may include semantic features, typographic parameters (bold, italics, color), layout and positional parameters (head of paragraph, head of section), lexical items such as *class words* (*process*, *method*, *function*). The following example is a metalinguistic pattern adapted from (Pearson 1998):

$$\{\text{indefinite article}\}^{?} + \{\text{term}\} + \{\text{connective verbs } (\textit{is a},\ \textit{consists}\ldots)\} + \quad (19.1)$$
$$\{\text{indefinite article} \vee \text{definite article}\}^{?} + \{\text{term} \vee \text{class word}\} + \{\text{past participle}\}$$

It leads to the acquisition of occurrences such as *A misdelivered frame* ***is*** *a frame transferred...*

Because they are only found within corpora, such patterns can only be defined through linguistic studies with a genuine concern for the detailed observation of textual data and not through blind language analyzers. Symmetrically, since the patterns defined at the linguistic level emerge from accurate and cautious analysis of corpus occurrences, they are likely to be useful to the design of tools for computer-aided terminology.

19.1.3 Terms in Thesauri

Among the terminological resources that can result from terminological investigation, thesauri are particularly interesting for computational applications such as NLP for specialized languages, automatic indexing, or automatic acquisition of terms and their relationships. For this reason, we have decided to present in detail the organization of a thesaurus in this section.

While terms in corpora are dependent on the context in which they are employed, terms in thesauri are autonomous units with attached information. According to (Sager 1990), the data used to describe terms can be classified into the following five categories: management data (numeric keys, record number, terminologist's name, date of coding, etc), conceptual data (subject, scope, definition, related concepts and type of relation, etc), linguistic data (lexical entries, synonymous entries, equivalents in other languages, variants, etc), pragmatic data (usage restrictions, contextual data, etc), and bibliographical data.

In order to illustrate the way in which a term can be documented in a thesaurus, we now turn to the description of the *Metathesaurus* in the *UMLS* project. As indicated above, such a thesaurus is not an absolute representation of the medical world. The *Metathesaurus* is designed to help the processing of medical data by providing practitioners with an authority list of concepts. The purpose of this project is to facilitate the information retrieval and the integration of all the linguistic data handled in the medical domain such as biomedical literature, clinical records, factual databanks, etc. There are three *UMLS* knowledge sources: a *Metathesaurus* which contains semantic information about the concepts and semantic relationships among them, the *Semantic Network*, an ontology of the categories and semantic types encountered in the *Metathesaurus*, and the *Specialist Lexicon* which contains syntactic information about the terms and covers all the terms encountered in the *Metathesaurus*. Thus, the *Metathesaurus* appears to be a central component of the project for it provides the meanings, the hierarchical connections, and other relationships between the concepts in the source vocabularies.

The *Metathesaurus* is organized into concepts. Each concept is an abstract representation of linguistic utterances which are considered as synonymous in the medical domain. Each concept is linked to several terms (or to only one term in the case where the term has no variant form). One of them is the preferred term according to usage observations in the vocabularies. Terms are in turn linked to alternative strings (including graphical variants, lexical variants, translations in other languages). For each term, one of the string is also considered as preferred. Table 19.1 illustrates one concept which is associated with two terms *Atrial fibrillation* and *Auricular fibrillation*. Each of these terms is in turn linked to two strings the singular—the preferred string—and the plural variant.

The *Metathesaurus* also contains several types of semantic relationships between concepts. For instance, *Atrial Fibrillation* is a type of *Arrhythmia* and is therefore linked to it by an `is-a` link. *Atrial Fibrillation* is in turn more generic than *Paroxysmal Atrial Fibrillation* which is related to it by a `narrower-than` link.

As indicated above, the major role of a thesaurus is to provide the user with details about the concepts it contains. In the *Metathesaurus*, several types of attribute are associated with concepts such as terms and strings. For instance, the concept *Atrial Fibrillation* shown in Table 19.1 has the semantic types *Pathologic Function* and *Finding* and the definition *Disorder of cardiac rhythm characterized by rapid, irregular atrial impulses and ineffective atrial contractions.*

The semantic types associated with concepts are organized into a *Semantic network* of 134 types and 54 relationships. Each concept is represented by the most specific type in the network which is linked to it. Figure 19.1 represents a fragment of a hierarchy of types in the *Semantic network*. As shown in Figure 19.2, relationships between concepts can also be organized into a hierarchical network.

Such a thesaurus can be used by a specialist in order to access information, control the quality of writing, get translation aids, etc. But, recently, thesauri have also proven to

Concepts	**Terms**	**Strings**
C0004238	L0004238	S0016668
(preferred) *Atrial Fibrillation* *Atrial Fibrillations* *Auricular Fibrillation*	(preferred) *Atrial Fibrillation* *Atrial Fibrillations*	(preferred) *Atrial Fibrillation*
		S0016669
		(plural variant) *Atrial Fibrillations*
	L0004327	S0016899
	(synonym) *Auricular Fibrillation* *Auricular Fibrillations*	(preferred) *Auricular Fibrillation*
		S0016900
		(plural variant) *Auricular Fibrillations*

Table 19.1: A sample concept in the *Metathesaurus* and associated terms and strings.

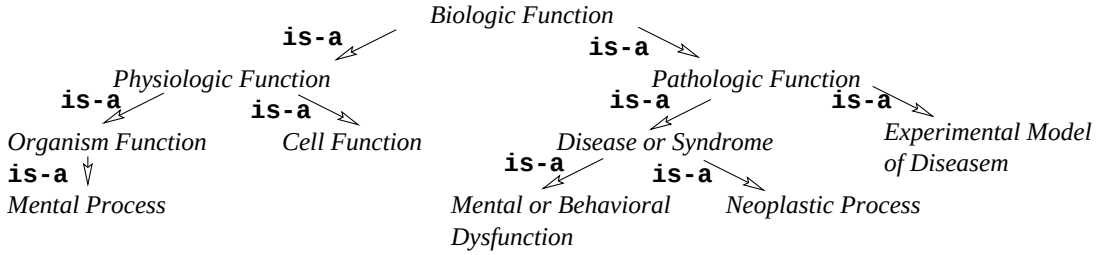

Figure 19.1: Is-a relationships among types of concepts in the *Semantic network*.

be useful resources in automatic text processing for the purpose of NLP and information retrieval (see Chapter *Ontologies in NLP*). Conversely, NLP and knowledge engineering techniques are developed for the automatic construction of thesauri. We now turn to the presentation of some tools for computational terminology at the crossroad of NLP and terminology, namely term acquisition (NLP for knowledge engineering) and automatic indexing (NLP for information retrieval).

19.2 Term-oriented NLP

The two main activities involving terminology in NLP are term acquisition, the automatic discovery of new terms, and term recognition, the identification of known terms

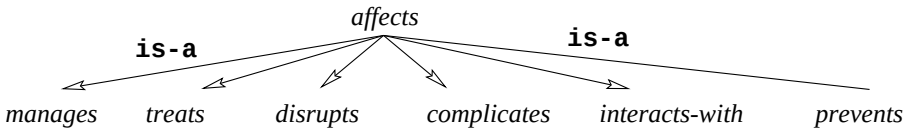

Figure 19.2: Is-a relationships among types of relationships in the *Semantic network*.

	Prior terminological data	No prior terminological data
Term discovery	*Term enrichment*	*Term acquisition*
Term recognition	*Controlled indexing*	*Free indexing*

Table 19.2: The four main sub-domains of term-based NLP.

in text corpora. According to this subdivision of the activities in computational terminology, a dozen systems are sketched out in this section. Half of them pertain to term acquisition (*ACABIT*, *LEXTER*, *TERMS*, *XTract,* etc.) and a half of them pertain to the recognition of terms, mainly for the purpose of automatic indexing, (*Constituent Object Parser*, *COPSY*, *FASIT,* etc.). Before presenting these systems, we first introduce more precisely the various trends in computational terminology.

19.2.1 Fields of Activity in Term-oriented NLP

The use of terms in NLP concerns either industrial applications of NLP, such as automatic translation, information retrieval, information extraction and knowledge management, or language planning, the description and the collection of terms in a given language and in a given area. Early developments of term acquisition, such as *TERMINO* (David & Plante 1990; Lauriston 1994), were motivated by terminography, an activity politically and linguistically crucial for a French speaking region in Canada. Recent research in term acquisition and term recognition are now mainly motivated by industrial applications, among which information retrieval and knowledge management are certainly the most prominent fields.

Basically, terms in NLP can be viewed as a particular type of lexical data. Contrary to general language lexicons, term bases contain mostly multi-word units and are prone to continuous evolutions: creation, modification, semantic shifts, neologisms, etc. Therefore term databases need to be constantly rebuilt, maintained and enriched in order to follow the thematic drifts in scientific and technical areas. Thus, a major domain of term-oriented NLP and term-oriented statistics is *term acquisition.* It is subdivided into two subfields, *initial term acquisition* and *term enrichment*, depending on whether or not initial terminological data are possessed. Both fields are described in Section 19.2.2.

The databases are constructed to be used for manual indexing, corporate knowledge representation, back-of-book indexing, etc. For example, the concepts of a thesaurus are used as abstracted descriptors of the content of documents. They can also serve as a basis for NLP. Since terms are condensed linguistic representations of major concepts in a field, they can be used as abstracted descriptors of the content of documents. In order to exploit the facilities offered by large-scale NLP, terms can be converted into lexical databases and used for automatic indexing through the automatic recognition of terms in text documents. In this case, indexing is called *controlled indexing* because it refers to an authority list of terms. It is called *free indexing* in the case where no preliminary terminological data is available.

To sum up, term-based NLP is divided as indicated by Table 19.2. We now turn to the description of some techniques involved in the acquisition of terminological data.

19.2.2 Some Milestones in Term Acquisition

There are basically two techniques for discovering terms in corpora: symbolic approaches that rely on syntactic descriptions of terms—mainly noun phrases, and statistical ap-

proaches that exploit the fact that the words composing a term tend to be found recurringly close to each other more frequently than would occur just by chance (see Chapter *Statistical Methods in NLP*).

A Grammar-based Approach: *TERMINO*

TERMINO is a seminal work in the domain of corpus-based terminology driven and funded by the *Office de la langue française*. It relies on the hypothesis that there are lexical and syntactic clues that can be used to detect terms in texts. The parsing module is divided in subcomponents in agreement with the hypothesis of modularity in X-bar theory (Chomsky 1981). The sequence of operations that perform the extraction of terms from raw corpora is:

1. **Preprocessing**. Text filtering and removal of format characters.
2. **Parsing and Term Extraction.**
   (a) Morphological analysis.
   (b) Noun phrase parsing.
   (c) Term generation.
3. **Interactive Term Bank Construction and Management**. An additional tool provides a user-friendly interface for the construction of term banks from terms extracted by the preceding steps.

TERMINO relies on a description of noun phrases and nominal compounds composed of a kernel structure, the *nucleus*, that can be enriched with pre-specifiers and post-complements (*localizers*). The noun phrases extracted through this local grammar are then filtered by a module which is in charge of comparing and evaluating terms. It relies on the categories of the constituents and on a comparison of ambiguous constructions with other nested non-ambiguous constructions.

Because of its applied orientation, *TERMINO* is also equipped with an interactive environment for the creation and the management of term descriptions. More generally, additional devices for knowledge construction and management appear to be major components of tools dedicated to the automatic acquisition of terms (*LEXTER*, *Termight*, etc.).

A Combination of Patterns and Learned Selection Restrictions: *LEXTER*

Since *LEXTER* was developed in an applied environment the tool is also accompanied by a rich visual interface for validating and organizing the candidate terms acquired from corpora. The major difference between *LEXTER* and *TERMINO* relies on the nature of the technology used for NLP. While *TERMINO* is based on generative grammar rules, *LEXTER* operates on a tagged corpus (see Chapter *POS Tagging*) with the help of lexico-syntactic patterns based on lexical elements and syntactic categories (Bourigault 1993; Bourigault 1996).

Apart from the validation and the structuring of a terminology, the acquisition of candidate terms from *LEXTER* is essentially made of two steps: firstly, maximal noun phrases are extracted through the detection of phrase boundaries and, secondly, these phrases are split into sub-constituents that are likely to be correct *candidate terms*. The output phrases of the termer are intentionally called candidates in order to stress the crucial role of human validation as post-processing of the automatic acquisition.

Maximal noun phrases The grammar of *LEXTER* is composed of transducers called *splitting rules* which indicate the most likely boundaries of noun phrases. Since some rules are over-permissive, they can be complemented with restrictions based on exception lists. For instance, a preposition following a past participle is a boundary. Thus *les clapets* ***situés sur*** *les tubes d'alimentation* (the valves **located on** the feeder pipes) is split into *les clapets* (the valves) and *les tubes d'alimentation* (the feeder pipes). This rule is applied unless the preposition is *de* (of) or unless the verb belongs to an exception list. The exception list is automatically acquired by the tool through an endogenous learning procedure which disambiguates problematic situations from unambiguous occurrences in the corpus.

Decomposition into candidates The maximal noun phrases produced by the preceding step are then split into binary phrases made of a head and an expansion. These nominal chunks together with the maximal noun phrases constitute the set of candidate terms produced by the system. For instance, from a noun-preposition-adjective-noun structure such as *pylône à haute tension* (high voltage pylon), the candidate *haute tension* (high tension) is generated. As in the preceding case, a learning procedure is used to choose between several potential positions for splitting through corpus investigation.

The candidate terms are then automatically organized into a network on the basis of shared lexical elements in similar syntactic positions. The thus structured set of terms is finally introduced into a database in order to be validated and modified by an expert.

A Combination of Syntactic Patterns and Statistical Filters: *ACABIT*

Since both *TERMINO* and *LEXTER* rely on lexico-syntactic analysis of the textual data, any occurrence—even if it appears only once in a corpus (a *hapax legomena*)—is likely to be a candidate term. While this can be considered as an advantage in some situations (no occurrence will be missed), it may be desirable in other situations to rely on ranking techniques which sort candidate by decreasing order of termhood. This purpose is achieved in *ACABIT* through a hybrid technique that combines analytical filtering and statistical ranking (Daille 1996). The acquisition of candidate terms in *ACABIT* is performed in two steps: a filtering step which extracts candidate terms from the corpus and a sorting step which ranks these candidates by decreasing order of quality.

Linguistic filtering As in *LEXTER*, the acquisition of candidate terms in *ACABIT* relies on finite state transducers (see Chapter *Finite-state Machines*). Positive patterns which describe term structures are used instead of the boundary rules of *LEXTER*. Three main binary noun phrase structures are considered for English (noun-adjective, noun-noun, and noun-preposition-noun) along with variation patterns such as coordination (*packet assembly/disassembly*), over-composition (*satellite transit network*), and syntactic modifications (*multiple satellite links*).

Statistical ranking The candidates acquired at the preceding step are then sorted according to several statistical criteria. Core frequency and *log-likehood ratio* (Dunning 1993) are among the two measures that are reported to produce the best classifications. The formula of *log-likehood ratio* for a binary term $W\,W'$ is given by Equation (19.2) in which $f(w, w')$ is the frequency of a contiguous pair of words w and w'.

$$\begin{aligned}\lambda(W,W') &= a\log(a)+b\log(b)+c\log(c)+d\log(d) \\ &\quad -(a+b)\log(a+b)-(a+c)\log(a+c)-(b+d)\log(b+d) \\ &\quad -(c+d)\log(c+d)+(a+b+c+d)\log(a+b+c+d) \\ \text{with} \quad & a=f(W,W'), b=\sum_{w\neq W'} f(W,w) \\ \text{and} \quad & c=\sum_{w\neq W} f(w,W'), d=\sum_{w\neq W, w'\neq W'} f(w,w')\end{aligned} \tag{19.2}$$

The statistical measure for term ranking is complemented with a *measure of diversity* which indicates how frequently the head or argument words of a term combine with other words in the same configuration. Head words with a high diversity denote central concepts in a domain (e.g. *network*, *antenna*, and *satellite* in telecommunication corpora), while arguments with a high diversity are of little terminological interest (e.g. *necessary*, *important*, and *following*).

For the English language, *TERMS* (Justeson & Katz 1995) is a tool developed for the acquisition of terminological data that bears some similarities to *ACABIT*. *TERMS* also combines the filtering of candidate structures through syntactic patterns. The regular expression used for extracting candidate patterns is:

$$((\mathrm{A}\,|\,\mathrm{N})^{+}\,|\,(\mathrm{A}\,|\,\mathrm{N})^{\star}\ (\mathrm{N}\ \mathrm{P})\ (\mathrm{A}\,|\,\mathrm{N})^{\star})\ \mathrm{N} \tag{19.3}$$

While *ACABIT* and *LEXTER* take unambiguously tagged corpora as input, *TERMS* uses morphologically analyzed corpora in which words may have several part-of-speech categories. The authors claim that their solution is more tolerant to tagging errors but however reaches a high level of precision. As in *ACABIT*, terms are ultimately filtered by a statistical measure in *TERMS*: candidate terms which occur only once in the text are rejected.

A Collocation Extractor: *Xtract*

In *Xtract* (Smadja 1993), the hybrid combination of a linguistic filter and a statistical measure is performed the other way round than in *ACABIT* and *TERMS*. First a statistical measure is applied to a corpus in order to acquire binary collocations. Then, the collocations are parsed by *Cass* (Abney 1990), a shallow and robust parser, in order to assign syntactic tags to the collocations. These tags denote the nature of the syntactic relation between the collocates.

The statistical measure used in *Xtract* differs from those used in *ACABIT*. In *Xtract* the strength of the association between two words is based on z-score. It is given by Equation (19.4) in which $f(W,W')$ is the frequency of co-occurrence of W and W' in a 10-word window, $\overline{f}=\sum_w f(W,w)$ is the average frequency, and σ the standard deviation of f.

$$z(W,W') = \frac{f(W,W')-\overline{f}}{\sigma} \tag{19.4}$$

For each word W, the words W' with a strength lower than a given threshold are eliminated. Then, the frequency of co-occurrence are analyzed more precisely in correlation with the relative positions of W and W'. Finally *Xtract* produces the pairs $W\,W'$ such that peaks of frequency are observed for certain values of the oriented distance from W to W'.

Contrary to *ACABIT*, *Xtract* is not focused on terms but on collocations, repeated associations of words. Only a subset of the collocations extracted by *Xtract* can be considered as candidate terms (e.g. *stock trader* or *last selloff*). Other structures are called grammatical associations and may correspond to compositional construction without terminological status such as *epic selloff* (for semantic reasons) or *investors awaited* (for syntactic reasons).

A Non-linguistic Approach: *ANA*

ANA (Enguehard & Pantera 1995) is at the opposite end of *TERMINO* in the spectrum of term acquisition. *ANA* is a language-independent tool for automatic term acquisition. It combines modules for approximate string matching and observations of recurring word co-occurrences. The tool is incremental in the sense that the terms acquired at a given step of the process serve as a basis for the acquisition at the following step. The incremental procedure is called *discovery module* and is in charge of acquiring candidate terms. It is preceded by a *familiarization module* in which three initial sets of data are defined.

Familiarization module The three sets of words defined at this step are:

1. a *stop list* composed of the most frequent words,
2. a set of *seed terms* manually collected which contains some of the major concepts in the domain of a corpus,
3. a set of *scheme words* found inside co-occurrences of seed terms, mainly prepositions and determiners in French.

Discovery module The set of seed terms from the familiarization module serves as a bootstrap for the incremental process for term discovery. It relies on three schemes of acquisition which use the terms acquired up to the preceding step to acquire new candidates at the current step:

1. *expressions* correspond to recurring co-occurrences of single-word terms,
2. *candidates* are single words which are repeatedly associated with a single-word term and separated by a scheme word,
3. *expansions* are binary terms composed of a single-word and a single word term repeatedly associated.

Since the textual data is not morphologically analyzed, morphological variants are conflated with the help of some kind of *string-edit distance* (Hall & Dowling 1980). In addition, the stop-list words are removed from the corpora for the purposes of simplification.

A Summary on Monolingual Term Acquisition

The tools for term acquisition presented in this section use a wide variety of NLP techniques: grammar-based parsing (*TERMINO*), finite state automata or transducers (*LEXTER*, *TERMS* and *ACABIT*), statistical measures (*ACABIT* and *Xtract*), part-of-speech tagging (*LEXTER* and *ACABIT*), text simplification (*ANA*), etc.

The major contrast between these approaches lies on the choice of whether or not to use a statistical filter. The tools which refuse the statistical filtering such as *LEXTER*

assume that any occurrence in a document can be a good candidate even if it appears only once. Such tools generally provide the user with interactive facilities for the selection and the structuring of the candidates. On the contrary, tools with a statistical ranking such as *ACABIT* are intended to be used in an environment in which it is preferable to offer an automatic pre-selection of the candidates.

Despite the apparent heterogeneity of text processing techniques, all these tools rely on the hypothesis that most terms are noun phrases or compound nouns. This hypothesis is criticized by recent studies on term variation which indicate that verb phrases can be conceptually equivalent to nominal terms (Jacquemin, Klavans, & Tzoukermann 1997).

19.2.3 Some Milestones in Monolingual Term Recognition and Automatic Indexing

Automatic indexing consists of associating text chunks with condensed descriptors. The simplest way to index a document in to build an inverted file and to associate this document with each of the words that it contains. This bag-of-word technique has the major drawback of ignoring the linguistic structure of the words. In this section, we intend to focus on the NLP techniques used for extracting indices. Contrary to the bag-of-word approaches, NLP-based indexers preserve word order and dependency relationships between words. Term recognition—also called phrase indexing—is certainly the most prevalent technique in NLP for automatic indexing (see Chapter *Information Retrieval*).

Our presentation will first introduce shallow NLP-based indexing techniques such as text simplification and window-based keyword recognition. Then, more sophisticated techniques such as dependency or transformation-based parsing will be presented.

Text Simplification: *FASIT*

Automatic indexing with *FASIT* (Dillon & Gray 1983) consists of two steps: an index extraction through syntactic patterns and a conflation of indices through text simplification and stemming.

Tagging and pattern matching Since taggers were not available at the time when *FASIT* was developed, the first step is a morphological analysis through suffix-based rules and exception lists for irregular suffixes. The tagset is made of classical syntactic tags (noun, adjective, verbs, etc.) and possible morphological features such as number or tense. Some of the ambiguities are resolved through the use of an exception list which contains frequent words, words with irregular endings, domain-dependent words, etc. Then, a set of contextual rules is used to discard some of the remaining ambiguities.

Finally, the tagged text is matched against a set of syntactic patterns which describe index structures—mainly single or compound nouns with structures such as Noun, Noun-plural, Noun Noun, Proper-noun Proper-noun, etc. A technique similar to *TERMS* is used to accommodate tagging ambiguities: syntactic patterns for index extraction contain alternatives which correspond to frequent ambiguities such as adjective/noun tags. For instance the (Adjective|Noun) Noun pattern is used to extract binary indices whatever the tag assigned to the first word in the case of ambiguity between adjective and noun tag.

Index conflation The phrase indices extracted at the preceding step are conflated through a classical text simplification technique composed of the following three steps:

1. deletion of stop words such as prepositions, conjunctions and general nouns,
2. stemming,
3. word reordering.

In addition, the bags of words produced by this technique are grouped if they share common words.

In *LinkIt* (Wacholder 1998), a system for the detection of significant topics in documents, major concepts are represented by noun phrases or named entities. As in *FASIT*, terms are grouped according to their lexical elements. *LinkIt* does not make use of statistical filtering but instead proposes a first organization of the terminological data.

Disambiguated Noun Phrases: *CLARIT*

The *CLARIT* system combines NLP techniques for morphological analysis (the initial system (Evans *et al.* 1991)) and shallow parsing and statistical filtering for compound noun disambiguation and decomposition (the latest version of the system (Evans & Zhai 1996; Zhai 1997)).

First the text is morphologically analyzed and unambiguously tagged. Then, a context free parser builds candidate noun phrase structures without considering structural ambiguities. For instance, *the redesigned R3000 chips from DEC* is analyzed as [*the*]$_{\text{Det}}$ [*redesigned R3000*]$_{\text{PreMod}}$ [*chips*]$_{\text{Head}}$ [*from DEC*]$_{\text{PostMod}}$ (PreMod and PostMod are pre- and post-modifiers).

The candidates produced by the parser are then sorted according to statistical measures which combine frequency counts, document counts, and domain specific measures established using domain corpora. Even though text structure is taken into account, the matching of query and document descriptors relies on partial overlaps of queries and indices. Different types of match are defined depending on the type of overlap between the descriptor of a query and the descriptor of a document. For instance, a *general match* occurs when a candidate NP is a substring of a controlled term.

Subsequent work on the *CLARIT* system has focused on indexing refinement. As indicated above, the output of the parser is a set of ambiguous noun phrases in which syntactic relations are expressed through dependency relations between head words and modifiers. In (Evans & Zhai 1996), a corpus-based technique for structural disambiguation is proposed. Repeated associations of word pairs are used to select subcompounds from the ambiguous structures produced by the parser. Another technique for choosing among several competing head/modifier relations inside noun phrases is also proposed in (Zhai 1997). It relies on an iterative algorithm that maximizes the probabilities of intra-NP head/modifier associations.

There is an important debate in the literature on information retrieval and automatic indexing about the respective merits of syntactic phrase indexing (based on grammar rules) and statistical phrase indexing (based on frequency counts and statistical measures). In (Fagan 1987) or (Mitra *et al.* 1997), two important studies on this topic, neither of the two techniques is reported to have a significant advantage over the other.

Parsing for Automatic Indexing: *TTP*, *COP*, *COPSY*, etc.

There is an important collection of studies that exploit large-scale parsers in order to extract noun phrases from documents. These studies rely on the output of shallow parsers in which dependencies between heads and arguments are given, whether they are words or constituents. The studies on large-scale parsers differ in the size of the textual data they have been applied to. Some of these tools have processed very large scale data such as the TREC collection for *TTP* (Strzalkowski 1995), other tools, such as the *Constituent Object Parser* (Metzler *et al.* 1990) are prototypes which have been only applied to rather small collections such as the CACM corpus.

There are two main ways of expressing grammatical relations in a language: the constituency approach based on the generative paradigm and the dependency approach resulting from the satisfaction of some linguistic constraints in a given configuration. For instance, the *TTP* outputs phrases while the *Constituent Object Parser* outputs binary dependency trees. We now present the major characteristics of some of the tools in these two families.

Constituency. The *TTP* parser (Strzalkowski 1995) produces parse trees of the sentences which may be incomplete. In the case where the input is ill-formed or ungrammatical, the parser manages to parse the sentences through a skip-and-fit recovery procedure. The parse trees consist of head and related arguments. For instance, *the former Soviet president* is analyzed as

$$[_{\mathrm{NP}}\ [_{\mathrm{N}}\ \textit{president}][_{\mathrm{T\_pos}}\ \textit{the}][_{\mathrm{Adj}}\ [\textit{former}]]\ [_{\mathrm{Adj}}\ [\textit{Soviet}]]] \tag{19.5}$$

The *TTP* parser relies on a large and comprehensive grammar of the English language (drawn from the Linguistic String Grammar (Sager 1981)) and contains subcategorization extracted from the Oxford Advanced Learner's Dictionary. The *TTP* parser is certainly one of the most ambitious and most complete NLP modules for automatic indexing.

The phrases in the parse trees produced by the *TTP* parser are used to generate indices for the documents. An index is either composed of a head noun and one of its adjacent adjective modifiers, or a head noun and the head of its right adjunct, or the verb of a sentence and the head word of its subject or its object phrase.

The *Constituent Object Parser* (*COP*) (Metzler *et al.* 1990) is another linguistically ambitious project for applying NLP to automatic indexing. Unlike the *TTP* parser, the *COP* only relies on binary dependencies. Since the dominance relationship is assumed to be transitive, any n-ary dependency can be transformed into a binary tree of depth $n-1$. Through the *COP*, the utterance *small liberal arts college for scared junior* is analyzed as follows

$$[\ \star[\textit{small}\ \ \star[\textit{liberal}\ \ \star[\textit{arts}\ \ \star\textit{college}]]]\ \ [\textit{for}\ \ \star[\textit{scared}\ \ \star\textit{junior}]]] \tag{19.6}$$

The starred branch in each binary subtree corresponds to the head and the non-starred one to the adjunct. For instance, in *liberal arts college*, *liberal* is the adjunct and *arts college* is the head word.

Modifier attachments are not disambiguated with the *COP*, each adjunct is attached to the rightmost constituent. The pairing procedure between the parse tree of a query and the parse tree of a sentence in a document preserves multiple interpretations. It produces all the possible structures by considering the transitivity of dominance. From

the reported experiments, it is however not certain that the application would be scalable to large document databases.

Dependency. More realistic are the dependency-based approaches to NLP for automatic indexing, such as *COPSY* (Schwarz 1988), which rely on a more partial representation of dependencies—only noun phrases are analyzed—and on a more ambiguous representation—several competing crossing dependencies are generated. In *COPSY*, the analysis of *problems of fresh water storage and transport in containers or tanks* results in the following dependencies:

fresh→*water*

water→*storage*→*problem*	*water*→*transport*→*problem*
container→*storage*	*container*→*transport*
tank→*storage*	*tank*→*transport*

The dependencies are established on the basis of configurational properties combined with word categories. In *COPSY* as in the *COP*, partial matches between the representation of a query and the representation of a document are used for information retrieval in order to offer a more flexible account of query/document similarity.

The formalism used by the *COP* belongs to dependency-based parsers even though it is restricted to binary structures. One of the most famous dependency-based computational description of grammar is the *Constraint Grammar* (Karlsson *et al.* 1995). It is used by two similar approaches to a dependency-based technique for automatic indexing in (Sheridan & Smeaton 1992) and in *NPTool* described in (Voutilainen 1993).

In (Sheridan & Smeaton 1992), the output of the *Constraint Grammar* is transformed into a partially ambiguous tree structure. All the content words are leaves in the tree structures in which dependency relations correspond to marked branches as in the *COP*. Another common characteristic between this work and the *COP* is that the structure is not exploited to build indices but is used as such in the retrieval procedure. Approximate matches between the document and the query structures allow to handle some types of variations in the expressions of the main subjects in the query and in the document.

In *NPtool* (Voutilainen 1993), the focus of the work is not automatic indexing but noun phrase extraction which can be however considered as some kind of automatic indexing procedure. The noun phrases extracted by *NPtool* are those which satisfy two complementary sets of rules: NP-friendly and NP-hostile rules. The NP-friendly rules accept the noun phrases with the highest number of words, while NP-hostile rules retain the analysis with the lowest number of words. The noun phrase structures that are agreed upon by both sets of rules are likely to be correct unambiguous noun phrases.

The dependency-based approach to language actually stems to earlier linguistic work such as (Tesnière 1959). The automatic indexer (Debili 1982) developed in the framework of the *SPIRIT* system (Andreewsky, Debili, & Fluhr 1977) is based on Tesnière's notion of structural dependency. Words are included into a lattice of dependency relations as shown above in the presentation of *COPSY*. Even though Debili's approach was developed relatively early, approximately ten years earlier than the other approaches described in this section, it already contained most of the techniques that were used in later study: morphological analysis, unambiguous tagging, extraction of dependencies through finite state techniques, and resolution of ambiguity through endogenous learning. Another important feature of Debili's indexing technique is the extrapolation of words in term occurrences to their morphological family. Through this technique and through the morphological relationship between *afficher* (to post) and *affichage* ([a] posting), the utterance *affichage sur les murs* (posting on the walls) is related with *afficher sur les*

murs (to post on the walls).

Recognition of Variation: *FASTR*

The recognition of variation appears to be a constant issue in NLP for automatic indexing. Either variation is explicitly addressed through the exploitation of word paradigms as in (Debili 1982) or variation is implicitly accounted for in approximate matching techniques between query and document representations as in the *COP* or in *COPSY*. In *FASTR* (Jacquemin 1999), the focus is to design and to implement a variationist description of terms in order to recognize term variants whatever the linguistic means involved in the paraphrase.

The description of variations in *FASTR* relies on metarules which combine structural transformations (the syntagmatic axis) and lexical relationships (the paradigmatic axis). Two main families of lexical relationships are used: morphological links as in (Debili 1982) and semantic links such as synonymy or antonymy. The following metarule denotes an adjective to noun transformation accompanied by a semantic variation:

$$\mathrm{Adj}_1\,\mathrm{Noun}_2 \;\rightarrow\; \mathrm{Noun}_1\ ((\mathrm{CC}\ \mathrm{Det}^?)^?\ \mathrm{Prep}\ \mathrm{Det}^?\ (\mathrm{Adj}|\mathrm{N}|\mathrm{Part})^{0-3})\ \mathrm{Noun}_2' \quad (19.7)$$

In this metarule, Adj_1 and Noun_1 are morphologically related, and Noun_2 and Noun_2' are semantically related. It is used, with the help of morphological and semantic relationships, to recognize that *malignancy in orbital tumours* is a variant of *malignant tumor*. *Malignancy* and *malignant* are morphologically related, *tumour* and *tumor* are semantically related, and $malignancy_{\mathrm{Noun}}$ in_{Prep} $orbital_{\mathrm{Adj}}$ $tumours_{\mathrm{Noun}}$ matches the target pattern.

There are three basic types of variations:

- *syntactic variations* which involve only structural transformations,
- *morpho-syntactic variations* which involve structural transformations and morphological relations, and
- *semantic variations* which involve only semantic relationships.

Hybrid variations are built by combining these elementary types into *semantico-syntactic* or *morpho-semantico-syntactic variations*.

While the parser-based approaches to automatic indexing, such as *TTP*, *COP*, or *COPSY* are geared towards free indexing, *FASTR* is designed for controlled indexing. It takes as input an authority list of terms transformed into computational data and generates candidate variants of these terms. The candidate variants are then paired with corpus sentences in order to retrieve actual variant occurrences.

The recognition of variants by *FASTR* or Debili's parser is differential: it relies only on lexical links and structural transformations. An alternative approach to variant recognition is proposed in (Sparck Jones & Tait 1984). It makes use of an explicit semantic representation of the base terms and their variants. Single words are labelled with semantic tags and semantic relations are added to the branches of the syntactic structures. Then variants are generated by transforming the semantico-syntactic representation of terms through paraphrase patterns. Similarly, Woods (1997) infers the semantic relationships between terms and variants through subsumption-based reasoning. This combines a semantic description of terms and a semantic lattice made of the single words contained in the terms and their variants.

19.2.4 Crosslingual Term Recognition

The programs for term alignment consist of establishing relationships between terms from two parallel corpora in two different languages. They generally operate in two steps: an acquisition step in which terms are acquired in each corpus and an alignment phase in which links are made between the terms in both languages. The acquisition phase is generally quite simple when compared with in-depth monolingual studies. Clearly, programs for term alignment focus on the alignment techniques.

In (van der Eijk 1993) terms are extracted through part-of-speech tagging and selection of $(\text{Adj})^{\star}\,(\text{Noun})^{+}$ patterns. Then, these candidate terms are aligned by comparing local and global frequencies of co-occurrence. The local co-occurrences are those which take place in fragments of texts that are aligned by the Gale-Church sentence alignment technique (Gale & Church 1991). The global co-occurrences are those which take place at the corpus level. In order to discard rare terms, the ratio of local to global frequencies must be higher than a given threshold.

The technique implemented in *Termight* (Dagan & Church 1994) for the alignment of candidate terms is based on the alignment of their words through the *Word-align* program (Dagan, Church, & Gale 1993). For each occurrence of a source term, *Termight* identifies candidate translations as sequences of words aligned with any of the words in a source term. The candidate term translations are collected in several places in the corpus and sorted in order of decreasing frequency. The user can interact with the proposed list through a graphical user interface in order to select or discard the candidate translations.

The technique used by Gaussier (1998) relies on corpora aligned at the sentence level. Association probabilities between single words are calculated on the basis of bilingual co-occurrences of words in aligned sentences. Then these probabilities are used to find the French correspondent of English terms through a flow network model. In this model, a graph between source and target words is weighted with association probabilities. The selected translations of the multi-word terms are those which correspond to the minimal cost flow in the graph.

Hull (1997) differs from Gaussier (1998) in that single word alignment, term extraction and term alignment are three independent modules. Terms and words are aligned through a greedy algorithm that scores the candidate bilingual pairs according to probabilistic data, chooses the highest scored pair, removes it from the pool, and repeatedly recomputes the scores and removes pairs until all the pairs are chosen. Since the system is not intended to be fully automatic, its design is made in such a way that it minimizes human labor.

19.3 Recent Advances and Prospects

The emerging techniques in NLP, such as shallow, robust, and/or partial parsing, hybrid models combining statistical and symbolic processing, high level taggers such as semantic taggers, and large scale morphological analyzers, are unique opportunities for better processing of specialized corpora (see Chapter *Sublanguages and Controlled Languages*) and providing better access to the information they contain. In addition, the increasing number of on-line documents makes it necessary to take into consideration the structure of documents in addition to traditional textual features for computational term processing.

Among the promising lines of research, the following issues should be considered:

- linguistic studies with a concern for specialized corpora and corpus investigation, possibly in cooperation with large scale parsers,
- construction of large-scale semantic and morphological resources for term and variant recognition,
- new hybrid solutions for term acquisition and recognition combining symbolic processing and machine learning techniques,
- semantic tagging and acquisition of semantic relationships from corpora,
- sophisticated techniques for corpus construction with fine grained parameters in order to build domain-oriented corpora,
- combination of textual and structural information for the recognition of rich contexts such as expositive or paraphrastic contexts,
- enhanced term acquisition procedures: extraction of terms at the verb or adjective phrase level, combination of acquisition and recognition for term structuring, enhanced interfaces for expert validation...

References

Abney, Steven P. 1990. "Rapid incremental parsing with repair". *Proceedings of the 6th New OED Conference: Electronic Text Research*, pages 1–9, University of Waterloo, Waterloo, Ontario.

Andreewsky, A., Fathi Debili, & Christian Fluhr. 1977. "Computational learning of semantic lexical relations for the generation and automatic analysis of content". *Proceedings, IFIP Congress*, pages 667–673, Toronto.

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

Bourigault, Didier. 1996. "LEXTER, a Natural Language tool for terminology extraction". *Proceedings, 7th EURALEX International Congress*, pages 771–779, Göteborg.

Chomsky, Noam. 1981. *Lectures on Government and Binding.* Dordrecht: Foris Publications.

Condamines, Anne & Josette Rebeyrolles. 1998. "CTKB: A corpus-based approach to a Terminological Knowledge Base". *Proceedings, 1st Workshop on Computational Terminology (COMPUTERM'98)*, pages 29–35, Montreal.

Dagan, Ido & Kenneth W. Church. 1994. "*Termight*: Identifying and translating technical terminology". *Proceedings, 4th Conference on Applied Natural Language Processing (ANLP'94)*, pages 34–40, Stuttgart.

Dagan, Ido, Kenneth W. Church, & William Gale. 1993. "Robust bilingual word alignment for machine aided translation". *Proceedings, Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1–8, Ohio State University.

Daille, Béatrice. 1996. "Study and implementation of combined techniques for automatic extraction of terminology". *The Balancing Act: Combining Symbolic and Statistical Approaches to Language* ed. by Judith L. Klavans & Philip Resnik. Cambridge, MA: MIT Press, pages 49–66.

David, Sophie & Pierre Plante. 1990. "De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes". *Intelligence Artificielle et Sciences Cognitives au Québec*, 3(3):140–154.

Davidson, L., J. Kavanagh, K. Mackintosh, I. Meyer, & D. Skuce. 1998. "Semi-automatic extraction of knowledge-rich contexts from corpora". *Proceedings, 1st Workshop on Computational Terminology (COMPUTERM'98)*, pages 50–56, Montreal.

Debili, Fathi. 1982. *Analyse Syntaxico-Sémantique Fondée sur une Acquisition Automatique de Relations Lexicales-Sémantiques.* Thèse de Doctorat d'État en Sciences Informatiques, University of Paris XI, Orsay.

Dillon, Martin & Ann S. Gray. 1983. "FASIT: A fully automatic syntactically based indexing system". *Journal of the American Society for Information Science*, 34(2):99–108.

Dunning, Ted. 1993. "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics*, 19(1):61–74.

van der Eijk, Pim. 1993. "Automating the acquisition of bilingual terminology". *Proceedings, 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, pages 113–119, Utrecht.

Enguehard, Chantal & Laurent Pantera. 1995. "Automatic natural acquisition of a terminology". *Journal of Quantitative Linguistics*, 2(1):27–32.

Evans, David A., Kimberly Ginther-Webster, Mary Hart, Robert G. Lefferts, & Ira A. Monarch. 1991. "Automatic indexing using selective NLP and first-order thesauri". *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'91)*, pages 624–643, Barcelona.

Evans, David A. & Chengxiang Zhai. 1996. "Noun-phrase analysis in unrestricted text for information retrieval". *Proceedings, 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, pages 17–24, Santa Cruz, CA.

Fagan, Joel L. 1987. "Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods". *Proceedings, 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'87)*, pages 91–101.

Felber, Helmut 1984. *Terminology Manual.* Paris: Unesco, International Information Centre for Terminology (Infoterm).

Gale, William & Kenneth W. Church. 1991. "A program for aligning sentences in bilingual corpora". *Proceedings, 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*, pages 177–184, Berkeley, CA.

Gaussier, Éric. 1998. "Flow network models for word alignment and terminology extraction from bilingual corpora". *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 444–450, Montreal.

Hall, Patrick A. & Geoff R. Dowling. 1980. "Approximate string matching". *Computing Surveys*, 12(4):381–402.

Hull, David. 1997. "Automating the construction of bilingual terminology lexicons". *Terminology*, 4(2):225–244.

Jacquemin, Christian. 1999. "Syntagmatic and paradigmatic representations of term variation". *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 341–348, University of Maryland.

Jacquemin, Christian, Judith L. Klavans, & Evelyne Tzoukermann. 1997. "Expansion of multi-word terms for indexing and retrieval using morphology and syntax". *Proceedings, 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL'97)*, pages 24–31, Madrid.

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

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, & Arto Anttila (Editors). 1995. *Constraint Grammar A Language-Independent System for Parsing Unrestricted Text.* Berlin: Mouton de Gruyter.

Lauriston, Andy. 1994. "Automatic recognition of complex terms: Problems and the TERMINO solution". *Terminology*, 1(1):147–170.

Metzler, Douglas P., Stephanie W. Haas, Cynthia L. Cosic, & Charlotte A. Weise. 1990. "Conjunction ellipsis, and other discontinuous constituents in the Constituent Object Parser". *Information Processing and Management*, 26(1):53–71.

Mitra, Mandar, Chris Buckley, Amit Singhal, & Claire Cardie. 1997. "An analysis of statistical and syntactic phrases". *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'97)*, pages 200–214, Montreal.

Pearson, Jennifer. 1998. *Terms in Context.* Studies in Corpus Linguistics. Amsterdam: John Benjamins.

Sager, Juan C. 1990. *A Practical Course in Terminology Processing.* Amsterdam: John Benjamins.

Sager, Naomi. 1981. *Natural Language Information Processing : A Computer Grammar of English and its Applications.* Reading, MA: Addison-Wesley.

Schwarz, Christoph. 1988. "The TINA Project: text content analysis at the Corporate Research Laboratories at Siemens". *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'88)*, pages 361–368, Cambridge, MA.

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

Smadja, Frank. 1993. "Retrieving collocations from text: Xtract". *Computational Linguistics*, 19(1):143–177.

Sparck Jones, Karen & John I. Tait. 1984. "Linguistically motivated descriptive term selection". *Proceedings, 10th International Conference on Computational Linguistics (COLING'84)*, pages 287–290, Stanford, CA.

Strzalkowski, Tomek. 1995. "Natural language information retrieval". *Information Processing & Management*, 31(3):397–417.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale.* Paris: Klincksieck. Fifth edition, 1988.

Voutilainen, Atro. 1993. "*NPtool*, A detector of English noun phrases". *Proceedings, Workshop on Very Large Corpora : Academic and Industrial Perspectives*, pages 48–57, Columbus, Ohio.

Wacholder, Nina. 1998. "Simplex NPs clustered by head: A method for identifying significant topics within a document". *Proceedings, COLING/ACL Workshoop on the Computational Treatment of Nominals*, pages 70–79, Montreal.

Woods, William A. 1997. "Conceptual indexing : A better way to organize knowledge". Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA.

Zhai, Chengxiang. 1997. "Fast statistical parsing of noun phrases for document indexing". *Proceedings, 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 312–319, Washington.