

Methods of Automatic Term Recognition

— A Review —

Kyo KAGEURA

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello St. Sheffield S1 4DP, UK

E-Mail: k.kageura@dcs.shef.ac.uk

(on study leave from National Center for
Science Information Systems, Japan)

Bin UMINO

Faculty of Sociology, Toyo University,
5-28-20 Hakusan, Bunkyo-ku, Tokyo 112, Japan

E-Mail: umino@hakusrv.toyo.ac.jp

May 27, 1996

Abstract

Following the growing interest in “corpus-based” approaches to computational linguistics, a number of studies have recently appeared on the topic of automatic term recognition or extraction. Because a successful term recognition method has to be based on proper insights into the nature of terms, studies of automatic term recognition not only contribute to the applications of computational linguistics but also to the theoretical foundation of terminology. Many studies on automatic term recognition treat interesting aspects of terms, but most of them are not well founded and described.

This paper tries to give an overview of the principles and methods of automatic term recognition. For that purpose, two major trends are examined, i.e. studies in automatic recognition of significant elements for indexing mainly carried out in information retrieval circles, and current research in automatic term recognition in the field of computational linguistics.

Keywords

Automatic term recognition, automatic indexing, automatic text processing, corpus-based approach, simple term, complex term, quantitative analysis, linguistic analysis

1 Introduction¹

In computational linguistics we have recently witnessed a growth in the interest in automatic treatment of terms, or linguistic units which characterise specialised domains, especially when NLP systems “are passing from the development stage to the application stage” (Ananiadou 1994:1034). Automatic term recognition (ATR) in particular is much needed because a simple but coherently built terminology is the starting point of many applications such as human or machine translation, indexing, thesaurus construction, knowledge organisation, etc. and because manual efforts cannot keep up with the rapid growth of technical terms.

Methodologically, current ATR research is situated within the broad category of “corpus-based” approaches to computational linguistics which has become very popular in the last five or six years. Many ATR studies reveal interesting insights into various aspects of terms, but in most cases they are just implied and not explicitly stated. Thus appropriate comparison or comparative evaluation of related studies is not easy.

Turning our eyes to the different but related research field of information retrieval (IR), we recognise a comparable research topic called automatic indexing or automatic keyword extraction. Research in automatic indexing has a long tradition and can be traced back to the late 1950's, when H. P. Luhn published a paper on extracting meaningful elements from texts for use in information retrieval (Luhn 1957). To date substantial efforts have been devoted to automatic indexing, and the ideas introduced by this work can substantially contribute to the ATR research. However, re-interpretation of the work within the current ATR research context is as yet outstanding².

The purpose of this paper is to give a clearer picture of the state-of-the-art of ATR methods as well as their principles and problems. For that purpose we examine the major work in the above two different research fields from the ATR point of view.

First we examine the basic quantitative approaches to automatic indexing relevant to ATR, clarifying the basic background and context of automatic indexing, so as to make it possible to view the relevant indexing studies from the ATR point of view. Then the more elaborate approaches in IR and computational linguistics, both linguistic and quantitative, are reviewed. Finally, various ideas and methodologies are summarised and examined.

2 Automatic Indexing: Basic Quantitative Approach

2.1 Context of IR and Indexing

Information retrieval, or more specifically document retrieval, is the process by which relevant documents (or information about them) are supplied to the person who needs them, by matching

the person's request with documents. Usually the matching between request and documents is not carried out directly but by means of their surrogates. Indexing refers to the process of choosing the surrogates which represent the topic or content of documents (Boyce, Meadow & Kraft 1994:102–103)³. Surrogates of documents are usually some sort of lexical units. Thus in indexing, there are two different levels or spaces, i.e. a document space consisting of a set of documents, and a vocabulary space consisting of lexical units.

The process of automatic indexing reviewed in this section is intended to:

- (1) Define the basic unit for indexing. The work reviewed in this section takes a simple and practical standpoint in this respect, regarding the sequence of letters separated by spaces, commas, periods, etc. as the basic unit. For the sake of simplicity, we call the unit thus defined a 'word', the word selected an 'index term', and this type of indexing 'single-term indexing'.
- (2) Calculate the values or weights of each word based on various frequency information related to the word.

This process is also called index term weighting. With a threshold which divides the weighted words into index terms and non-index terms, it becomes automatic indexing in its literal meaning.

It should be noted here that only the work directly relevant to ATR is examined. This limits the range of the studies reviewed here, because a number of highly elaborate studies on indexing are inherently related to the overall retrieval process and are not readily exportable to ATR⁴.

In IR, index terms are evaluated by their retrieval performance, namely recall and precision. Recall is the proportion of relevant documents retrieved from among the total number of relevant documents in the database, and precision is the proportion of the relevant documents retrieved from among the total number of retrieved documents (Salton 1989:277–278). They are closely correlated with how well an index term represents documents of which it is a surrogate and how well it can discriminate the documents from others. Hence the degrees of representation and discrimination are two of the most important characteristics of index terms. (van Rijsbergen 1979:29–30).

2.2 Quantitative Measures for Single-Term Indexing

As already mentioned, the quantitative approach to single-term indexing can be traced back to late 1950's. A great number of studies have appeared since then, of which we only review major or representative ones. Some of them use rudimentary linguistic information such as stop-word lists consisting of function words, but we focus here on the quantitative measure.

2.2.1 Notational Preparations

For the preparation of our description, let us introduce some basic notations. First, to denote the document space which consists of all the documents in the database we use D , and use d_j to denote each document in D . Thus the document space is:

$$D = \{d_1, d_2, \dots, d_j, \dots\}$$

The vocabulary, or the set of all the different words or word types w_i appearing in D , is denoted by W_D . Thus:

$$W_D = \{w_1, w_2, \dots, w_i, \dots\}$$

If necessary, we use W_{d_j} to denote the vocabulary of a document d_j . When talking of a word w_i with respect to a specific document d_j , we use two subscripts to denote the word, i.e. w_{ij} .

We also define three basic functions as follows:

$n(S)$, which gives the number of elements of a set S

$f(w_{ij})$, which gives the number of occurrences of word w_i in d_j

$g(w_{ij})$, which gives 1 when $w_i \in W_{d_j}$ and 0 when $w_i \notin W_{d_j}$

Lastly, the weight of a word w_i in a document d_j is denoted by I_{ij} . When the document is not specified, we use I with one subscript, i.e. I_i^5 .

2.2.2 Weighting by Occurrence in Document

The simplest weighting measures are based on the occurrence of words in a document, without referring to the occurrence information in other documents. Luhn (1957) was the first exponent of this approach. Various measures have been proposed since then, some based on the existence or non-existence of a word in a document, while some are based on frequency.

Two of the most straightforward measures for the weights of w_i for d_j are $g(w_{ij})$ and $f(w_{ij})$. The former gives 1 to all the words appearing in the document, while the latter gives the frequency of the word in the document as its weight. The following is a little more elaborate, but still fairly straightforward (Sparck-Jones 1973; Noreault, McGill & Koll 1977):

$$I_{ij} = \frac{g(w_{ij})}{\sum_i g(w_{ij})} = \frac{1}{\sum_i g(w_{ij})} \tag{1}$$

$$I_{ij} = \frac{f(w_{ij})}{\sum_i f(w_{ij})} \tag{2}$$

They take into account the size of the document, i.e. the number of different words and the number of running words in the document respectively, in order to normalise the weight. Variations exist, e.g. taking log of the denominator of (2).

2.2.3 Weighting by Occurrence in a Database

Some proposed measures are based on the occurrence of a word in a document relative to its occurrence in the database. For instance, both Sparck-Jones (1973) and Noreault, McGill & Koll (1977) introduce the idea of using the number of documents in which the word occurs or the frequency of the word in the database, instead of or in addition to the denominators of (1) and (2):

$$I_{ij} = \frac{g(w_{ij})}{\sum_i g(w_{ij}) \cdot \sum_j g(w_{ij})} \quad (3)$$

$$I_{ij} = \frac{f(w_{ij})}{\sum_i f(w_{ij}) \cdot \sum_j f(w_{ij})} \quad (4)$$

Variations exist again, e.g. taking a square of the numerator or log of the denominator of (4).

Edmundson & Wyllys (1961), Damerou (1965), and Carroll & Roeloffs (1969) present several formulae based on similar idea, of which we presented only one:

$$I_{ij} = \frac{f(w_{ij})}{\sum_i f(w_{ij})} - \frac{\sum_j f(w_{ij})}{\sum_i \sum_j f(w_{ij})} \quad (5)$$

Noreault, McGill & Koll (1977) also introduce a measure based on the same type of information:

$$I_{ij} = f(w_{ij}) \cdot \log \frac{\sum_i \sum_j f(w_{ij})}{\sum_j f(w_{ij})} \quad (6)$$

A famous weighting measure called inverse document frequency is similar to this, though based on the existence or non-existence of a word in the documents (Sparck-Jones 1972). Salton & Yang (1973) presents the measure, in which the inverse document frequency is multiplied by the frequency of the word in the document:

$$I_{ij} = f(w_{ij}) \cdot \log_2 \frac{n(D)}{\sum_j g(w_{ij})} \quad (7)$$

The frequency of the word in the document and the number of documents in which the word occurs are used simultaneously in this measure.

2.2.4 Weighting by Cross-Document Distribution

Some researchers use variance of word distribution among documents in the database. For instance, Dennis (1967) proposed the following measure to distinguish ‘content word’ and ‘non-content word’ in documents:

$$I_i = \frac{\sum_j f(w_{ij})}{{}_r f(w_i)^2 / {}_r \sigma_i^2} \quad (8)$$

where ${}_r \bar{f}(w_i)$ is the mean, and ${}_r \sigma_i^2$ is the unbiased variance, of relative frequency of w_{ij} for all d_j . Stone & Rubinoff (1968) proposes the measure which uses variance of word frequency distribution as well.

Nagao, Mizutani & Ikeda (1976) uses the statistical measure of χ^2 to calculate the weight of words:

$$I_i = \chi_i^2 = \sum_j \frac{f(w_{ij}) - m_{ij}}{m_{ij}} \quad (9)$$

where

$$m_{ij} = \frac{\sum_j f(w_{ij})}{\sum_i \sum_j f(w_{ij})} \cdot \sum_i f(w_{ij})$$

Salton, Yang & Yu (1975) introduces the concept of term discrimination value based on a different viewpoint but nevertheless has intuitive similarity to those based on cross-document distribution (Salton, Yang & Yu 1975; Salton 1989). The idea is that a word which reduces the document density is a good index term. A simple way of calculating document density Q is to use the average pairwise similarity between all pairs of distinct documents:

$$Q = \frac{1}{n(D)(n(D) - 1)} \sum_j \sum_k sim(d_j, d_k)$$

in which $sim(d_j, d_k)$ can be calculated, in the simplest situation for instance, by

$$sim(d_j, d_k) = \sum_i d(w_{ij})d(w_{ik})$$

Finally, the weight of each word can be calculated by

$$I_i = Q - Q_i \quad (10)$$

where Q_i and Q are the document densities with and without w_i taken into consideration.

These measures calculate the weight of each word with respect to the database, and not to each document. Some propose the multiplication of these measures by document dependent measures such as $f(w_{ij})$, in order to give the weight for each document.

2.2.5 Weighting by Distribution of Relevant and Non-Relevant Documents

Harter (1975a; 1975b) proposes a method based on the distinction of the documents which treat the topic represented by the word (class I documents) and those which do not (class II). Assuming that the frequency of a word follows Poisson distributions in both classes, he proposes the model of word distribution by combining two Poisson distributions:

$$Pr(f(w_{ij}) = x) = \pi \frac{e^{-m_{1i}} \cdot m_{1i}^x}{x!} + (1 - \pi) \frac{e^{-m_{2i}} \cdot m_{2i}^x}{x!}$$

where $Pr(f(w_{ij}) = x)$ is the ratio of the number of documents which have x occurrence of w_i to the total number of documents, m_{1i} and m_{2i} are means of the occurrences of w_i in documents in class I and class II respectively, and π is the ratio of class I documents to the total number of documents.

The measure for weighting a word is proposed on the basis of this model, as follows:

$$I_i = \frac{m_{1i} - m_{2i}}{\sqrt{m_{1i} + m_{2i}}} \quad (11)$$

It is necessary to calculate the values of m_{1i} and m_{2i} in order to obtain the weight. Harter explains how to derive these values from the moment generating function of the two Poisson models.

Others also propose weighting measures based on the distinction of relevant and non-relevant documents (Bookstein & Swanson 1974; Bookstein & Swanson 1975; Cooper & Maron 1978). The basic measure is as follows (summarised in Salton (1989:289)):

$$I_i = \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \quad (12)$$

where p_i is the probability of the occurrence of w_i in relevant documents, and q_i is the probability of w_i in non-relevant documents. Because these probabilities cannot be known straightforwardly in an operational IR environment, much efforts have been paid to estimate them.

2.3 Summary of Single-Term Indexing

The basic measures for automatic indexing listed above can be applied to automatic term recognition. The simplest way is to regard index terms and technical terms as being the same. Or, if we use a set of documents belonging to the same domain instead of individual document, which is similar to the idea adopted in (11) and (12), the result can be interpreted as the weighting measure of terms specific to the domain⁶. In this sense, we can recognise a naive but strong parallelism between index terms and terms, and consequently the potential applicability of automatic indexing methods to ATR.

2.3.1 Characterisations of Index Terms

The question “what is an index term?” can and must be approached from at least two points of view, i.e. the status of index terms as syntagmatic linguistic units (unithood), and the status of index terms as units of representative of document contents (termhood)⁷. The methods reviewed above only address the latter aspect.

Despite the variations in weighting measures, the basic ideas on which they are based can be classified into a few basic groups:

- (1) A ‘word’ which appears in a document is likely to be an index term for that document.
- (2) A ‘word’ which appears frequently in a document is likely to be an index term for that document.
- (3) A ‘word’ which appears only in limited number of documents is likely to be an index term for these documents.

- (4) A ‘word’ which appears relatively more frequently in a document than in the whole database is likely to be an index term for that document.
- (5) A ‘word’ which shows a specific distributional characteristic in the database is likely to be an index term for the database.

All the measures introduced above are based on one, or a combination, of these five basic ideas.

(1) and (2) emphasise the ‘representation’ aspect of index terms, while (3) and (4) emphasise the ‘discrimination’ aspect. While (1) to (4) focus on an individual document, (5) takes into account the relationships between documents as seen from the overall distribution of words. Therefore, (5) has the advantage of considering topics as represented by a group of documents, while (1) to (4) only treat each document as a basic topical unit. Accordingly, the measures based on (5) vary considerably, both in theoretical viewpoints and methods, and also in the resultant weights given to words.

As for the content of documents, represented by index terms, all the methods above explicitly or implicitly approximate the content of a document to a set of words appearing in the document. Theoretically, therefore, single-term indexing reviewed here is based on rather fragile grounds on both ends, i.e. of vocabulary space and of document space, because neither the word–concept relation nor the word-sets–content relation has a theoretical or intuitive foundation⁸.

2.3.2 Status of Measures and Framework of Evaluation

Most measures proposed so far take an empirical or pragmatic standpoint regarding the meaning of weight. This can be seen from the existence of variations based on the same idea, e.g. taking or not taking log or square, or applying division or subtraction to the same factors. In that sense, they are essentially the application of statistical or quantitative methods but not of statistical models, although in the broadest interpretation they can be claimed to be based on Bernoulli experiments and thus presuppose Poisson or normal approximation to a binomial distribution.

Some use well established measures such as information theory (in case of (6), neglecting the base of log, and (7)), χ^2 (in case of (9)) or decision theoretic measure (in case of (12)), whose theoretical background is well known in general. However, as has been pointed out (Dunning 1993; Cohen 1995), most of them are rather *ad hoc* with respect to the mathematical validity of the measure. Some researchers, e.g. Damerau (1965) and Harter (1975a; 1975b), are concerned with this point, and give relatively detailed examination of the statistical models from which the weighting measure is derived. However, even in these cases, the status of the model with respect to the selection of index terms is not completely clear.

This problem can also be observed at the stage of evaluating the resultant index terms.

They are evaluated either by subjective comparison with man-made lists of index terms or by retrieval performance, i.e. recall and precision, in experimental IR contexts. In either case, what is examined and evaluated is the result obtained by the application of the measure, and not the theoretical validity of the measure or underlying model itself.

These criticisms are, of course, not fair to the work reviewed here, because, firstly, as long as they are situated in the context of IR applications, the ultimate purpose of automatic indexing is the practical improvement of IR performance, and secondly, they have to pay attention to the applicability of their methods within practical IR environments, which necessarily imposes a certain restriction on their approach. Nevertheless, for further improvement of indexing, not to speak of ATR, totally effect-oriented evaluation has its own limitations. At that point, the problem of the theoretical validity of the statistical model, with respect to language in general and index term in particular, has to be addressed.

3 Approaches to Automatic Term Recognition

In 2.3.1 we described that in single-term indexing the content of a document approximates to the set of words, and an index term approximates to a simple word, where a ‘word’ is a character string delimited by space, comma, etc. as defined in 2.1. The limitations of this viewpoint have long been recognised in IR circle, and various paths have been pursued to overcome them, i.e. by applying linguistic or statistic analysis to recognise linguistically relevant units, grouping index terms by word occurrence patterns, or incorporating dependence information in index term weighting (Salton 1989:290). In the first path, IR work shares the assumption with terminology that a certain type of lexical unit can express concepts or topics without ambiguity. Accordingly, it is expected that recognising the proper lexical unit contributes to indexing, even if the assumption with respect to the content of a document is not much elaborated⁹.

With the growth of practical systems such as machine translation or natural language interface in the field of computational linguistics, the need for managing lexical units specific to an application domain or a certain sublanguage is increasing. One of the major problems is the treatment of terminology which represents domain specific knowledge or concepts. Thus it has become a concern for some parts of computational linguistics to extract domain-specific lexical units. It is usually within this context that the work is pursued under the name of automatic term recognition.

As IR work and CL work share a great deal both in their purpose, i.e. extracting a properly content-bearing lexical unit for a certain topic or a domain, and in their methodology, i.e. using linguistic and/or statistical methods, we will review them together. Although many approaches use both linguistic and statistical information, the emphasis is usually on one or the other of

them. So in the following we categorise the work into linguistically-oriented and statistically-oriented.

Before the examination of the existing studies, it is convenient here to clarify some basic terminology:

- Following the convention of section 2, we use the term ‘word’ to refer to a character string delimited by space, comma, period, etc. However, we also sometimes use ‘word’ to refer to the conventional linguistic concept of ‘word’, e.g. ‘content word’ or ‘word formation’. This is partly to simplify expressions, and partly the reflection of the convention of ATR or automatic indexing studies, where the linguistic status of words (not to speak of lexemes, etc.) has rarely been clarified and is usually approximated to character strings.
- We use ‘term’ in its terminological sense (e.g. Sager (1990) or Kageura (1995)). We use ‘simple term’ and ‘complex term’ to refer to a term consisting of one word and a term consisting of two or more than two words, respectively.

3.1 Linguistically-Oriented Approach

3.1.1 Overview of IR-oriented Work

Some of the early linguistic attempts to extract indexing units consisting two or more words are reported in Earl (1970;1972) and Klingbiel (1973a;1973b). Earl’s (1970;1972) PHRASE system uses a parser to obtain grammatical information about words, and to select noun phrases as index term candidates. Final index terms are selected from candidates according to the frequency of their constituent noun elements. Klingbiel (1973a;1973b) also uses grammatical information. The ‘recognition dictionary’ attaches parts-of-speech to each word in the text, from which certain grammatical sequences are selected as index term candidates. Parts-of-speech suitable for indexing purposes were used, which are different from conventional parts-of-speech in linguistics.

Similar work has continued to appear to date, with some refinements in various aspects, e.g. portability, computational effectiveness, or range of coverage of index term forms. Dillon & Gray (1983) reports a simple parts-of speech based automatic index system called FASIT, which extracts complex index terms. In this system formal variations are normalised into canonical form, e.g. ‘book review’ and ‘review of books’. Salton (1988) proposes a system for book indexing in which sentences are fully parsed and the combination of the heads of constituents are taken as possible complex index terms. They are then weighted by equation (7) (using chapters instead of documents). Evans (1995) introduces the CLARIT system, which uses a parser to extract both simple words and complex noun phrases which are then also weighted by equation (7).

3.1.2 Overview of CL-oriented Work

Bourigault (1992) argues that terminological units represent fixed concepts without ambiguity and take certain syntactic forms, namely noun phrases. The system described in Bourigault (1992), LEXTER, extracts French complex terms, first by extracting maximum-length noun phrases based on the ‘surface grammatical analysis’ (not full parse but using some heuristics) and then, parses the maximum-length noun phrases to extract suitable units as term candidates.

Justeson & Katz (1995) points out two characteristics of complex terms, i.e. (1) terms are not subject to formal variation in usage and thus appear in the same form throughout a text, and (2) grammatical constructs of complex terms are restricted to a certain form of noun phrases. After tagging the text, they extract units as candidate terms which have the form $((A|N)^+ | ((A|N)^*(NP)^2)(A|N)^*N$ and appear more frequently than specified by a threshold. Dagan & Church (1994) adopts basically the same method.

Daille, Gaussier & Langé (1994) recognises two levels in complex terms (or multi-word units (MWU)), i.e. base MWU consisting of two content words, and MWU of length greater than two which are formed on the basis of base MWU by overcomposition (e.g. ‘[side lobe] regrowth’), modification (e.g. ‘interfering [earth station]’), or coordination (e.g. ‘packet assembly/disassembly’). They focus on extracting base MWUs, whose forms were defined, for instance, as ‘N Adj’, ‘N1 N2’, ‘N1 de (det) N2’, ‘N1 prep (det) N2’ in case of French and as ‘Adj N’ and ‘N1 N2’ in case of English. Like Dillon and Gray (1983), they allow for formal variations, e.g. ‘N2 N1’ and ‘N1 of N2’. Valid morpho-syntactic forms are extracted from a tagged corpus, to which statistical scorings are applied to weight the candidate base MWUs. This will be discussed below.

Lauriston (1994) points out various theoretical problems in automatic term recognition, e.g. lexicalisation and terminologisation, syntactic ambiguity, formal variations, etc., and then describes the ATR system called TERMINO, which is based on a detailed morpho-syntactic analysis and uses not only patterns of parts-of-speech but also the syntactic structure of possible terminological units, although he does not give detailed explanation of the grammar used in TERMINO.

Ananiadou (1994) elaborates the internal structure of complex terms, based on the linguistic theory of word formation. She introduces the idea of level-ordering to describe English term formation patterns, i.e. level 0 for non-native (neoclassical) compounding, level 1 for class I affixation, level 2 for class II affixation, and level 3 for native compounding. She also recognises that some elements are mainly used for forming terms, and marks them as such, which can then be used in the recognition of termhood in the relevant grammatical construction.

Enguehard & Pantera (1994) uses both statistic and linguistic methods. First, simple terms are extracted from the corpus, according to the frequency. They are then used in combination

with linguistic heuristics and frequency considerations to extract other term candidates. Three basic ways of extracting term candidates are proposed, i.e. (1) when two known terms appear frequently together, they constitute a complex term, (2) when a word frequently appears with some known terms in some specific manner, the word becomes a simple term, and (3) when a word appears frequently with a known term they together constitute a new complex term. Although (1) and (3) are based on the same basic linguistic considerations as the other work, (2) is different in that it refers to external relationships with known terms as the key for recognising terms.

3.1.3 Summary of the Linguistic Approach

Before summarising the approach, let us define two concepts which so far have not been clearly defined, i.e. ‘unithood’ and ‘termhood’. ‘Unithood’ refers to the degree of strength or stability of syntagmatic combinations or collocations. Thus ‘unithood’ is not only relevant to complex terms, but to other complex units as grammatical collocations or idiomatic expressions. On the other hand, termhood refers to the degree that a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts. Thus ‘termhood’ is not only relevant to complex linguistic units, but also to simple units.

According to this distinction of unithood and termhood, the primary concern of most of the work reviewed above is unithood of terms, thus focusing on complex terms. They seem to agree on one point, i.e. terms are linguistically materialised as noun phrases, though the degrees of descriptive minuteness differ, from simple parts-of-speech based description of collocational patterns to detailed structural considerations or patterns of term formation. At the processing stage, accordingly, some use simple pattern-matching while other assume a full parse.

There are two aspects where researchers take different viewpoints: one, concerning formal variations and the other, concerning discontinuous units. Although not all the studies mention it, most of them allow for basic singular-plural variation. However, as for such variations as ‘book review’ and ‘review of books’, some (e.g. Dillon & Gray (1984) and Daille, Gaussier & Langé (1994)) consider them to be variations of the same unit, while others (e.g. Justeson & Katz (1995)) apparently do not. The possibilities of discontinuous units are most notably addressed by Salton (1988), though his aim is indexing.

As for termhood, most take a pragmatic standpoint, simply admitting noise or leaving the final decision to human evaluation. Accordingly, many use frequency of the whole unit or a part as a measure for termhood. Ananiadou (1994) uses the existence of certain elements as a key for deciding termhood. This implies that terms are formed in the same way as words but sometimes with different substances. Enguehard & Pantera (1994) combines them in two steps, first evaluating termhood of simple terms by frequency and then uses them together with the

frequency of candidate patterns of complex terms. The assumption is that complex terms tend to be made from existing simple terms.

3.2 Statistically-Oriented Approach

3.2.1 Overview of Major Work

Some statistical studies for recognising complex units take a straightforward standpoint concerning the unithood of complex units¹⁰. For instance, Salton, Yang & Yu (1975) indicates the method of extracting and weighting complex index terms by simply extracting two adjacent words (or two words with one intervening element)¹¹, and give weights on the basis of their constituent elements¹²:

$$I_{(ik)j} = \frac{f(w_i) \cdot f(w_k)}{2} \cdot \left(\log_2(n(D)) - \frac{\log_2 \sum_j g(w_{ij}) + \log_2 \sum_j g(w_{kj})}{2} \right) \quad (13)$$

Damerau (1993) derives a measure for evaluating two-word units from a mutual information measure (Church & Hanks 1990). The mutual information for w_i and w_k is defined as:

$$I(w_i, w_k) = \log_2 \frac{P(w_i, w_k)}{P(w_i)P(w_k)}$$

where $P(w_i)$ and $P(w_k)$ are the probabilities of w_i and w_k , and $P(w_i, w_k)$ is the probability that the words w_i and w_k occur together (adjacently).

Assuming that w_i and w_k occur independently in the corpus, while $P(w_i, w_k)$ is biased to a certain subset of the corpus, he introduces the following measure:

$$I_{(ik)s} = \log_2 \frac{P_s(w_i, w_k)/P(w_i)P(w_k)}{P(w_i, w_k)/P(w_i)P(w_k)} = \log_2 \frac{P_s(w_i, w_k)}{P(w_i, w_k)} \quad (14)$$

where $P_s(w_i, w_k)$ is the probability that w_i and w_k occur together in the subset s of the corpus.

Jones, Gassie & Radhakrishman (1990) describe their INDEX system, which extracts repeated sequences of words and gives weights. For a complex unit of length N , i.e. $w_{i_1}w_{i_2} \dots w_{i_N}$, the weight is given by the following equation:

$$I_{i_1 \dots i_N} = \sum_{i=i_1}^{i_N} f(w_{i_N}) \cdot f(w_{i_1}w_{i_2} \dots w_{i_N}) \cdot N^2 \quad (15)$$

They explain that the measure is strictly empirical, chosen from some competing measures, e.g. those using N or N^3 instead of N^2 . A similar but much simpler approach is adopted by Steinacker (1974), which uses phrase frequency as the weighting measure.

Cohen (1995) starts from weighting sequences of n -characters (n -grams). The basic procedure to obtain term weight is: (1) count n -grams both for a document and the whole corpus (called ‘background’), and give weight to the n -grams for the document (n set to 5), (2) assign the value of each n -gram to the character at the center of the n -gram, (3) extract words

which includes a character exceeding the threshold, or word sequences when the adjacent edge characters of their constituent words both exceed the threshold, (4) give these words or word sequences a score by calculating the average score of the characters they contain.

For weighting n -grams of a document with respect to background, he uses a log-likelihood ratio, which is defined as follows:

$$\Psi_i = c_i \log(c_i/s) + b_i \log(b_i/r) - (c_i + b_i) \log[(c_i + b_i)/(s + r)] \quad (16)$$

when $c_i/s \geq b_i/r$, and 0 otherwise. Also, c_i and b_i are the frequencies of the i th n -gram in the document and in the background, and s and r are the frequencies of all the n -grams in the document and in the background, respectively.

Kit (1994) proposes a method of extracting multi-word basic text units (BTUs, which roughly correspond to terms as terminologists understand) from noun phrases extracted by the CLARIT system mentioned above (Evans 1995). He uses a few criteria, e.g. a BTU must occur as an independent NP, structural dependency of the constituents of a BTU must be kept throughout the occurrences. The structural dependency is obtained statistically from the occurrences in noun phrases. For instance, the structural dependency (*sdep*) of a word sequence $w_1 w_2 w_3$ is decided by the following rule:

$$\text{take } sdep(w_1, w_2) \text{ if } f_{np}(w_1, w_2) > 1.5 \cdot f_{np}(w_1, w_3) \text{ otherwise } sdep(w_1, w_3)$$

where $f_{np}(w_i, w_k)$ is the co-occurrence frequency of words w_i and w_k in noun phrases.

Frantzi & Ananiadou (1995) use two statistical measures in addition to the linguistic constraints for extracting complex terms. One is mutual information described above, extended to accomodate more than three words. The other is cost criteria, introduced by Kita, Kato, Omoto & Yano (1994):

$$K(a) = (|a| - 1) \cdot (f(a) - f(b)) \quad (17)$$

where a is a word sequence, b is every word sequence that contains a , and $|a|$ is the length of a .

Frantzi, Ananiadou & Tsujii (1996) introduces the measure called C-value, which is applied after noun phrases have been extracted by simple syntactic rules:

$$C - value(a) = f(a) - \frac{t(a)}{c(a)} \quad (18)$$

where a is the examined n -gram (sequence of n words), $f(a)$ is the frequency of a in the corpus, $t(a)$ is the frequency of a in longer candidate terms, and $c(a)$ is the number of these longer candidate terms.

Daille, Gaussier & Langé (1994), in addition to the linguistic criteria described in 3.1.2, tested some statistical measures to give weights to the base MWUs, e.g. frequency, log-likelihood

ratio, and mutual information. The log-likelihood ratio is basically the same as the one used by Cohen (1995), but the application is different. In case of Daille, Gaussier & Langé (1994), it is defined and applied as follows (Dunning 1993):

$$\begin{aligned}
 & -2 \log \lambda \\
 & = 2[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)] \quad (19)
 \end{aligned}$$

where k_1 is the frequency of w_i followed by w_k , n_1 is the frequency of w_k , k_2 is the frequency of w_i followed by the words other than w_k , and n_2 is the frequency of the words other than w_k . Also,

$$\begin{aligned}
 \log L(p, n, k) &= k \log p + (n - k) \log(1 - p) \\
 p_1 &= \frac{k_1}{n_1} \\
 p_2 &= \frac{k_2}{n_2} \\
 p &= \frac{k_1 + k_2}{n_1 + n_2}
 \end{aligned}$$

They indicate that frequency and log-likelihood ratio are possible quantitative measures to augment the syntactic analysis.

3.2.2 Summary of the Statistic Approach

These studies are roughly classified into two, according to which aspects of the units they measure, i.e. termhood or unithood, though this distinction is not completely straightforward as we shall see below.

Salton, Yang & Yu (1975), as well as Salton (1988) and Evans (1995) examined in 3.1, are typical approaches which give termhood weights to complex units. In the case of Salton, Yang & Yu (1975), unithood is not really considered as they mechanically extract two adjacent words. Damerau (1993) starts from mutual information, which is a typical method of measuring collocational strength, i.e. unithood, of word sequences. However, what Damerau (1993) proposed is a measure of termhood, essentially comparable to the measures introduced in Damerau (1965) (e.g. equation (5) in 2.2.3). Like Salton, Yang & Yu (1975), Damerau's (1993) measure is not concerned with unithood.

The measure introduced by Jones, Gassie & Radhakrishman (1990) is also inclined to the measurement of termhood rather than unithood. Although the theoretical interpretation is not easy, this point can be seen for instance from the fact that the independent occurrences of constituent elements are positively taken into account. Cohen's (1995) measure is also difficult to interpret. Although the weights given to n -grams are comparable to those in 2.2.3 and thus

related to the concept of termhood, the subsequent procedures to weight and recognise simple and complex terms make the meaning of weight unclear.

Kit (1994), Frantzi & Ananiadou (1995), Frantzi, Ananiadou & Tsujii (1996) and Daille, Gaussier & Langé (1994), on the other hand, apply the measures which have been proposed to quantify the collocational strength, or unithood, of word-sequences in general, to recognise terms.

As for the validity of statistical methods or models, most of them unfortunately exhibit the same problem as we identified in earlier approaches to single-term indexing. The log-likelihood ratio adopted by Cohen (1995) and Daille, Gaussier & Langé (1994) is considered to be statistically valid for binomial distributions with low-occurring events (Dunning 1993), and is thus in a stronger theoretical position. However, in the case of Cohen (1995), this is adopted at n -gram characters, thus it cannot be said to be statistically valid with respect to terms.

4 Summary and Discussion

In the process of reviewing individual studies, we did not give their detailed background, especially with respect to their practical and engineering status. Nor did we compare their performance. Instead we focused on the basic idea underlying this work and the methods they propose. While fully admitting that our present approach is not completely fair in evaluating each of these studies, we believe it is the best way of drawing a consolidated picture of what has been done so far. Although we give brief summaries and examinations in each section, we here want to summarise the important points and problems raised by these studies.

4.1 Nature of Terms

Firstly, many characteristics of terms, which we can call linguistic in its broadest sense, are implicitly or explicitly introduced by ATR studies, which are related to one another in complex ways. They can be summarised as follows:

Characteristics of Terms by Linguistic Structure:

By linguistic structure we mean here the internal structure of complex terms as seen from the point of view of the linguistic system. This point is addressed in most linguistic approaches. All of them agree that most terms take the form of noun phrases (especially those without prepositional modifications in the case of English or French). Most descriptions are relatively simple and straightforward, by the standard of recent linguistic theory. Ananiadou (1994) went further than that, while Daille, Faussier & Langé (1994) touched on more general patterns of term formation. Some studies treat conceptual patterns of complex terms, e.g. Kageura (1993),

but they have not been successfully utilised in ATR work.

Another point raised is that of the possible formal variations, including the continuity or discontinuity of terms. The problem of surface adjacency vs. structural dependency concerning unithood also belongs to the same type of problem. These are also partly related to usage of terms. Most studies are IR-oriented which accept grammatical (not inflectional) variations or discontinuous elements as possible terminological units.

Characteristics of Terms by Usage:

All the statistical work in ATR explicitly or implicitly adopts some assumptions concerning quantitative aspects of term usage. The assumptions are roughly parallel to those summarised in 2.3, which are re-stated from an ATR viewpoint as follows:

- (1) A unit which appears frequently in a domain is likely to be a term of that domain.
- (2) A unit which appears only in one domain is likely to be a term of that domain.
- (3) A unit which appears relatively more frequently in a specific domain than in general is likely to be a term of that domain.
- (4) A unit whose occurrence is biased in some way to (a) domain(s) is likely to be a term.

All these seem to be intuitively reasonable. They remain, however, no more than reasonable ideas so far, and the task of proper theorisation is yet to be carried out.

Another point raised by some studies is the relationships between linguistic-form and usage. Should we admit the different but related syntactic forms as usage variations of the same term? If yes, to what extent are variations admitted? Do typical syntactic variations such as ‘book review’ and ‘review of books’ show the same structural dependency? Is the concept of the same structural dependency really valid for recognising formal variations of terms? As mentioned above, this is an interesting question closely related to the structure and formation of terms.

Enguehard & Pantera (1994) also uses the possible relationships of the usage of two terms, though in an extremely restricted range. This in fact raises the question of how we regard the corpus, i.e. just as an unordered set of words or something more structured.

Characteristics of Terms by Termhood:

As long as a term is understood to be some sort of linguistic unit with termhood, and as long as the above mentioned aspects of characterisation are about terms as distinct from corresponding general linguistic units, all the characterisations are by definition characterisations of terms by termhood.

What we mean here is more restricted, i.e. unithood and termhood with respect to complex units and their constituent elements. From this point of view, we see several different ideas:

- (1a) The termhood of a complex term has no relationships with the termhood of its constituent elements.
- (1b) The termhood of a complex term is at least partially related to the termhood of one or more of its constituent elements.
- (2) The termhood of a complex term has no relationships with that of its constituent elements, but is related to how strong they are combined together, i.e. unithood is termhood.

The pursuit of these ideas leads us to the well known problem of term formation and word formation, e.g. the relationships between classificatory principle and syntagmatic rule in term formation.

4.2 Methodological Background of Term Extraction

Regarding the theoretical and methodological background of term extraction, specifically when statistical methods are involved, we can distinguish two aspects, at least conceptually, i.e. (1) the validity of statistical methods in itself with respect to the nature of data to be treated, and (2) the validity of the way statistical methods are applied with respect to the characteristics of terms to be clarified by the term extraction process. Little needs to be said about (2) now that we have examined the general meaning of statistical measures and summarised the various characterisations of terms assumed in the statistical approach.

Point (1), part of which has already been examined in 2.3.2 and 3.2.2, is not only a concern of ATR but of corpus-based CL work in general (Dunning 1993; Lauer 1995). At least two points need examination, i.e. the validity of statistical methods or models, and the validity of data or corpora. The latter is then related to point (2) above.

As for the validity of statistical methods or models, we have seen that many use intuitively reasonable but mathematically unfounded measures. This has long been recognised and tackled in automatic indexing, and recently in computational linguistics, and has led to some progress. This fact has always to be kept in mind even in the most pragmatic approach, because, without a proper theoretical framework, it is difficult to accumulate and utilise the output of previous work. From the weighting measures we can see that with time more complex measures were proposed, but we cannot really know whether there has been a real advance there, though this is not to deny the practical importance of the various studies.

As for the data or corpora used in individual work, we have said little so far because the methods examined are either claimed to be data-independent or themselves show what types of data are to be used. In most indexing and ATR work, the selection of data is based on practical considerations or availability. In actual experiments, some use abstracts, some use

individual article, some use a set of articles or books, etc. But whether the same method is really applicable to different types of data is not clear.

5 Conclusion and Outlook

This review gives a brief overview of the state-of-the art of ATR work, as well as the possible directions for further research related to ATR. From a more traditional terminological point of view, ATR has most typically shown the potential contributions and problems of corpus-based terminological research.

Undoubtedly the practical ATR work will find some useful application areas such as aids for terminology compilation or information retrieval, and on the way suggest various ideas concerning the nature of terms. To what extent ATR work has to consider the theoretical aspects of the characterisations of terms or the validity of its methodology is, of course, a matter of practical decision for each individual ATR application. For individual applications, there is no guarantee that ATR systems which are based on a more plausible theory perform better, especially given the practical situation that the result of ATR is evaluated in the wider context of IR or machine translation.

However, viewing ATR as a whole, an overall framework within which different approaches are compared or evaluated from comparable points of view is badly needed. It is necessary, therefore, to pursue theoretical work concerning various characterisations of terms assumed or proposed in ATR research. Some of the most notable are:

- **Unithood of terms:** There are many studies concerning the syntagmatic structure of complex terms, but most of them are based on pre-existing list of terms. In relation to ATR and to corpus-based study, such topics as formal variations or discontinuous occurrence of complex terms are addressed. They need serious consideration from a theoretical point of view. It is possible that rather static description of the structure of terms can benefit from studies of their actual usage in texts and so produce a more dynamic characterisation of term structure or term formation patterns, which in turn contribute to practical ATR.
- **Termhood of terms:** Much has been said about the basic nature or termhood of terms, in the form of abstract definitions of terms. But when it comes to the identification or characterisation of termhood of individual terms or actual terminology, this problem has rarely been satisfactorily elaborated (see, for instance, Kageura (1995)). When it becomes necessary to examine the concept of termhood of terms vis-à-vis non-terms in certain corpora in ATR applications, the problem arises again. While this is a real problem for ATR, it provides a suitable opportunity for a proper examination of the concept of

termhood with concrete data.

- The usage of terms in various types of texts or corpora. As far as we know, little if any work has been done concerning this point from a terminological point of view. Nor has ATR addressed this, as we saw above. There are many topics to be addressed, beginning from the difference of quantitative distribution of terms in different types of texts or domains to the occurrence characteristics with respect to the functional or grammatical positions of texts.

It is only after these points are examined and clarified that the evaluation of ATR can, besides establishing the practical validity and effectiveness of the results, lead to valid conclusions on the underlying theory and the methodology applied. From the terminological point of view, it is not surprising that these are essential theoretical problems, because, as one of the authors has argued (Kageura 1995), the proper theorisation of terms has to take into consideration the wider context including non-terms or discourse, which is precisely the condition from which practical ATR work has to start.

Notes

1. The authors would like to thank many people who read and commented on the draft, especially Professor J.C. Sager of UMIST. The first author would also like to thank Professor Yorick Wilks and all the staff of the Department of Computer Science, University of Sheffield, for giving him a wonderful research environment.
2. In fact many researchers on both sides begin to recognise the relations between these two fields from various perspectives, e.g. Grefenstette (1994), Kit (1994), Lewis & Sparck-Jones (1993), Salton (1989), Wilks, Slator & Guthrie (1995). As for ATR or automatic indexing, occasional reference exists on both sides (e.g. Justeson & Katz (1995), Kit (1994)), but they are fragmentary.
3. On the other hand, the process of choosing surrogates for user requests is called query or search formulation (Boyce, Meadow & Kraft 1994:103).
4. There are many textbooks on IR which give an overview of automatic indexing research including the work not covered here, e.g. van Rijsbergen (1979), Salton & McGill (1983), Salton (1989). Salton & Buckley (1988) also gives a brief overview of automatic indexing.
5. Other notations can be defined to make each formula less complex at first glance, while some can be simplified, e.g. $f(w_{ij})$ into f_{ij} . However, we decided to keep the notation minimal and use standard mathematical notations, because this is a much easier way of showing the meaning of the formulae.
6. In fact, some studies talk of ‘content words’ or ‘speciality words’ instead of index terms, e.g.

Dennis (1967), Stone & Rubinoff (1968), Nagao, Mizutani & Ikeda (1976). Damerau (1990) actually applies the measure proposed for automatic indexing to extract ‘domain-oriented vocabularies’ from a set of texts classified according to subjects.

7. To be strict, termhood is a characteristic of terms, not of index terms. For simplicity, however, we use termhood of an index term to refer to the characteristic of the index term of representing and discriminating the topical concept of the document.

8. It should be emphasised again that the measures introduced in this section are only those of direct relevance to ATR. In IR in general and in indexing in particular, much effort has been devoted to overcoming this problem, which, however, is not directly relevant here. Some of the directions are briefly sketched in Lewis & Sparck-Jones (1993)

9. In fact, as long as the content of a document is represented by a set of linguistic symbols, the effort of extracting complex indexing units must be correlated with a particular viewpoint of structuring document content. This applies to all the paths for more elaborated indexing methods. For instance, discarding the assumption of independence of words and use co-occurrence or context to calculate term dependence or relations (e.g. van Rijsbergen 1977; Yu, Buckley, Lam & Salton 1983; Deerwester, Dumais, Furnas, Landauer & Harshman 1990; Grefenstette 1994) can be interpreted as an approach to representing the structure of documents contents by relating simple words.

10. It is therefore closer to or basically the same as single-term indexing methods, and included here only for convenience.

11. In their paper they suggest taking adjacent words from a query to the IR system, thus it is not directly applicable to ATR applications. Here, however, we abstract the difference away for the sake of explanation.

12. The notations f , g and n are to be understood by analogy with those defined in section 2.

References

- Ananiadou, S. 1994. ‘A Methodology for Automatic Term Recognition.’ *Proceedings of COLING94* 1034–1038.
- Bookstein, A. and Swanson, D. R. 1974. ‘Probabilistic Models for Automatic Indexing.’ *Journal of the American Society for Information Science* 25(5), 312–318.
- Bookstein, A. and Swanson, D. R. 1975. ‘A Decision Theoretic Foundation for Indexing.’ *Journal of the American Society for Information Science* 26(1), 45–50.
- Bourigault, D. 1992. ‘Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases.’ *Proceedings of COLING92* 977–981.
- Boyce, B. R., Meadow, C. T. and Kraft, C. T. *Measurement in Information Science*. New York:

Academic Press.

Carroll, J. M. and Roeloffs, R. 1969. 'Computer Selection of Keywords Using Word-Frequency Analysis.' *American Documentation* 20(3), 227–233.

Church, K. W. and Hanks, P. 1990. 'Word Association Norms, Mutual Information, and Lexicography.' *Computational Linguistics* 16(1), 22–29.

Cohen, J. D. 1995. 'Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting.' *Journal of the American Society for Information Science* 46(3), 162–174.

Cooper, W. S. and Maron, M. E. 1978. 'Foundation of Probabilistic and Utility Theoretic Indexing.' *Journal of the Association for Computing Machinery* 25(1), 67–80.

Dagan, I. and Church, K. 1994. 'Termight: Identifying and Translating Technical Terminology.' *Proceedings of the Fourth Conference on Applied Natural Language Processing* 34–40.

Daille, B., Gaussier, E. and Langé, J-M. 1994. 'Towards Automatic Extraction of Monolingual and Bilingual Terminology.' *Proceedings of COLING94* 515–521.

Damerau, F. J. 1965. 'An Experiment in Automatic Indexing.' *American Documentation* 16(4), 283–289.

Damerau, F. J. 1990. 'Evaluating Computer-Generated Domain-Oriented Vocabularies.' *Information Processing and Management* 26(6), 791–801.

Damerau, F. J. 1993. 'Evaluating Domain-Oriented Multi-Word Terms from Texts.' *Information Processing and Management* 29(4), 433–447.

Dennis, S. F. 1967. 'The Design and Testing of a Fully Automated Indexing–Searching System for Documents Consisting of Expository Text.' In Schechter, G. (ed.), *Information Retrieval — A Critical Review*. Washington D. C.: Thompson Book.

Dillon, M. and Gray, A. 1983. 'FASIT: Fully Automatic Syntax-based Indexing.' *Journal of the American Society for Information Science* 34(2), 99–108.

Dunning, T. 1993. 'Accurate Methods for the Statistics of Surprise and Coincidence.' *Computational Linguistics* 19(1), 61–74.

Earl, L. L. 1970. 'Experiments in Automatic Extracting and Indexing.' *Information Storage and Retrieval* 6(X), 273–288.

Earl, L. L. 1972. 'The Resolution of Syntactic Ambiguity in Language Processing.' *Information Storage and Retrieval* 8(6), 277–308.

Edmundson, H. P. and Wyllys, R. E. 1961. 'Automatic Abstracting and Indexing — Survey and Recommendations.' *Communication of the ACM* 4(5), 226–234.

Enguehard, C. and Pantera, L. 1994. 'Automatic Natural Acquisition of a Terminology.' *Journal of Quantitative Linguistics* 2(1), 27–32.

Evans, D. A. and Lefferts, R. G. 1995. 'CLARIT-TREC Experiments.' *Information Processing and Management* 31(3), 385–395.

- Frantzi, K. T. and Ananiadou, S. 1995. 'Statistical Measures for Terminological Extraction.' *Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data (JADT 1995)* 297–308
- Frantzi, K.T., Ananiadou, S., and Tsujii, J. 1996. 'Extracting Terminological Expressions.' *The Special Interest Group Notes of Information Processing Society of Japan, 96-NL-112*, March 14-15 at Tokushima University, Tokushima, Japan.
- Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Boston: Kluwer Academic.
- Harter, S. P. 1975a. 'A Probabilistic Approach to Automatic Keyword Indexing — Part I. On the Distribution of Speciality Words in a Technical Literature.' *Journal of the American Society for Information Science* 26(4), 197–206.
- Harter, S. P. 1975b. 'A Probabilistic Approach to Automatic Keyword Indexing — Part II. An Algorithm for Probabilistic Indexing.' *Journal of the American Society for Information Science* 26(5), 280–289.
- Jones, L. P., Gassie, Jr., E. W. and Radhakrishnan, S. 1990. 'INDEX: The Statistical Basis for an Automatic Conceptual Phrase-Indexing System.' *Journal of the American Society for Information Science* 41(2), 87–97.
- Justeson, J. S. and Katz, S. M. 1995. 'Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text.' *Natural Language Engineering* 1(1), 9–27.
- Kageura, K. 1993. *A Conceptual Analysis of Japanese Complex Noun Terms with Special Reference to the Field of Documentation*. PhD Thesis. Manchester: University of Manchester.
- Kageura, K. 1995. 'Toward the Theoretical Study of Terms — A Sketch from the Linguistic Viewpoint' *Terminology* 2(2) 239–257.
- Kit, C. 1994. *Reduction of Indexing Term Space for Phrase Based Information Retrieval*. Interim Memo of Computational Linguistics Program. Pittsburgh: Carnegie Mellon University.
- Kita, K., Kato, Y., Omoto, T. and Yano, Y. 1994. 'A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria.' *Journal of Natural Language Processing* 1(1), 21–33.
- Klingbiel, P. H. 1973a. 'Machine-Aided Indexing of Technical Literature.' *Information Storage and Retrieval* 9(2), 79–84.
- Klingbiel, P. H. 1973b. 'A Technique for Machine-Aided Indexing.' *Information Storage and Retrieval* 9(9), 477–494.
- Lauer, M. 1995. 'Conserving Fuel in Statistical Language Learning: Predicting Data Requirements.' Interim Memo. North Ryde, Australia: Microsoft Institute.
- Lauriston, A. 1994. 'Automatic Recognition of Complex Terms: Problems and the TERMINO Solution.' *Terminology* 1(1), 147–170.

- Lewis, D. D. and Sparck-Jones, K. 1993. *Natural Language Processing for Information Retrieval*. Cambridge: University of Cambridge Technical Report 307.
- Luhn, H. P. 1957. 'A Statistical Approach to Mechanized Encoding and Searching of Literary Information.' *IBM Journal of Research and Development* 2(2), 159–165.
- Nagao, M., Mizutani, M., and Ikeda, H. 1976. 'An Automated Method of the Extraction of Important Words from Japanese Scientific Documents.' *Transactions of Information Processing Society of Japan* 17(2) 110–117. (in Japanese)
- Noreault, T. McGill, M. and Koll, M. B. 1977. 'A Performance Evaluation of Similarity Measure, Document Term Weighting Schemes and Representations in a Boolean Environment.' In Oddy, R. N. (ed.), *Information Retrieval Research*. London: Butterworths.
- van Rijsbergen, C. J. 1979. *Information Retrieval*. 2nd ed. London: Butterworths.
- Sager, J. C. 1990. *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.
- Salton, G. 1988. 'Syntactic Approaches to Automatic Book Indexing.' *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*. 204–210.
- Salton, G. 1989. *Automatic Text Processing*. Reading: Addison-Wesley.
- Salton, G. and Buckley, C. 1988. 'Term-Weighting Approaches in Automatic Text Retrieval.' *Information Processing and Management* 24(X), 513–523.
- Salton, G. and McGill, M. J. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw Hill.
- Salton, G. and Yang, C.S. 1973. 'On the Specification of Term Values in Automatic Indexing.' *Journal of Documentation* 29(4) 351–372.
- Salton, G., Yang, C.S. and Yu, C. T. 1975. 'A Theory of Term Importance in Automatic Text Analysis.' *Journal of the American Society for Information Science* 26(1), 33–44.
- Sparck-Jones, K. 1972. 'A Statistical Interpretation of Term Specificity and Its Application in Retrieval.' *Journal of Documentation* 28(1), 11–21.
- Sparck-Jones, K. 1973. 'Index Term Weighting.' *Information Storage and Retrieval* 9(11), 619–633.
- Steinacker, I. 1974. 'Indexing and Automatic Significance Analysis.' *Journal of the American Society for Information Science* 25(4), 237–241.
- Stone, D. C. and Rubinoff, M. 1968. 'Statistical Generation of a Technical Vocabulary.' *American Documentation* 19(4) 411–412.
- Wilks, Y. A., Slator, B. M. and Guthrie, L. M. 1995. *Electric Words: Dictionaries, Computers, and Meanings*. Cambridge, Mass: MIT Press.