

What is the role of NLP in text retrieval?

Karen Sparck Jones
Computer Laboratory, University of Cambridge

March 1997

Appeared in
Natural language information retrieval, Ed. T. Strzalkowski, Dordrecht: Kluwer, 1999, 1-24.

Abstract

This paper addresses the value of linguistically-motivated indexing (LMI) for document and text retrieval. After reviewing the basic concepts involved and the assumptions on which LMI is based, namely that complex index descriptions and terms are necessary, I consider past and recent research on LMI, and specifically on automated LMI via NLP. Experiments in the first phase of research, to the late eighties, did not demonstrate value in LMI, but were very limited; but the much larger tests of the Nineties, with full text, have not done so either. My conclusion is that LMI is not needed for effective retrieval, but has other important roles within information-selection systems.

The rapid growth of full text databases, together with developments in natural language processing (NLP) technology, has prompted those engaged with NLP to suggest that it could be usefully applied to text retrieval, primarily for indexing purposes but perhaps also for more or less related tasks such as document ‘abstracting’ or extracting; it could be applied at shallow text as well as at deep content levels, and for user display or for database creation. Retrieval itself has various modes, including filtering or routing as well as one-off searching; the family of information and text processing tasks in which it figures includes categorisation for various purposes; and the material dealt with extends, for example, into hypertext. The claim is that *linguistically-motivated analysis*, and hence NLP, is needed not only for tasks like information extraction but even, in the current demanding circumstances where vast volumes of machine-readable material are becoming available, for such simple ones as document retrieval. Further, NLP may also be needed to provide useful linkage between data and task types within the family of information-selection tasks, by supplying concept representations that may be exploited for related purposes (as with data and document query), as well as for more friendly information supply for the user (as with a display of key document concepts).

In what follows I shall concentrate on document (text) retrieval, returning briefly to other tasks later. Thus the questions to be examined are:

1. What indexing and searching devices depend on linguistically-motivated analysis?
2. How do these devices depend on such analysis?
3. Does automation (i.e. NLP) affect the type of device used, or way it is created and manipulated?

I begin with basic concepts and then consider the arguments for complex index descriptions requiring linguistically-motivated indexing. I then review experiments with providing and using such descriptions, making necessary comparisons with less exigent approaches. I conclude by considering my findings in a wider information-management context.

1 Linguistically-motivated indexing

1.1 Basic concepts

I shall refer to *index terms* and to *index descriptions* made up of terms. Such terms and descriptions are obtained by text processing in some way or other, whether of a document text (either full text or surrogate e.g. abstract text) or of a *request* expressing a user's information need. The terms and descriptions supplied for documents may be formed at *file time*, when a document enters the data base, or at *search time*, stimulated by a user's request. The terms and descriptions constructed from user requests, at search time, form one or more *queries*.

In general in modern, research-based systems as opposed to older conventional automated systems or traditional manual ones, there is more of a distinction between an *initial* document or request description and the *final* one that determines a document's matching score or 'retrieval status value'. This is because indexing and searching are tightly bound together in the manipulation of given descriptions so as to achieve matches. The major reason for this development has been the shift from *precoordinate* to *postcoordinate* indexing, i.e. from fixed to variable term combination. This corresponds to a move from file time to search time indexing, further emphasised by the adoption of techniques like statistical term weighting. With the arrival of full text, the need for a distinction between initial and final description becomes even more marked. The implications of search-time processing for the applicability and utility of a linguistically-motivated approach to indexing, even if it appears desirable in itself, are therefore important.

The notion of 'linguistically-motivated indexing' may appear quite straightforward, or the expression even pleonastic, but precisely how it is interpreted has a significant bearing on claims for its potential or actual utility. We can define an index term as a *content-bearing key* such that if document and request descriptions share a key the presumption is that the information content of the document has some relevance to the user's information need. An index description as a whole is then some combination of keys that provides a fuller indication of content and hence allows a closer approach to relevance. Then *linguistically-motivated indexing* means at least the application of text interpretation techniques designed to identify such keys and construct descriptions from them. But it is typically taken, more specifically and aggressively, to mean the application of syntactic and/or semantic techniques for term identification and description formation. In this paper I shall use linguistically-motivated indexing (LMI) in the latter stronger sense. In addition, since index terms need not be single words but may be multi-word units, while descriptions are often no more than simple sets of terms, LMI often refers to the use of linguistic analysis for identifying such terms and characterising their internal structure.

Thus we may suppose, to take a minimal example, given the text

There are urban centres. These are decaying.

LMI at its most ambitious could deliver an index description which we will represent (as it is to be understood) as

CENTRES that are URBAN are DECAYING

(thus making it clear that the index terms and descriptions derived from ordinary source texts are properly now to be viewed as being in a *document representation language*). The more limited version of LMI would just give

[CENTRES that are URBAN] [DECAYING] .

The presumption in all this is that the linguistic analysis is *explicit*. For example, analysis could identify an adjective-noun group, and label its members as modifier and head respectively for the multi-element term. Then, assuming the unit thus identified is not subject to further processing, its structure could be used, in the event of a failure to match on the term, to allow a match on the head only. Thus given some general assumption about the sort of content-bearing key needed for retrieval - in this case that they are grounded in such natural source units as noun groups, - LMI is a *direct* way of identifying such keys.

This all seems oppressively obvious. But matters are in fact more tricky. Take the text sentence:

Architects have made a lot of money from urban centre redevelopments.

Syntactic analysis would identify “urban centre developments” as a unit, giving the term

URBAN CENTRE REDEVELOPMENTS

with REDEVELOPMENTS marked as head. Full semantic and pragmatic analysis (assuming this were possible) could proceed further, and define a multi-element term with a marked bracket structure and supplied internal relations that we will represent as

REDEVELOPMENTS *of* [CENTRE *of the sort* URBAN] .

We thus have direct determination of an explicitly characterised term.

But this has now to be contrasted with the alternative *implicit, indirect* approach. This is illustrated by the simple strategy that fragments the example sentence by deleting stop (function etc) words and retaining remaining word strings as terms. This would give the fixed sequence

URBAN CENTRE REDEVELOPMENTS

as a term. An analogous, though somewhat less restrictive, effect is achieved when a query is constructed that specifies that individual words have to be adjacent or within some specified proximity, for example:

[URBAN =0 CENTRE] <4 REDEVELOPMENTS .

This strategy, often found in conventional retrieval systems, would not allow any words to intervene between URBAN and CENTRE, but would allow up to three between this pair and an occurrence of REDEVELOPMENTS either before or after it. It thus (retrospectively) defines a set of matching word strings for document texts. For convenience I shall label the style of indexing illustrated in these last two examples *non-linguistic indexing* (NLI),

while recognising that the distinction is not absolute, only relative. Thus the retention of all words but stop words as terms relies on facts about English as a language. More specifically, taking multi-element strings as terms depends on the linguistic basis for co-location. But by comparison with the LMI case, the linguistic grounding for selecting terms and, more importantly for forming index descriptions, is very vague. An NLI description will normally consist simply of an unstructured set of terms (indeed retaining the source text order for the terms in a description, even if considered useful, would barely invoke linguistic considerations).

There is, however, also another factor to consider. Indexing as so far described is on a *local* basis, i.e. words or word sequences generate terms if they occur in a document (or request) text. But experience has repeatedly shown that it is essential to take the *global* as well as local presences of source words into account in term generation. For direct (and explicit) terms the linguistic ground remains the same; but an additional constraint is imposed, namely that a *candidate* term only survives if it satisfies some frequency condition(s), signalling a larger and thus putatively more helpful content significance. Less stringently, statistical information may be used to weight, i.e. prefer one term to another. However further explicitly-linguistic processing may interact with global considerations. Thus we may specify equivalence classes of expressions, so the same term is derived from different initial text expressions (or candidate terms). For example:

$$[\text{ADJ} + \text{NOUN}] = [\text{NOUN BE ADJ}] .$$

(Of course this may be done in various ways; if term structure were already specified in the more abstract head-modifier style, it would subsume this equivalence statement.) Naturally, just as with the original decision that certain syntactic categories and structures are useful for indexing, we have to decide which forms shall be subject to *conflation*. The important point however is that here linguistic criteria interact with frequency ones in relation to global term significance, since the definition of equivalence will influence the resulting term frequencies.¹

Statistical criteria may clearly also be applied to the complementary, NLI, case. Again, this may be limited to selection or weighting. But we may choose, more subtly, to apply this information to check the validity of candidate terms or to derive new, more solid terms. For example we may reject very infrequent word strings on the grounds that they are spurious units and hence not proper content indicators; or we may choose to select shared substrings from overlapping strings on the grounds that these are likely to be proper or potentially valuable indicators. Just as with the initial way of proposing terms, such procedures have some underlying linguistic rationale, and would only be expected to work because they have such a rationale.

Overall, and as a *general* statement, the distinction between LMI and NLI as I have defined these applies at the local level, and to identifying index terms and forming descriptions for individual source texts. Exploiting global information derived from more than one document and in particular statistical information based on whole files of documents may, in the present context, be viewed as applying NLI overlays on candidate terms and descriptions. This nevertheless still leave open whether the result is an initial description or a final one as far as the interaction between indexing and searching is concerned.

Finally, the contrast between LMI and NLI applies even when single word terms are concerned, as indeed illustrated by the earlier ‘Architects making money’ example. Thus with

¹Taking terms only if they satisfy some text-locational or other other discourse condition can be viewed either as a delimitation of indexing sources or as a further refinement of LMI.

LMI, ARCHITECTS and MONEY would be taken as terms only if they met conditions on syntactic category or sentence function, while with NLI as illustrated they would be automatically adopted as terms. However, as the next section makes plain, there is less manifest requirement for LMI as opposed to LNI if terms are based on single words. The main modern rationale for LMI is in capturing multi-element terms effectively. The single-word examples nevertheless emphasise the fact that NLP could have two distinct roles. One is simply to identify those word(s) in the source text that generate, or are adopted as, index terms. The second goes beyond this, in characterising the source items in such a way, particularly by defining their structural properties, that this information can be exploited in the form of document description and that is created in searching.

Linguistic analysis may of course also be applied at the subword level, for morphological decomposition and *stemming*. This is a way of defining word equivalence classes. The value of LMI here, whether stemming refers to a lexicon or makes use only of an affix-stripping algorithm (Porter, 1980), is much less controversial than the cases considered so far. The experimental evidence, accumulated over a long period, shows that stemming even when only algorithm base is normally, though not necessarily largely, a positive contributor to retrieval performance (Harman, 1991; Hull, 1996). I shall therefore assume in what follows that in LMI stemming plays its own considered role as an integral part of the larger linguistic processing; while in NLI it is the only explicitly linguistic operation, but again interacting with other procedures. From this point of view, the earlier examples might have been alternatively shown as

[CENTR* that are URBAN] [DECAY*]

or

URBAN CENTR* REDEVELOP* .

In later examples the possibility of stemming will be taken for granted even if it is not shown.

1.2 Complex descriptions and terms

I have so far, in discussing the two approaches to index terms and descriptions given by LMI and NLI respectively, considered these solely from the point of view of their derivation from a source document or request text. It is now time to look more carefully at the claim that LMI rather than just NLI is required for effective retrieval, by examining the *functions* terms and descriptions are intended to have.

This is very usefully done in a historical perspective. Essentially, the assumption in indexing has been that as documents (texts) deal with conceptual structures, i.e. concepts and the relations between them, index descriptions will only be adequate for the task of leading the user to the documents, and hence to the information they contain, if the descriptions indicate this conceptual structure. Two further assumptions have also been made. One, very natural in the days of book indexing but also with its own justification, is that index descriptions should concentrate on the source's main concepts, i.e. should be selective from or reductive of the source text content. The other assumption is that as natural language is highly varied as a means of expression, so shared underlying concepts may surface in different expressions and in particular differ between document and request texts, index descriptions should be *normalised*. Index terms and descriptions should be lexically unambiguous and

structurally regular, achieved by the application of an *indexing language*, and specifically a *controlled language*.

Simplifying grossly, this view is seen in its classical form in the use of such major indexing languages as the Library of Congress Subject Headings (Chan, Richmond and Svenonius, 1985)). Thus the heading **Church architecture - Italy** may be assigned as a description to a document on the architecture of churches in Italy. The internal structure of the heading has two roles, both aimed at promoting precision in retrieval output. The first is to ensure that those documents and requests that involve architecture and churches and Italy only match if the conceptual structure in both cases is that of the heading and e.g. the request is not about the influence of Italy on church architecture. The second is to constrain description modification if no match is obtained. Thus a new request description might be formed consisting just of **Church architecture**.

Clearly, human understanding is required to provide such descriptions, and has been in the past; and NLP would correspondingly be needed for automated indexing (assuming this were feasible, which the reductive aspect makes especially difficult).

With the growth of the scientific literature in particular, such highly reductive, aggressively normalised, precoordinate descriptions were found to be both insufficiently discriminating as a means of differentiating documents and over-constraining in not allowing a range of approaches to them. What I have labelled conventional, as opposed to traditional, systems have tended to compromise by constructing document descriptions from separate terms, which may individually be multi-element ones but which are combined in the description only as a set. Lexical and structural normalisation thus applies primarily to the terms, provided as *descriptors* by a *thesaurus*, for example **Infrared radiation** (Euratom) - read as **radiation of type infrared** - (Lancaster, 1972). However while a document description has such an initial structure, the actual final structure of a description is defined at search time via the matching conditions imposed or allowed by the query, within a general postcoordinate framework. This may select terms either via a strict Boolean expression or in a more relaxed way via simple coordination and quorum matching for a ranked output. From one point of view such operator relations are normalising ones, but linguistically-significant normalisation is within rather than between terms.

I shall use the expression *complex description* for cases where terms from the indexing vocabulary are tied by linguistic relations like case-type ones. Where the terms themselves may have an internal structure indicated by linguistic relations, I shall refer to these as *complex terms*. Clearly this is not an absolute distinction: for instance a complex description may be formed from a single complex term. But alternatively, within-term and between-term relations may be of different types; and the distinction is convenient given that the perceived value of fuller and more flexible descriptions has led to the use of simply coordinated terms, rather than fully-integrated structures, as descriptions. When a previously-defined term vocabulary is available, complex terms may be assigned to documents in statistical grounds (e.g. Biebricher et al., 1988). But in general it is reasonable to assume that if complex terms are used, even if not in complex descriptions, NLP would be required to determine these for a document, i.e. to ground assignment.

It is important to recognise what an enormous range of index language designs (and hence of description forms) has been advocated in the past, and how many and varied the languages actually in use are, especially those represented by controlled vocabularies (thesauri, subject heading lists) that allow complex terms (Lancaster, 1972; Hutchins, 1975; Chan, Richmond and Svenonius, 1985). Methods of indicating *syntagmatic* i.e. temporary connective or as-

ceptual relations, as with the use of so-called links and roles and of faceted schemes, have been extensively studied and widely applied: work on the use of relations between terms has not been confined to the more familiar *paradigmatic* case defining permanent classificatory relations. In traditional and conventional systems the latter may, however, also have syntagmatic implications. Thus where a controlled vocabulary has complex terms but these are treated as fixed precoordinate units, using hierarchical relations like that between **Infrared radiation** and **Radiation** to allow term substitution in searching is logically equivalent to a linguistically overt decomposition of **Infrared radiation** to extract **radiation**.

In the approaches so far described the index language, and especially its controlled vocabulary, is independently defined, even if the vocabulary is justified by reference to a subject's literature.² But constructing and maintaining a large controlled vocabulary is extremely effortful. More importantly, there are good arguments in favour of a more *derivative* approach, which is also more practicable for automation. Thus the vocabulary, including complex terms, of a system should be drawn from the documents in the file as better reflecting their real conceptual content, with control applied to the selection of terms and type of normalisation applied, as in stemming for example, but also association-based clustering. In such a situation, LMI would be essential not only for the description of an individual document but for building the indexing vocabulary, according to the language specification, that is applicable to a document file (and search requests). While the resulting vocabulary could in principle be the same as one created a priori, the expectation would be that it would be less controlled, especially in the degree of lexical normalisation covering synonyms, for example. This would be likely in practice, at any rate, whether or not it was thought desirable in principle. Thus given the use of complex terms but not necessarily complex descriptions, the purpose of LMI would be in identifying and characterising these in ways that allowed for normalisation and for *modulation* in searching when descriptions are altered, for instance by dropping modifiers, to achieve matches.

This traditional and conventional indexing background may not seem especially pertinent to modern text retrieval conditions and aims for LMI (regardless of what current NLP capabilities are). But it is useful for two reasons. One is that it provides a context for interpreting the results of research on NLP-based indexing so far. The other is that it emphasises the fact that it is always necessary to be clear what assumptions about retrieval are being made, whether these are used to justify an attempt to implement LMI, or to support analysis of actual performance if LMI does not live up to expectations.

2 Research and tests

For future discussion, since postcoordinate indexing is the dominant style, it is useful to make specific distinctions as follows, and to abandon the earlier loose notion of "multi-element" term. Thus I shall distinguish descriptions using *compound terms* from those using *simple terms* (single words, stems), and further distinguish, with respect to compound terms, *complex terms* from *joined terms*. The former are defined by LMI, the latter by NLI, i.e. by co-location, whether this is strict adjacency or proximity.³

Then in considering past tests, comparisons at the most general level are simply between LMI and NLI, regardless of whether descriptions are precoordinate or postcoordinate. At a

²Of course not surprisingly given the fact of manual indexing in the past.

³The expression "joined term" is deliberate because other common labels are variously used.

lower level, for the latter, the comparison is between compound and simple terms, Finally, for compound terms, there is a further comparison between complex and joined ones.

2.1 Phase 1 : Experiments from the 1960s to the 1980s

Research on NLP for retrieval is nearly as old as research on automated retrieval in general. Early work was clearly based on the goal of achieving, as far as possible, the same kind of indexing as was done manually. This is implicit in both Salton’s early research (summarised in Salton, 1968), and on the work on automating SYNTOL-type indexing done by Bely et al. (1970). These two projects are not only historically important but are exemplars, occupying somewhat different points in the space of possibilities mentioned earlier, of types of LMI that have been continuously invoked since (see e.g, Strzalkowski, 1994; Mauldin, 1991). It is therefore worth considering them in a little more detail.

Bely et al.’s work was the more ambitious of the two, in seeking a deeper analysis of the source text. Thus NLP was directed towards identifying manifestations of thesaurus concepts (as defined in a prior domain descriptor vocabulary), *and* their semantic (“logical”) relationships. The descriptors were simple, while the relationships were of a generic kind, for instance **cause**. The expressions (“syntagms”) constituting an index description were essentially atomic propositions with the relational predicates linking descriptor arguments. As the initial texts were already abstracts, the descriptions were reductive primarily through the concept selection imposed by the use of the thesaurus, rather than through any analysis of the text alone, but the degree of reduction was not necessarily large. In the limit, all the propositions in a description might be linked by shared arguments, giving a graph-style structure, though in practice fragments were more common. The main value of the connectivity was in allowing very weak inference (for instance transitive). In searching, descriptions could also be modulated by invoking the thesaurus taxonomy, or by generalising on individual relations (e.g. replacing **cause** by **association**. The NLP involved was very elaborate, since it was necessary to identify instances of thesaurus concepts and the relations between them in many guises. Bely et al. did not carry out any retrieval performance evaluation but found automatic and manual descriptions, when compared, fairly similar. However the SYNTOL work as a whole suggested that the relational structure in descriptions had to be weakened or disregarded to obtain sufficient matches, even though it was of only a simple and general kind to start with.

Salton also used a thesaurus, with syntactic analysis to establish that thesaurus terms were in a proper dependency relationship, as defined by “criterion trees” for types of syntactic structure, covering many possible specific text instantiations. The results of the analysis, so-called *syntactic phrases*, were therefore normalised complex units. Salton compared performance for these with *statistical phrases*, defined in opposition to syntactic ones as meaning simply cooccurrence of the constituents of a thesaurus descriptor within sentence boundaries. The conclusion Salton drew from his early tests was that syntactic phrases were no better than statistical ones, with the control given by the use of the thesaurus a more helpful contributor to performance, for example in comparison with simple word stems.

In both of these projects, the linguistic analysis was substantial. Hillman (1968) sought a similar indexing outcome by simpler means, using only partial parsing to identify “canonical components”, i.e. propositional structures centred on verbs from which nominal arguments could be extracted as terms. In contrast to the two previous projects, the analysis was applied to full texts. But there was no organised performance evaluation.

As noted, this early work sought if not to simulate at least to emulate human indexing: it was based on the same assumptions about what indexing should be like, i.e. that complex, normalising descriptions would be useful. Thus the relational normalisation achieved by complex terms could in principle be supported (if not facilitated by a reference thesaurus) by word clustering, though in practice work on term grouping was largely a separate enterprise and was not normally done using syntactically-constrained association contexts, or indeed co-location within narrower limits than whole abstracts.

But at the time in the Sixties when NLP-based indexing was proving effortful and not demonstrably valuable, the first serious evaluation tests were also being done with different types of (manual) indexing languages. Thus the Cranfield experiments were intended to provide concrete evidence for the relative merits of a range of indexing languages and devices. In fact, the Cranfield 2 tests (Cleverdon, 1967) showed that very different languages gave very similar performance, and that descriptions using heavily controlled languages with complex terms were no better than (manually supplied) natural language descriptions, even when these consisted only of simple terms (stems). Other similar experiments like Salton's comparing the MEDLARS system's controlled indexing with natural language showed a similar picture (Salton, 1972).

Such findings flew in the face of received wisdom about indexing and retrieval; and in the period up to the late Eighties, when the context and conditions for retrieval research began to change significantly, there were several attempts to apply NLP for LMI, at least to obtain complex terms; and indeed it was used in at least one operational system. NLP technology was less powerful than it is now, and certainly slower, but as in some cases only crude analysis was required the state of the NLP art was not too limiting. Thus Dillon and Gray (1983) applied a relatively shallow processing strategy to obtain complex terms. They used tagging, reducing tag ambiguity by tag-string constraints, and then applied a set of rules defining acceptable tag sequences in order to select phrases. Normalisation conflated phrases with the same members after stemming. Performance was somewhat better than that for simple terms. The DDC operational system (Klingbiel, 1973) used a similar strategy to process abstracts to find phrases as candidate terms for manual indexer consideration. The words in strings left after text segmentation at stop words were checked against a selective lexicon, and then strings against a set of syntactic formats. As a subsequent evaluation (Klingbiel and Rinker, 1976) implies, the results were presumably useful enough, given the vast numbers of documents to be processed, for operational purposes. These tests showed that even straight machine indexing could give respectable performance, while with editing plus suitable search strategies performance was comparable with that for manual indexing (though differently balanced between recall and precision).

The continuing improvement in NLP technology stimulated rather more work on NLI in the Eighties. Sparck Jones and Tait (1984), for example, used NLP to determine sentence propositional structures from which complex terms could be extracted. Normalisation was determined by equivalence classes. This was because analysis was applied only to requests, and the classes were used as the basis for generating alternative proper surface word strings to be sought in document texts. Linguistically, the work was quite ambitious, but it was not taken as far as systematic retrieval testing; and even with normalisation the method imposed constraints on the forms in which phrasal concepts could occur in documents. Smeaton and van Rijsbergen (1988) also processed only requests, but in much shallower style designed to identify nominal phrases, with co-location of phrase elements as the matching condition on documents. Retrieval testing here showed a modest performance improvement over simple

terms.

The work done by Fagan (Fagan, 1987; 1988), and by Croft, Turtle and Lewis (1991) can be taken as marking the end of this first phase of research.

Fagan's experiments consolidated earlier tests not only in exploring options more fully than before but also through his comparisons with earlier studies; and his analysis of his and other results is valuable both in its own right as a commentary on the work hitherto and as suggesting important points to be taken into account in any new work on indexing using compound terms.

Fagan's primary concern was to evaluate non-syntactic, i.e. joined terms, under alternative settings for the system parameters involved e.g. degree of proximity, frequency properties of term members, etc. Thus he lays out very clearly the various issues to be addressed in defining joined terms and in treating them in relation to their components and other single terms in matching. His results, across several collections, showed that joined terms could perform rather better than simple terms, though under different parameter settings for different collections. These tests were then complemented (for some of the collections) by indexing with complex terms obtained by applying the powerful IBM PLNLP parser, with subsequent normalisation by stemming and reduction to head-modifier term pairs. Such complex terms could give better (though not much better) performance than simple terms, but were either the same as or somewhat inferior to the joined terms. So the overall ordering of approaches was joined terms \geq complex terms \geq simple terms. Moreover after examining earlier work like Dillon and Gray's, Fagan found that when factors he concluded were important were taken into account, all of these tests conformed to the same general pattern.

Thus his conclusions were that compound terms should be restricted to pairs (i.e. candidate terms containing more elements should be decomposed); weighting should use $tf \times idf$ ⁴ and not idf alone; and compounds should be treated as *additions* not *substitutions* (i.e. their members should be retained as simple terms in their own right). These recommendations apply to both complex and joined terms, and have in fact been generally followed (whether or not under Fagan's influence) in later research with compounds.

But Fagan's analysis of why compound terms did not contribute more conspicuously to performance (the general effect was of order 5-10%) is also very valuable. He concluded that they could not contribute much, for his test data, simply because they were not shared between queries and documents: they could therefore not increase matching scores. This was primarily because the documents were all very short (abstract length). However this effect was reinforced by the fact that the best compounds had absolutely low collection frequency. Fagan nevertheless believed that LMI could be helpful in filtering out bad joined terms, and in capturing phrases beyond narrow proximity limits, while they might also be helpful if used to expand descriptions. Lewis (1991) subsequently attempted to compensate for low complex term frequencies by clustering, but to no effect because frequencies were also too low to generate useful cooccurrences.

Croft, Turtle and Lewis also lay out the questions to be asked about compound terms (phrases), for example what evidence there is for them in documents and requests, how they should be weighted, and again how they should be treated relative to their components and other single terms: they use their inference net model to illustrate types of treatment. But their experiments too (though on fewer collections than Fagan's) showed little gain from

⁴ tf = within-document term frequency; idf = inverse document frequency, a log function where document frequency is the number of documents in the file containing the term.

phrases of any sort and no relative merit in syntactic as opposed to non-syntactic ones even when the proximity constraints on the latter were as weak as mere cooccurrence in documents. However like Fagan they note that the results may be attributed to short documents and small files, and suggest that proximity-based joined terms, using a small window, may be useful for large files with longer documents.

2.1.1 Implications of Phase 1 results

Overall, the work done up to the end of the Eighties examined a whole range of analysis and representation methods. Indeed in further contrast to the work just described, Mauldin (1991) combined robust parsing with AI-type case frames, though his evaluation was too restricted for comparison with other tests, and Hahn (1990) in addition exploited discourse-level information. But when retrieval performance was compared with that given by other types of postcoordinate indexing, simple or compound but using NLI rather than LMI, there was no convincing demonstration that LMI worked better, or certainly much better, than NLI, or indeed that either was conspicuously superior to the simplest approaches confined to single words (stems).

However to offset this negative finding, three points need to be taken into account. The first is that, as Sparck Jones and Tait, Fagan, and Lewis all note, the range of system parameter settings, and of environment variable values, is so great that to ensure adequate control the number of tests required is very large. Thus it is at least possible that all the work missed good ways of doing and using LMI. The second point is that the tests were all on a very small scale. This not only made them suspect in themselves. It also meant that frequency effects could not be properly studied. In this connection Fagan pointed out that the notion of “term discrimination value” invoked in earlier SMART work to restrict common words to phrases (and thus to some extent acting as a statistical successor to a concept lexicon in filtering compounds) did not behave predictably. The third important point is that testing was normally with abstracts. This also affected the frequency data, and could in addition mean there might be little scope or need for sophisticated indexing: the content words in abstracts are ‘good’ in themselves, and may be sufficiently captured by simple or just joined terms.

But even while allowing for the effects of scale and document type, this work could also, under a broader view of things, prompt further observations as follows.

1. That insofar as NLP is intended to achieve the same results as human indexing, and assuming the latter really gives better performance than ‘raw’ automation, the NLP actually done was unsatisfactory or inadequate.
2. That regardless of the merits of human indexing, automated LMI could offer something different (for instance in more sensitivity to text usage), but the precise form it should take has not yet been determined.
3. That since retrieval performance may depend in a quite complicated way on the interaction between various devices e.g. weighting, query modulation, the right way of combining the information embodied in NLP-based terms with other information has not yet been established.
4. That even if NLP-based indexing could give better descriptions than others, whether human or NLI ones, the effect on search outcomes could be small simply because the

other factors affecting retrieval, like the character of the initial user request, or the use of feedback, dominate performance.

5. That on the assumption that NLP is successfully simulating/emulating human indexing, the reason why it does not deliver much is the same as that accounting for the lack of manifest superiority in ‘fancy’ human indexing, namely that retrieval does not in fact need complex descriptions or terms.

On the whole, though NLP is not able to achieve, for general text, the kind of deep interpretation that humans can, it is nevertheless fair to say there is no strong evidence for (1) or (2). It is true that the power of NLP up to the mid-Eighties was not very great, and that syntax without semantics has its limits, but NLP capabilities were adequate for the task of identifying simple constituents of the kind leading to complex terms. Again, setting the use of statistical information aside, it is not clear how far automated LMI could deliver something significantly different from humans: thus the independent arguments that complex terms should be generated by e.g. nominal groups is quite compelling (even if the mistaken assumptions that automatic term classifications should look like human ones are a salutary reminder that the obvious is not always the right).

We are thus left with (3) - (5), Point (3) is a good reason for continuing research, especially since, more generally, the interactions between different conditions and devices is not fully understood. Point (4) has a similar implication, to see what responses to major conditions, e.g. very short requests, can most usefully be made. Nor is point (5) an argument for abandoning LMI altogether, until it can be demonstrated, rather than believed, that this is the case.

The tests considered so far also raised other issues, for instance those connected with users’ performance requirements. Even allowing for normalisation, complex descriptions and terms are precision devices. In general, laboratory retrieval testing has considered both precision and recall, in a fairly even-handed way, and it is possible to argue that this does not capture what LMI is intended to achieve for the user. As against this, one of the persistent findings in retrieval research has been how much recall has to be sacrificed to achieve high precision. It is also not the case that LMI is the only viable strategy for achieving high precision.

2.2 Phase 2 : The Nineties

The important new factor affecting work on LMI at the end of the Eighties, along with retrieval research (and practice) in general, was the arrival of full text. Indeed it might appear that the older trials and results, just considered, are no longer interesting or pertinent to current research. As against this, it has yet to be shown that full text introduces such new conditions specifically bearing on indexing that the previous assumptions, arguments, strategies and tests no longer apply. Thus it is essential to ask what these new conditions are, whether they could make a difference to the observed value of LMI and, as part of this inquiry, to reexamine the assumptions on which LMI is based. Thus it is not sufficient to start, uncritically, from the same assumptions as were found wanting in the earlier period. At the same time, advances in NLP make it more reasonable to attempt serious and systematic LMI, even for masses of full text, than before. In particular, while there may not have been any major changes in, for instance, styles of syntactic representation, LMI should benefit from the much greater experience we have now of robust processing able to deliver at least partial analyses if complete sentence processing fails, and of the use of corpus data to establish

preferences in analysis. Much greater machine power is also bringing what were formerly unattainable processing scales and speeds within practical reach.

Thus with the arrival of full text, LMI appears to be needed for the following reason:

There is more detail in the documents, implying a need for both more refined index descriptions and more elaborate normalisation than before. The former is to satisfy the greater requirement for precision, the latter the concomitant one for recall.

But the reason for LMI is also its justification:

Because there is more detail in the documents, there is the information required to make LMI effective.

The context in which LMI is to be applied has also changed in another important respect. This is that while the value of index term weighting has long been known, weighting now has a more conspicuous role in retrieval systems. This is not just because continuing experiment has confirmed its value. The more important reason is that with full text much more occurrence data is available for weighting, and in particular within-document frequencies can vary much more than with abstract-length texts. However, when compared with the single-term cases, weighting for compound terms can be complicated. More generally, working with full text involves a larger range both of environment variables and system parameters, making testing LMI, in conjunction with other system devices, more challenging.

2.3 The TREC Programme

Fortunately, the Text REtrieval Conference (TREC) Programme underway since the early Nineties has provided an unrivalled opportunity to study indexing questions (TREC, 1993-1997). The Programme has involved very large files containing hundreds of thousands of documents, many requests (“topics”), extensive relevance data, and specified retrieval tasks and standard performance measures. More significantly, it has drawn in many participating teams applying very many different indexing and searching strategies: these range from using sophisticated LMI to using just simple single terms, and from searching with manually-constructed queries to using plain automated ones, along with many mixed approaches e.g. supplementing LMI with terms drawn from available manually-constructed thesauri.

In the present context TREC is especially important as a field for testing full-text retrieval using LMI, and as involving enough text to provide statistical data to guide or amplify LMI. The evaluation has covered both “ad hoc” searching and routing: I shall concentrate on the former here, as comparable with earlier work. Several teams, notably at University of Massachusetts at Amherst (UMass), Carnegie Mellon University and Claritech (CLARIT) and New York University and General Electric (NYU/GE), have applied LMI, in different ways, over several TREC cycles (Callan, Croft and Broglio, 1995; Evans and Lefferts, 1995; Strzalkowski, 1995). Taking their results together with those from other teams, it is possible to compare various approaches to LMI with one another, to compare LMI with NLI for compound terms, and to compare compound terms with simple single terms. In addition, though it is not possible to compare automatic NLP-based indexing with manual indexing for descriptions formed at file time, in the classical manner, it is possible to make some comparisons between automatic and manual search time indexing via query formation. However, while successive TREC cycles have allowed those engaged with NLP to refine and consolidate

their approaches, and have also made it possible to identify consistent trends and patterns in overall test performance, because all systems combine a range of indexing and searching devices it is impossible to attribute system performance levels very specifically to particular LMI strategies: only broad brush statements are possible,

Subject to this qualification, the particular questions on which the TREC evaluations may be able to throw light are as follows. For full text retrieval, what does LMI contribute to:

1. the determination of an indexing vocabulary, i.e. a chosen rather than simply accumulated vocabulary?
2. the provision of individual document descriptions?
3. the construction of individual request descriptions, i.e. queries?

These questions may be further specialised to refer to the use of compound terms, and hence as involving comparisons between LMI and NLI compounds and between compounds and simple terms. The questions are properly separate. Thus with respect to (1), with such a vocabulary consisting of (normalised) complex terms, when documents are indexed only those compounds in the list will be assigned, regardless of others occurring in the source texts. This represents the classical application of a vocabulary, except that we may expect the vocabulary itself to be ultimately derived from the file, by formation of candidate list and purging on simple frequency or to retain only shared subcompounds (e.g. pairs common to triples). With respect to (2), where there is no vocabulary a document is indexed by any compounds it contains subject e.g. to frequency criteria referring to the individual text. However the form of the term as assigned to a document may embody compound normalisation, or this may be achieved by query term modulation at search time.

Clearly, when indexing is done with compounds this applies to both documents and requests. However the effects are different when documents are indexed independently of requests, or only via the matching process when the document presence of query compounds is checked for documents. The difference is that where documents are not independently indexed, query expansion cannot add compounds, only single terms. This applies whether expansion is done as a prepass, or after a search iteration. It is also a logical rather than practical point as far as the distinction between file time and search time indexing is concerned. It would be practically feasible to process selected documents to find compounds on the fly at search time, though it is more likely to be done at file time. Building a vocabulary requires file processing, though it does not follow that individual descriptions have to be formed at file time. The practical issues are associated with the relative costs of fully indexing all documents once, offline, as opposed to doing online analysis, and with the implications of whether frequency data is too hard to compute term weights and hence document scores for ranking. Thus without weights it is not possible to decide which are the best documents to exploit for expansion.

Of the teams mentioned, GE/NYU and CLARIT both opt for document analysis, with the former engaging in more elaborate processing including extensive head-modifier normalisation for variant text expressions: the output is used both for primary indexing and as a basis for establishing file-based vocabulary relations. UMass represents a minimalist approach, with requests linguistically processed for compounds checked against document texts only through proximity. From the point of view of indexing with compound terms, these systems can be set specifically against the Cornell use of joined terms via a statistical phrase vocabulary

(Buckley, Allan and Salton, 1995). However when comparisons are made between the NLP-based systems, or against others using NLI even just with single terms, there is no manifest and consistent advantage evident for NLP-based LMI. It appears that system effectiveness is more determined by other factors, notably the use of weighting.

The most convenient view of the TREC results for present purposes is that given by the summary comparisons shown in Appendices to the TREC Proceedings (for TREC-3 - TREC-5). These show the status of the better performing teams over successive cycles, and thus give a more reliable picture of the relative merits of different approaches than a single evaluation. The comparisons focus on precision at a document rank cutoff of 30. They deliberately play down fine performance distinctions as these are not necessarily significant, while the use of the cutoff is appropriate for approaches to indexing that place considerable emphasis on precision. The main features of the comparison are that the figures show very similar performance from quite different indexing strategies, upper-bracket performance for NLI approaches, and fair consistency in relative performance over time (and hence with different data) for leading teams.

It is possible to interpret earlier TREC results in less damaging ways for LMI than may first appear, once the nature of the TREC requests (topics) is taken into account. These were initially long and very carefully formulated. In the first cycles the requests included so-called ‘concepts’ supplied by the user and often figuring as two-or three-word phrases. These on the one hand made NLP unnecessary, since the phrases were evident from the request format, and on the other supplied such rich information that even single-term approaches could be expected to work well with the carefully-selected set of words supplied, particularly since individual words might recur in different phrases, increasing their weighting. However in later cycles concepts were removed (as making life too easy), and in subsequent cycles requests were either further reduced (TREC-4) or made harder to satisfy (TREC-5). Even so, while in both cases performance levels dropped substantially, the general comparative picture remained as before.

2.3.1 Implications of Phase 2 results

It may nevertheless still be premature to conclude that compound terms, and especially ones given by LMI, could not improve performance at the margin: because TREC systems typically apply many devices in combination (as the UMass system is explicitly designed to do), it is on the one hand difficult to see what the precise effect of LMI is, and on the other as difficult to be certain that there is no productive way of combining LMI with other devices. Thus if we think of the effect of indexing and searching devices on the retrieval of relevant documents, we have various possibilities. Two devices may be alternative ways of reaching the same concept, and hence same individual document and hence document set. Or they may access different concepts, but in the same individual document and document set. Or they may access different concepts in different individual documents, and hence different document sets. Thus in retrieval devices may duplicate, reinforce or complement one another with respect to relevant (or to non-relevant) documents, but may also, as soon as different concepts are involved, cancel one another out in matching both relevant and non-relevant documents. It is therefore hard to determine what precise effects, and effects holding as a rule for files of documents and series of requests, LMI or any other form of indexing will have.

However it is impossible not to conclude, for ‘ordinary’ document/text retrieval, that where compound terms are concerned the statistical facts about term occurrences help as

much to make joined terms linguistically legitimate as NLP can (see e.g. Damerau, 1994), as well as helping to make them effective through weighting. Of course, some effort is required to keep a statistically-controlled compound term vocabulary reasonably up-to-date for a file, both for indexing given requests and to support query expansion. But this is much less effort than the LMI analogue. Moreover, if NLI is affected by poor starting requests, so is LMI: two-word requests do not provide much retrieval leverage, whatever the indexing strategy.

Yet in paradoxical fashion, improving requests by supplying more words may make simple-term indexing effective enough, via the *redundancy* that allows multiple term matches, to render actual compound terms unnecessary. Thus it would appear that the primary role of compound terms in indexing, and even for LMI if this can be shown to be superior to NLI, would be in conjunction with simple terms. This is quite different from the traditional approach to indexing with which I began. It has become customary when compounds are used, as in TREC systems, to include both compounds and their constituents in queries. This still, with weighting, naturally gives a preference to the usually more discriminating compounds; in particular, they will normally be less frequent in the file than their members. But including the member terms means that matching is also allowed for individual terms (even if their actual text occurrences are within phrases), which if their within-text frequency is high may be advantageous. Treating phrases, even normalised ones, as fixed and undecomposable units is an extremely stringent retrieval strategy, and the default strategy that automatically adds all the single terms into a description is much more flexible. Doing this has the advantage of promoting both precision, through multiple single term matches, and recall. Treating all indexing terms according to the same general principles in weighting appears to work well enough, though it could be that more refinement in differentiating weight levels for compound and simple terms may be useful.

3 Other roles for NLP

The conclusion that LMI is not needed for mainstream retrieval within the modern postcoordinate framework is practically convenient; and it is reinforced by such operational evidence as that provided by Silvester, Genuardi and Klingbiel (1994), where even the crude LMI applied in DDC, described earlier, has been abandoned in the successor NASA system. Document retrieval, even with full text, remains an intrinsically coarse task. Thus claims that linguistic analysis is needed for other indexing purposes than the provision of structured index descriptions on which I have concentrated have not yet been substantiated. For example, it may appear obvious that explicit word sense selection, for which linguistic analysis would be required, would improve retrieval performance. But Krovetz and Croft's (1991) investigations suggest this is far from obvious. It seems that the coordination effect, enhanced by redundancy in simple term indexing, can be quite sufficient for sense disambiguation, at least for single language databases, though whether explicit disambiguation is needed for cross-language searching with multi-lingual databases is an open question. Even where sense discrimination may add something to performance, Schütze and Pedersen (1995) show that this can be achieved by statistical rather than linguistic methods.

It is not clear, either, that NLP is required for some tasks that are closely related to ordinary retrieval. Thus with routing (filtering) as studied within the TREC Programme, the same general pattern of relative merit holds as for the ad hoc case, with the major additional contributor to performance the amount of training data available: this can indeed

be expected to be the dominant factor in performance. Similarly, Lewis's (1991) work on text categorisation, i.e. automatic assignment to prior headings, suggested that LMI, i.e. using complex terms, had no extra beneficial effect, in line with Hayes' experience (Hayes, 1992).

Thus the real contribution that NLP may make to *selective* content-based information management has to be sought elsewhere. One potential line is in displaying information about whole documents to the user, where giving selected phrases may be more informative about document content than highlighting matching terms, or simply listing key individual words. This has been illustrated for CLARIT, for example, in system demonstrations. By a natural extension this kind of processing leads on to information extraction and summarising. For some purposes it may be feasible and useful to extract source text sentences or passages (just using statistical, text-locational, or "cue-word" criteria). But it is clear that selecting key information and using it to form an information base, for future question-answering, in general depends on linguistic analysis, even if this may sometimes be done by linguistically-shallow means. This is well illustrated by the line of work from Sager (1978) through e.g. Jacobs and Rau (1988) to the Message Understanding Conference applications (cf. MUC-6, 1996). As the effort embodied in the TIPSTER Programme (TIPSTER, 1996) implies, the real role of NLP must be in supporting such more exigent information-management functions within a larger, multi-functional whole.

4 References

Bely, N., Borillo, A., Virbel, J. and Siot-Decauville, N. (1970) *Procédures d'analyse sémantique appliquées à la documentation scientifique*. Paris: Gauthier-Villars.

Biebricher, B. et al. (1988) The automatic indexing system AIR/PHYS - from research to application. *Proceedings of the 11th International Conference on Research and Development in Information Retrieval (ACM SIGIR)*, 333-342.

Buckley, C., Allan, J. and Salton, G. (1995) Automatic routing and retrieval using SMART: TREC-2. *Information Processing and Management*. 31 (3), 315-326.

Callan, J., Croft, W.B. and Broglio, J. (1995) TREC and TIPSTER Experiments with INQUERY. *Information Processing and Management*, 31 (3):385-395.

Chan, L.M., Richmond, P.A. and Svenonius, E. (Eds.) (1985) *Theory of subject analysis: a sourcebook*. Littleton, CO: Libraries Unlimited.

Cleverdon, C.W. (1967) The Cranfield tests on index language devices. *Aslib Proceedings*, 19, 1967, 173-192.

Cleverdon, C.W. (1977) *A computer evaluation of searching by controlled language and natural language in an experimental NASA data base*. Report ESA 1/432, European Space Agency, Frascati, Italy.

Croft, W.B., Turtle, H.R. and Lewis, D.D. (1991) The use of phrases and structured queries in information retrieval. *SIGIR 91, Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*. 32-45.

Damerau, F.J. (1993) Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, 29(4), 433-447.

Dillon, M. and Gray, A.S. (1983) Fully automatic syntax-based indexing. *Journal of the American Society for Information Science*, 34 (2), 99-108.

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

Fagan, J.L. (1987) *Experiments in automatic phrase indexing for document retrieval: a comparison on syntactic and non-syntactic methods*. PhD Thesis, Department of Computer Science, Cornell University; TR 87-868.

Fagan, J.L. (1989) The effectiveness of a non-syntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40 (2), 115-132.

Hayes, P.J. (1992) Intelligent, high-volume text processing using shallow, domain-specific techniques. In *Text-based intelligent systems*, Ed. P.S. Jacobs, Hillsdale NJ: Lawrence Erlbaum Associates, 227-241.

Hahn, U. (1990) Topic parsing: accounting for text macro structures in full-text analysis. *Information Processing and Management*, 26, 135-170.

Harman, D. (1991) How effective is suffixing? *Journal of the American Society for Information Science*, 42 (1), 7-15.

Hillman, D.J. (1968) Negotiation of inquiries in an online retrieval system. *Information Storage and Retrieval*, 4, 219-238.

Hull, D.M. (1990) Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47 (1), 70-84.

Hutchins, W.J. (1975) *Languages of indexing and classification*. Stevenage, Herts: Peter Peregrinus.

Jacobs, P.S. and Rau, L.F. (1988) Natural language techniques for intelligent information retrieval. *Proceedings of the 11th International Conference on Research and Development in Information Retrieval (ACM SIGIR)*, 85-99.

Klingbiel, P.H. (1973) A technique for machine-aided indexing. *Information Storage and Retrieval*, 9 (9), 477-494.

Klingbiel, P.H. and Rinker, C.C. (1976) Evaluation of machine-aided indexing. *Information Processing and Management*, 12 (6), 351-366.

Krovetz, R. and Croft, W.B. (1992) Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10 (2), 115-141.

Lancaster, F.W. (1972) *Vocabulary control for information retrieval*. Washington, DC: Information Resources Press.

Lewis, D.D. (1991) *Representation and learning in information retrieval*. PhD Thesis, Department of Computer and Information Science, University of Massachusetts at Amherst, TR 91-93.

Mauldin, M. (1991) Retrieval performance in FERRET: a conceptual information retrieval system. *SIGIR 91, Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*. 347-355.

MUC-6 (1996) *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Francisco, CA: Morgan Kaufmann,

Porter, M.F. (1980) An algorithm for suffix stripping. *Program*, 14, 130-137.

Salton, G. (1968) *Automatic information organisation and retrieval*. New York: McGraw-Hill.

Salton, G. (1972) A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science*, 23 (2), 75-84.

Salton, G. and McGill, M.J. (1983) *Introduction to modern information retrieval*. New York: McGraw-Hill.

Schütze, H. and Pedersen, J.O. (1995) Information retrieval based on word senses. *Fourth Annual Symposium on Document Analysis and Information Retrieval*, Information Science Research Institute, University of Nevada, Las Vegas, 161-175.

Silvester, J.P., Genuardi, M.T. and Klingbiel, P.H. (1994) Machine-aided indexing at NASA. *Information Processing and Management*, 30 (5), 631-645.

Smeaton, A.F. and van Rijsbergen, C.J. (1988) Experiments in incorporating syntactic processing of user queries into a document retrieval strategy. *Proceedings of the 11th International Conference on Research and Development in Information Retrieval (ACM SIGIR)*, 32-51.

Sparck Jones, K. and Tait, J.I. (1984) Automatic search term variant generation. *Journal of Documentation*, 40, 50-66.

Strzalkowski, T. (1994) Robust text processing in automated information retrieval. *Proceedings of the 4th Conference on Applied Natural Language Processing (Stuttgart)*, Association for Computational Linguistics, 168-173.

Strzalkowski, T. (1995) Natural language information retrieval. *Information Processing and Management*, 31 (3), 397-417.

TIPSTER (1996) *Tipster Text Program, Phase II. Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996*. San Francisco, CA: Morgan Kaufmann.

TREC (1993-1997) *Proceedings of the First Text REtrieval Conference (TREC-1)*. Ed. D.K. Harman, Special Publication 500-207, National Institute of Standards and Technology, Gaithersburg, MD, 1993; *Second (TREC-2)*, 500-215, 1994; *Third (TREC-3)*, 500-225, 1995; *Fourth (TREC-4)*, 500-236, 1996; *Fifth (TREC-5)*, 1997.