Bibliometric mapping as a science policy and research management tool

Noyons, E.C.M.

11 Towards automated field keyword identification

11.1 Introduction

In the SIB project, experts have been consulted to make a proper selection of publications to cover the field. Moreover, they could give their comments to the results during the study and afterwards (see previous section). More than once, they have expressed their concern about the selection of keywords and (thus) about the coverage of research in the maps. Not surprisingly, the choice of keywords to create both the overview map and the subdomain maps is of vital importance. Expert consultation during the process of creating the field structure revealed the difficulty in making a proper selection. Simply presenting a list of 'candidate keywords' does not work. Even though we were dealing with bibliometricians, contradictory input was given[24]. In some cases this may be due to ambiguity of candidate keywords.

In this section, a procedure for pre-selection of candidate keywords is proposed. This pre-selection aims at providing the expert with information to help him to make a more well-grounded decision to select or reject a candidate. The setup for such a procedure is based on three principles:

1. The structure (clusters of words or terms) to be generated, should be recognized by researchers in the field or other users of the maps (Section 3.2);
2. The interference of the creator of the map should be limited to a minimum or, ideally, be zero (the objectivity principle);
3. The words used to create a structure should be extractable from any 'standard' bibliographic database (Chapter 2).

The first principle covers the utility of the maps. When the map user recognizes the generated map as a reasonable representation of the field, not only his or her willingness to co-operate will be higher, also the interpretability of the maps will be benefited. As pointed out in Chapter 2, a structure is particularly useful when changes of the structure (in time) can be studied. However, the changes of a structure can only be interpreted if at least *one* situation in the evolution (one picture in the film) is recognized.

The second principle is mainly pragmatic in nature. Complete dependence on field experts is unwanted. It is difficult to find experts who are objective and willing to evaluate the maps and additional information (Chapter 3). Partly because they lack

[24] The keyword *mapping* seems a relevant candidate. In 1998 however, a report from the Welcome Trust was published entitled: 'Mapping the Landscape' in which not even *one* science map is found. The word *mapping* refers, apparently, to a much broader activity than *science mapping* in the 'cartographic' sense, although in bibliometric research it refers indeed in most cases to *cartography*.

time, and partly because they are not acquainted with the method. Moreover, reasonable objectivity is assured if the maps are generated using as little as possible 'human interference'. This objectivity is an important reason for the success of bibliometric methods as an evaluative tool.

The third principle refers to the applicability of bibliometric mapping. As pointed out in Chapter 2, the structure depends partly on the bibliographic database chosen to map the field. In order to be independent from the database producer, the content descriptive elements (CDEs) should be extractable from any 'standard' bibliographic database. A database producer often adds elements to a bibliographic item in order to improve the retrievability (classification codes; indexed terms). These elements improve the recall and precision of a user's search. However, in most cases they are database-specific, so that a map based on these elements, reveals a database-specific structure, rather than a structure based on the research output itself. Moreover, the usage of these database-specific elements preclude the creation of a map based on documents from more than one database. If a study requires that a research field can only be mapped on the basis of publications from several databases, this may become a serious drawback. Particularly in the case of database-specific classification, concordance with other systems can be very problematic.

11.2 From CDE to field keyword (FKW)

Under the assumption that we wish to create a science map based on keywords, there are several bibliographic fields that can be used to extract appropriate FKWs from the publication database representing the field under study. The most important ones are:

- controlled terms (indexed);
- (names of) classification codes;
- titles;
- abstracts.

Consideration of the third principle mentioned in the foregoing section, puts forward *titles* and *abstracts* as the most important elements to be used. Both aim at disclosing the contents of an article, and they are available in most bibliographic databases. Moreover, as they are usually drawn up by the authors themselves, a bibliometric mapping analysis based on titles and abstracts, sticks as close as possible to the original data. A drawback of using these 'free' text elements is the lack of a standardized style and the jargon used by the authors in a field. The first principle suggests that controlled vocabulary and indexed terms are more appropriate CDEs to create science maps. Field experts are more likely to recognize keywords from an indexed list, being generated by other experts. Free terms may not be recognized due to an alternative word choice by authors. Also, the second principle appears to be in

favor of the indexed terms. The lack of standardized jargon in any field affirms the need for a controlled vocabulary. It should be noted however that expert interference has already taken place on indexed terms. This is one of the main reasons why indexed terms are not to be preferred, in particular where studies based on multiple bibliographic databases are concerned.

As the use of a 'controlled' CDE encounters principle objections, and the use of a 'free text' CDE only makes the selection procedure more complicated, we explore the feasibility of using the latter to select the relevant FKWs to create the maps. The issue concerning the recognition of the controlled terms as opposed to free text terms is merely challenging towards the use of free text. The arguments against the usage of controlled terms are more fundamental by nature. An *indexed* (controlled) term is per se database-dependent and therefore more subjective. A bibliometric mapping study based on *free* text terms may be conducted with any publication corpus provided by researchers *themselves* in a specific field, as long as titles and abstracts are available.

The selection of keywords to describe the main contents of a publication document starts with *title* and *abstract*. They are elements describing the publication but are too specific[25] for describing the core activities in the science field to which the publication belongs. The 'candidate keywords' should therefore be meaningful parts *from* titles and abstracts. The smallest independently meaningful element in a title or abstract is one single *word*. The most common method used in the early years of co-word analysis based on 'free text', and still used in the present, can be described as follows. From publication titles and abstracts all individual words are identified[26]. Highly frequent, redundant words like *the*, *and*, *can* etc. are removed by using a 'stop word list'. Subsequently, the list of most frequent words is cleaned by removing further 'non-specific' words, such as *case*, *study* etc. The list of remaining words is input for co-word analysis.

To this method two important objections can be raised. First, the usage of 'stop word lists' and lists of 'too general words' requires the input of a field expert. This expert is prompted with relatively much (in his view) redundant information. Expert input should be reduced to a minimum and focused at the most relevant issues. Second, single words cause too much ambiguity. For this reason a word may be excluded from the list of candidates. For example, in the SIB project the word *performance* is a relevant and even central topic. Experts acknowledge this, but for various reasons. For researchers in information retrieval (IR) this word refers to the performance of the *computer* or *software* (speed and quality of results). But in science studies, it refers to the performance of *scientists* (scientific production and impact). A co-word clustering analysis of a set of core terms including *performance*, will probably put these different kinds of research topics together into an 'artificial' cluster.

[25] A publication (including title and abstract) is a unique contribution to the scientific dispute.
[26] Common word boundaries are spaces, hyphens, comma's and dots.

The above observations point out that a co-word analysis should at least allow *phrases* (word combinations, e.g., *research performance* and *system performance*) to join in. As a result a candidate keyword used to generate a science field structure and its representation by a map may be a word or phrase. We consider therefore as the smallest independently meaningful element in a title or abstract, a *word* or a *phrase*. But as long as a phrase is no more than a group of words (within a sentence), it should be determined what kind of phrases *should* and what *should not* be included. On the one hand, identifying any group of adjacent words (see Zamir and Etzioni, 1998) may cause severe interpretation or processing problems if the number of elements within a phrase is only limited by the number of words within a sentence. In that case any sequence of n words (n>1) within a sentence is a possible phrase. On the other hand, limiting the number of elements within a phrase to an absolute number may lead to interpretation problems (c.f., *journal impact factor*, would lead to *journal*, and *impact factor* or *journal impact* and *factor* if the maximum words in a phrase would be 2).

Thus far, two main issues involved in selecting the proper keywords to generate a science map were identified: (1) bibliometric distribution (number of occurrences in publications), and (2) semantic scope (exclusion of non-specific words, and specificity of phrases as opposed to single words). At this point a third characteristic of words and phrases is introduced: syntactic, or in a broader sense, linguistic properties (for instance, lexical category of words). This is particularly useful where the identification of phrases is concerned.

In the remainder of this chapter, the setup of the procedure to select 'appropriate' field keywords (FKWs) is discussed using these three characteristics of words and phrases:

- linguistic properties;
- semantic scope; and
- bibliometric distribution.

On the basis of the SIB study data, a procedure is proposed. Next, discussions with field experts in the SIB project and with a field expert in a project for the European Commission on neuroscience are used to validate the procedure in its basic elements. Rather than presenting a completed study, this section outlines a direction for future research.

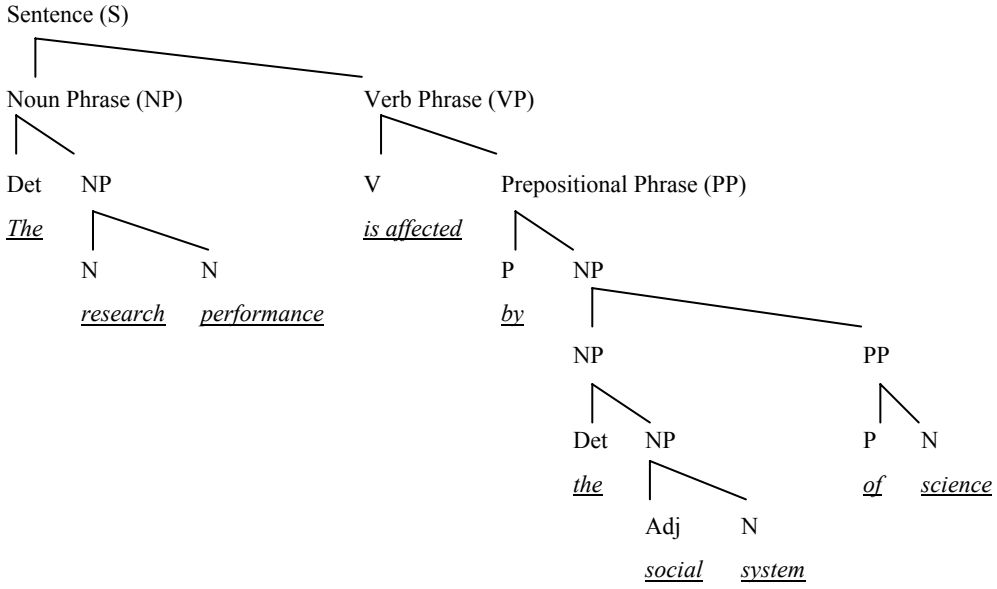
11.3 Linguistic characteristics

The linguistic analysis of titles and abstracts precludes the need for extensive lists of stop words (i.e., words that are to be excluded). These lists contain words like *the*, *and*, and *very*. These unwanted words are more easily detected by determining their lexical category (determiners, modifiers). The morphological and lexical part of a linguistic analysis can easily mark such words.

Moreover, a lot of words in a sentence can be ruled out because of their syntactic characteristics. Adjectives *as such*, for instance, very rarely contribute to the main issues in an abstract of an article. The same holds true for verbs. Consider the following sentence.

> "The research performance is affected by the social system of science"

The syntactic structure of this sentence may be described as follows:



As discussed above, an adjective like *social* as such is evidently not a favorable candidate to describe the article in which this sentence appears. But in the context *social system of science*, the relevance of the word becomes immediately obvious. The syntactic structure identifies the noun phrases *social system* and *social system of science* as coherent, and thus meaningful word-combinations, or phrases. Therefore, these phrases *are* relevant candidates. On top of that, this syntactic structure provides an excellent starting-point to identify phrases within sentences. By using the syntactic structure, we reduce the number of possible phrases within a sentence significantly. Moreover, phrases will more easily be interpreted. The smallest meaningful elements within a sentence should therefore be words and *syntactic* phrases.

Finally, it has been argued that *nouns* are particularly interesting for describing the main contents of an article. In an ESPRIT project in the early nineties (Karlsson, 1990; Karlsson et al., 1995; Voutilainen, 1993), a software tool was developed to extract noun phrases from English sentences. This tool was developed to be used for automated indexing for information retrieval. It has been argued that the noun phrase (NP) plays an important role to identify the main issues of an article. This is supported

by the finding that between 80 and 95 of the terms listed in thesauri, or indexed lists of bibliographic databases, are nouns or noun phrases (Arppe, 1995). For our purposes we use a slightly adjusted version of the application developed in the ESPRIT project (NPtool). This NP extracting tool has an excellent performance (Bennet et al., 1997). It is grammar rule-based rather than being lexically-based. The flow chart below describes the process from text to NPs.

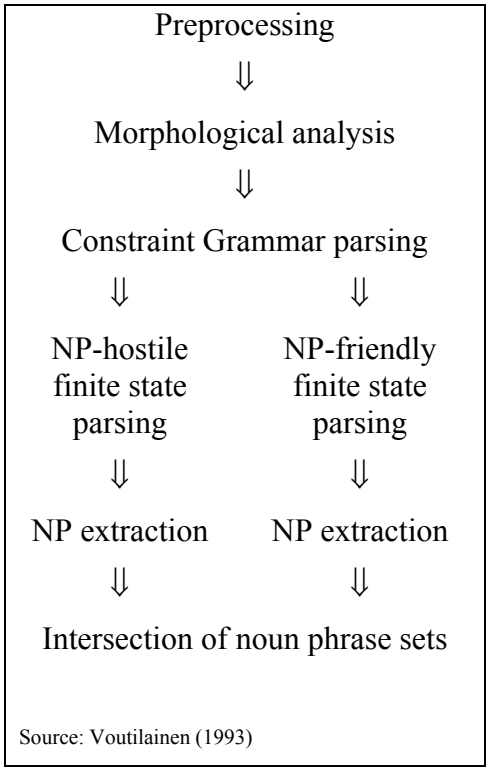


Figure 11-1 NPtool system flowchart

The results of an analysis by NPtool on two selected titles plus abstracts, given in Table 11–1 and Table 11–2. In the texts, noun phrases (NP's) identified by NPtool are underlined.

Table 11–1 Example of a parsed title and abstract of an article (sample 1)

An experiment in science mapping for research planning

This paper considers the recent attempt of the UK Advisory Board for Research Councils to test the usefulness of various citation, co-citation and co-word bibliographic analysis techniques for evaluating the state of various scientific disciplines, including potential areas for useful investment; and in general as an aid to research planning by science policy-makers in a period of steady (or even relatively declining) resources. Results of a study involving the examination of five important scientific fields are considered, and discussion focuses on the advantages and disadvantages of each technique.

(text from: Healey, Rothman, and Hoch, 1986)

Table 11–2 Example of a parsed title and abstract of an article (sample 2)

Where is science going?

Do researchers produce scientific and technical knowledge differently than they did ten years ago? What will scientific research look like ten years from now? Addressing such questions means looking at science from a dynamic system perspective. Two recent books about the social system of science, by Ziman and by Gibbons, Limoges, Nowotny, Schwartzman, Scott, and Trow, accept this challenge and argue that the research enterprise is changing. This article uses bibliometric data to examine the extent and nature of changes identified by these authors, taking as an example British research. We use their theoretical frameworks to investigate five characteristics of research said to be increasingly persuasive – namely, application, interdisciplinarity, networking, internationalization, and concentration of resources. Results indicate that research may be becoming more interdisciplinary and that research is increasingly conducted more in networks, both domestic and international; but the data are more ambiguous regarding application and concentration.

(text from: Hicks and Katz, 1996)

Table 11–3 Comparison of sample 1 and sample 2

Indicator	*Sample 1*	*Sample 2*
Number of sentences	3	6
Number of words in document	96	152
Number of identified NPs	34	41
NP density (ratio NP to total number of words)	0.35	0.27
Number of words covered by NPs	48	57
Average number of words per NP	1.41	1.39
Number of words not covered by NPs	47 (49)	92 (61)

The results in terms of NP's identified by the automated parser are not perfect. For instance, in the last sentence of Table 11–2, the NP *ambiguous regarding application* has been identified. On semantic grounds, only *application* should be identified. On syntactical grounds, however, the NP is correct (c.f., *serious looking woman*). In spite of such 'failures', we use the identified NP's without correcting them. The above described inaccuracy is rare and does not affect the overall quality of the tool. In Table 11–3, some simple statistics of the parsed titles and abstracts are presented.

Evidently, the style of these two abstracts is very different. Sample 1 is rather technical and short, whereas sample 2 is longer and more comprehensible. Apart from such common (quantitative) style characteristics (number of words per sentence or use of passive voice), the difference seems to be indicated by (other) statistics in Table 11–3. With the average number of words within an NP being equal (around 1.4), the number of NPs per document differs (0.35 in sample 1 vs. 0.27 in sample 2). As a result, the ratio of non-NP words to the total number of words in sample 2 is higher (0.61 in sample 2 vs. 0.49 in sample 1).

In spite of the different abstracting styles, the NP extraction tool identifies for both samples a reasonable number of relevant NP's which can be used to describe the contents of each article. Moreover, the words not covered by the NP tool are equally non-informative, and would normally appear in a 'stop word' list, containing words to be removed from the list of most frequently used words by authors in their titles and abstracts (see section 11.1). Seemingly, the NP extraction tool is capable of providing a set of publication keyword (PKW) candidates (Table 11–4).

Table 11–4 Most relevant publication keywords from sample 1 and sample 2

Sample 1	*Sample 2*
Bibliographic analysis technique	application
co-citation	bibliometric data
co-word	British research
mapping	concentration
research planning	dynamic system perspective
science policy-maker	interdisciplinarity
UK	internationalization
Useful investment	networking
Various scientific discipline	research enterprise
	scientific and technical knowledge
	scientific research
	social system
	theoretical framework

In this case, with just two abstracts, it was possible to select these keywords by reading the titles and abstracts and than checking the list of candidates 'manually'. But

obviously, in bibliometric studies involving many thousands or sometimes even hundreds of thousands of publications, this is undoable. Therefore, a powerful and fast algorithm is absolutely needed to select these words automatically. To meet this strict requirement, we need more information about the semantic context of the candidates.

Obviously, the opinion of a field expert is crucial for this issue. In the case of SIB – our own field of research - we were able to make the selections ourselves. But generally, we depend on an 'external' field expert. As pointed out in Chapter 3, the way in which the data are presented to the expert will influence the quality of the results. If an expert is prompted with every possible keyword and is kindly asked to give his opinion, you may end up with no input at all. If there has been a considerable pre-selection, results will be more to the point. In the next section, a new procedure will be proposed in which this pre-selection reduces the work for the expert significantly. As these pre-selections are based on objective criteria, and as the expert will be informed about them, the discussion will be focused on the relevant issues.

11.4 Semantic scope

The primary **meaning** of a word or phrase has at least three levels:

A. Meaning as such (in a dictionary: more or less general);

B. Meaning within a field (specific, 'field jargon');

C. Meaning within the article (contextual, sometimes even metaphoric).

In the process of selecting FKWs, all these levels must be taken into consideration. The following example will illustrate these levels and their contribution to the selection procedure.

Table 11–5 Meaning overview of a noun phrase sample

Candidate keyword: Impact

Context level		*Meaning*	
A	Dictionary	*Noun*:	hit, influence
		Verb:	implant, hit
B	Noun within field (SIB)	1.	Citations per publication
		2.	Scientific influence
C	Noun (citation per publication) within article (function within discussion)	*Result*:	(…) that the impact of institute X is below the impact of Y (…)
		Method:	(…) using the ISI impact factor (…)
		Object:	(…) creating a new impact factor (…)
		Neutral:	(…) that this does not have any impact on the usage of (…)

First, the word *impact* appears in two lexical categories: noun and verb. The meaning in both is related. As a noun, it has a both a *physical* and a *mental* connotation (A level). In the scientific field of SIB and related fields, *impact* has a general meaning *and* a specific meaning (B level). The impact of publications is measured by the number of times it is cited. The bottom line is: the more often an article (or person) is cited, the more impact it has (on other researchers). Moreover, the function within the article (C level) can be of importance when examining candidate FKWs to generate a field structure.

In the above configuration, we identify three different context levels in which the candidate has its 'meaning'.

- The A level, which is the 'whole world' or, at least, the world of science;
- the B level, which is the 'world' of the science field (for instance SIB);
- the C level, which is the very limited 'world' of the article, so that 'meaning' should rather be described as 'function'.

In view of the objective of this chapter, the B level plays a crucial role, as we are trying to identify *field keywords* rather than *publication keywords* (PKW, c.f., Chapter 2) or *science keywords*. If a word primarily has its meaning on this B level, it will be a relevant candidate to be used to structure the field. If candidate primarily has its

meaning on the A level, it is most probably removed from the list on the basis of being non-field specific. In the samples (Table 11–1 and Table 11–2), words such as *article, author* and *data* are examples of such non-specific words. The meaning the C level seems too specific in order to be used for the objective. Is the 'meaning' or 'function' of *impact* within the article of any relevance to structure the field of SIB? In particular cases, where the structure has a certain purpose, this may be the case. Leydesdorff (1997) pointed out that particular words may appear in the introduction, method, results, or conclusions section. He argues that this particular phenomenon proves that words should not be used to map (the dynamics of) science. In his reply, Courtial (1998) argues that this wandering in particular is in favor of co-word analysis as a tool to visualize the dynamics of science. The shift of a topic from one section to another will cause changes in the word co-occurrence data. These changes reflect the developments in the particular science field. As long as the data are robust enough, the results will not suffer from inadequacies due to sociolectic and jargon matters. So, in studies where mapping concerns the visualization of the field dynamics, this 'function' issue does not seem to create a problem, because the aim is to generate a map on the B level, not on the C level.

Still, there is an important aspect about the C level. In relation to this aspect, Leydesdorff (1989) did a significant observation. It concerned the applicability of title words on the one hand, and abstract words on the other, for co-word maps. Leydesdorff (1989) reached the conclusion from a biochemistry publication corpus, that on average abstract words are *less specific* than title words. That is, the selection of most frequent abstract words was less specific than the selection of most frequent title words. But Peters and Van Raan (1993a) reach the opposite conclusion, comparing the title co-word map with the abstract co-word map.

> Comparing the maps, we see that in many places the same words are visible (…) Furthermore, title words cover more general words than abstract-words, such as "chemical", "experiment-", "measurem-", or "using". Thus, title-words appear to be somewhat more general in scope, but differences are not as clear-cut as found in the study of Whittaker (1989).
>
> (Peters and Van Raan 1993a, p. 31)

In their study, however, Peters and Van Raan used the uncontrolled terms of the Chemical Abstracts database. These uncontrolled terms are keywords directly extracted from abstracts and not unified or indexed. In that sense, they are similar to the keywords that Whittaker (1989) compared to title words. In both studies *not all* abstract words have been considered, but both seem do have used *all* title words. Peters and Van Raan found the word "using" as one of the most frequent title words but not as a (abstract) keyword. The reason for this is, of course, that this word already has been filtered out by the database producer as being too general to be a (abstract) keyword. Leydesdorff (1989) filtered out the same (stop) words from titles

and abstracts and *then* compared the specificity of the most frequent ones. In that sense, it is understandable that Peters and Van Raan find that title words are more general, whereas Leydesdorff (1989) found the opposite.

The title is both an advertisement-board of the article and, in most cases, the shortest possible descriptor of its contents (Whittaker, 1989). These two objectives of a publication title, require that it refers to the context either on the C level (description) or on the B level. Generally speaking, the audience addressed in scientific journals is from the B context and should be able to read the main issues addressed in the article from the title. Together with the descriptive requirement, this accounts for the specific character of title NP's. The fact that the most frequent NP's in abstracts are less specific than title NP's is primarily due to the requirement that the readability (as the knowledge of language goes beyond the B context, it refers to the A context) of abstracts should be higher than the readability of titles[27]. In the project SIB, reported in section 10 of this chapter, we generated some general statistics concerning the NP usage in titles and abstracts (Table 11–6).

Table 11–6 General statistics of NP's in SIB abstracts (1992-1997)

Indicator	*Titles*	*Abstracts*
Total number of publications	2,503	2,503
Average number of NP's	3.95	29.66
Average number of words	10.42	113.61
Average NP density	0.39	0.27
Average number of words per NP	1.87	1.68

These figures show that, on average, four NP's are included in titles and almost thirty in abstracts. The average NP density (ratio NP to the total number of words) of titles is significantly higher (0.39 vs. 0.27). The latter observation supports the finding in Leydesdorff (1989). The specificity of title words is much higher because, on average, less non-informative words are involved. Furthermore, these overall statistics in SIB show that the NP density of the abstract in Table 11–1 (sample 1: 0.35) almost reaches a title average of 0.39. The absence of 'non-informative' words in this abstracts does not improve its readability.

[27] In most cases the title is not a sentence but rather an elliptic phrase 'mapping the field of SIB' rather than 'In this article we will give a report of the study concerning the mapping of the field of SIB'.

This observation indicates that the title seems more appropriate to extract the most specific keywords from the candidates than the abstracts. The second sample, however, shows that the title is useless for that matter. The latter title is primarily used to attract readers, rather than to cover the main issues addressed (see Whittaker, 1989 and Peters and Van Raan, 1993a). A selection of most specific keywords per document (PKW) exclusively based on their occurrence in title (or abstract) does not seem reasonable.

11.5 Bibliometric distribution

An overview structure of a field based on keyword co-occurrence, requires that the most central keywords are used. Bibliometrically speaking this means that the most frequent keywords are used. In previous sections, it has been argued that the selection of nouns (NPs) may be a good pre-selection. Moreover, it has been argued that on average the NP in a title yields more specific information about the contents of an article than the abstract. As a result, we argue that most frequently used NP's in abstracts tend to be less specific. All kinds of 'common' methods, data and tools are being discussed in abstracts. These methods and tools are of great importance for the development in a research field, but do not contribute per se to the identification of core topics and areas in a field. Moreover, it has been argued previously that the readability requirement plays a role in the specificity of the average abstract NP. Thus, we conclude that the distribution of abstract NPs may be of importance to identify PKWs, but it seems not to contribute directly to the identification of FKWs. For the latter the distribution on the B level context is significant. title NPs seem to be more appropriate at this level. Moreover, at this level of aggregation the 'advertisement' titles (c.f., "where is science going?") will not interfere because of their low frequency. For these reasons, we select high frequent NP's for the overview structure from titles only. Then, to create the structure based on the selected FKWs, we use their co-occurrences in both titles *and* abstracts.

11.6 Combining the three aspects

In view of the objective in this chapter, the proposed selection procedure requires the following information about each candidate:

- lexical category;
- syntactic structure;
- distribution in titles within field;
- distribution in abstracts within field;
- distribution within field as opposed to distribution in whole of science.

The lexical category and syntactic structure is needed to select NPs only. The former is to identify the NPs, the latter has appeared recently to be useful regarding the semantic scope of the NPs. After having created a list of relevant candidates for the SIB study and having created a list of NPs to be excluded, we discovered a certain pattern. There appeared to be a clear correlation between complexity of the NP structure (in terms of number of words included) and the likeliness to be selected as being specific enough. The list of NPs to be excluded contained almost exclusively single word NPs (SWNP), whereas the list of selected NPs almost exclusively contained multiple word NPs (MWNP). In a presently performed study for the EC in the field of neuroscience, we encountered a similar finding. On the basis of our findings in SIB, we created a list of MWNPs on the one hand as most likely candidates, and a list of SWNPs with candidates to be excluded on the other. A field expert evaluated both lists. At first, he was rather skeptic about the pre-selection. After having evaluated the two lists, he noted about 10 MWNPs which had better be removed from the list of keyword candidates. Moreover, he selected only 40 words from a list of 500 SWNPs which, in his opinion, should be included in the list of keywords. In some of the latter cases, it concerned SWNPs with a 'complex' morphological structure (e.g., *immunoreactivity*). In these cases, the SWNP is composed of two 'virtually independent' words. As a result, we should consider to regard the morphosyntactic characteristics of an SWNP as well[28].

Two additional findings support the proposed pre-selection. For the NPs identified by NPtool in the samples of Table 11–1 and Table 11–2, we searched in the Social Science Citation Index (SSCI), and within a subset of documents included in the journal set of the SIB study. The SSCI we take as a representation of the 'whole of science'[29] (the context on A level of Table 11–5), the journal subset as a representation of the field (B level). For each identified NP from sample 1 and 2, the ratio of occurrences in B to occurrences in A was calculated ('specificity index'). Furthermore, the NPs were categorized on the basis of their syntactic structure and selection by experts, and for each category an average is calculated (Table 11–7).

[28] Note that the morphological difference between 'immunoreactivity' and 'science policy' does not exist in, for instance, Dutch ('immunoreactiviteit' vs. 'wetenschapsbeleid'). The linguistic treatment of word compounds plays an important role in this discussion and will be taken into consideration in future research.

[29] Of course, it would be more accurate to take the whole set of ISI products to represent the whole science output if not all available scientific output databases. For my point here, the SSCI seemed enough.

Table 11–7 Specificity Indicators for NPs in two SIB samples (Table 11–1 and Table 11–2)

NP type	*Specificity Average*
All NPs in sample 1 and 2	0.06
Single Word NPs	0.02
Multiple Word NPs	0.35
Expert-selected NPs	0.22
Expert-excluded NPs	0.02

The results show that the specificity of both multiple word NPs as opposed to single word NPs, and expert-selected NPs as opposed to non-selected NPs is significantly higher. On the one hand this may indicate the syntactic structure to be a useful source of information to make a pre-selection of candidate NPs. On the other hand, it suggests that the distribution of any candidate NP within the field as opposed to its distribution within the whole of science may indicate its relevance to the field, in terms of representing the core activity of a science field.

Furthermore, we calculated the specificity index for 120 SWNPs from the SIB study. It showed an average value of 0.03, whereas the specificity of MWNPs in SIB revealed 0.21 on average.

Egghe (1999) shows that the rank-frequency distribution of multi word phrases (not *noun* phrases only) is significantly different from the distribution of single word phrases. Referring to Smith and Devine (1985), he indicates that the exponent β of multiword phrases decreases with an increasing number of elements within the phrase. A similar finding is in the graph of Figure 11-2, where the distribution of SWNPs and MWNPs in SIB is shown. It shows a slope of SWNPs that is much steeper than the slope of MWNPs. The head of their distribution (up to rank 100), however, is similar. This phenomenon has to be studied in more detail.

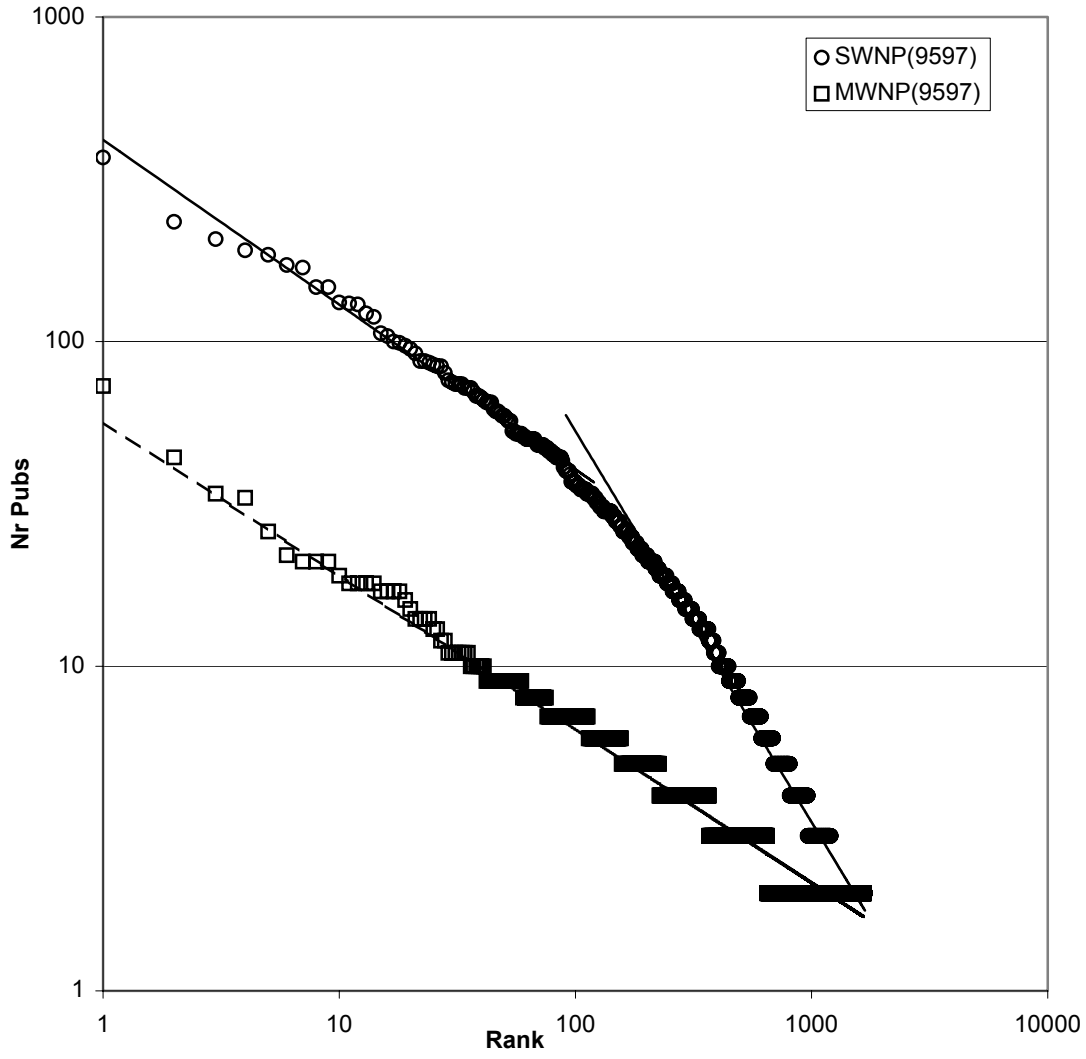


Figure 11-2 Zipf distribution of single word NPs and multiple word NPs in SIB (1995-1997)

On the basis of the above observations a pre-selection of NPs on the basis of their syntactic properties seems justified. Furthermore, in the neuroscience project, we experienced a smooth validation by making certain pre-selections and by providing the expert with well-structured and well-documented information about the candidate keywords. In particular the division of pre-selected keywords (within a generated cluster structure) on the one hand, and the initially rejected NPs on the other, speeded up the procedure considerably. Thus, well-aimed questions could be asked to the expert with regard to the required input. Moreover, by structuring the data, the expert

had a kind of preview into the results. A summary of the proposed selection procedure is sketched in Figure 11-3.

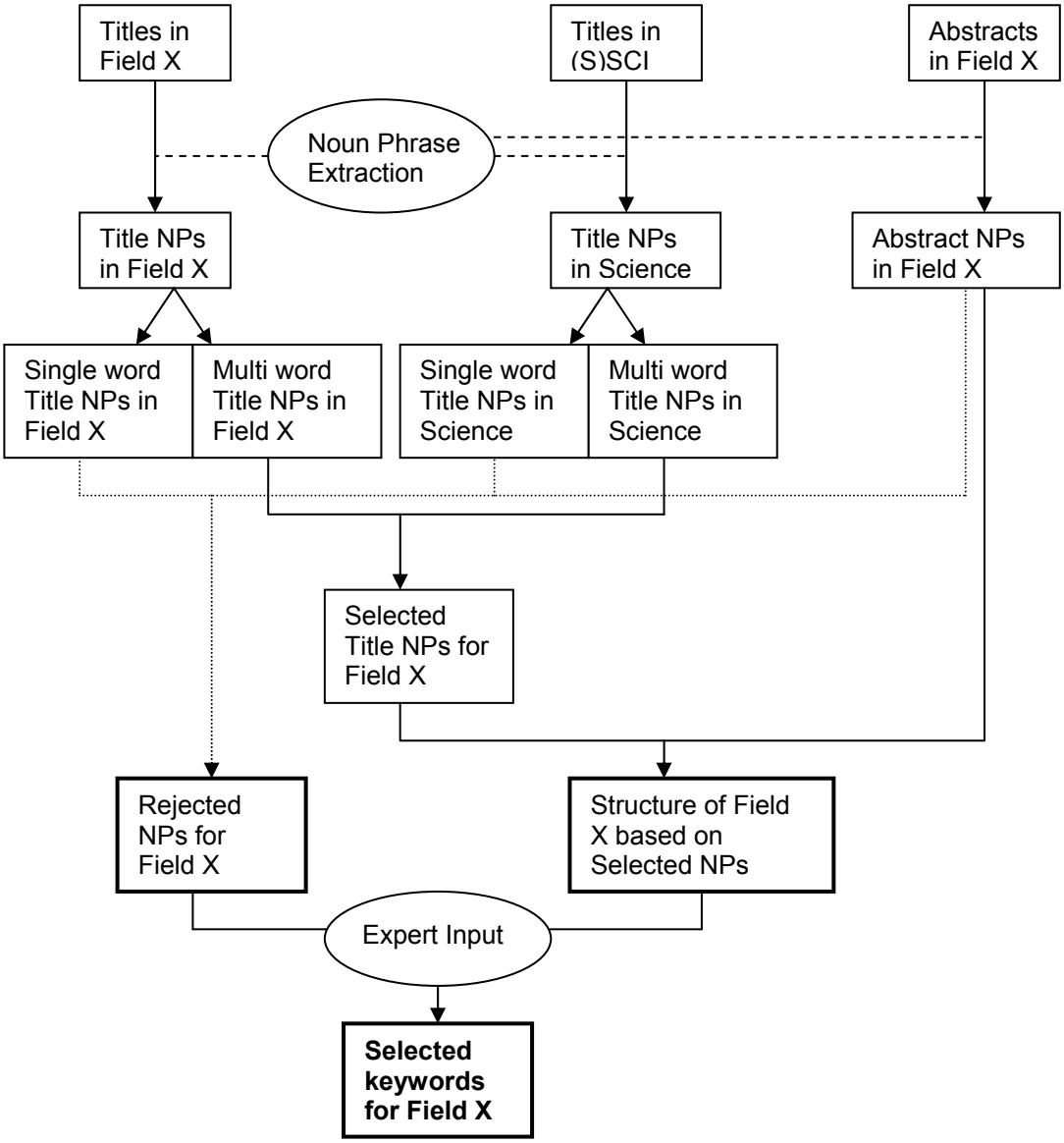


Figure 11-3 Flowchart of the keyword selection procedure

In the procedure, we start with the titles and abstracts from publications representing a particular field *X*. From the titles of these publications, the NPs are identified by linguistic analysis. Of the most frequently used NPs we collect the numbers of hits within the field and within the 'whole of science' (indicated with (S)SCI). As

discussed earlier, the ratio for each NP indicates its specificity. Furthermore, the NPs are divided into multiple word NPs and single word NPs. The specificity index of multiple word NPs is used to exclude certain candidates being too less field-specific. From the list *single* word NPs those with a significant specific character are also transferred to the selected candidates. The selected candidates are subject to a co-word clustering analysis. The frequencies of these NPs are retrieved by searching in titles *and* abstracts. The generated structure (keywords assigned to clusters) are proposed to the expert. The list of initially rejected candidates is also presented to the expert.

In this chapter we propose a procedure to select field keywords from a publication database by field experts. The procedure is based on the specific character of title NPs and the combination of several characteristics of words. Moreover, by presenting the lists of keywords in a structured way, we claim that the results will be most valuable.

References

Arppe, A (Web Page). Term Extraction from Unrestricted Text. A short paper presented at the 10th Nordic Conference on Computational Linguistics 1995. http://www.lingsoft.fi/doc/nptool/term-extraction.html.

Bennett, N.A., Q. He, C. Chang, and B.R. Schatz (1997). Concept Extraction in the Interspace Prototype.

Courtial, J.P. (1998). Comments on Leydesdorff's article. *Journal of the American Society for Information Science* 49. 98-98.

Egghe, L. (1999). On the Law of Zipf-Mandelbrot for Multi-Word Phrases. *Journal of the American Society for Information Science* 50. 233-241.

Healey, P., H. Rothman, and P.K. Hoch (1986). An experiment in Science Mapping for Research Planning. *Research Policy* 15. 233-251.

Hicks, D. and J.S. Katz (1996). Where is Science Going?. *Science, Technology & Human Values* 21. 379-406.

Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Running Text. In: *Papers presented to the 13th International Conference on Computational Linguistics*, Vol. 3. 168-173.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995). *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin and New York.

Leydesdorff, L. (1989). Words and Co-words as Indicators of Intellectual Organization. *Research Policy* 18. 209-223.

Leydesdorff, L. (1997). Why Words and Co-Words Cannot Map the Development of the Sciences. *Journal of the American Society for Information Science* 48. 418-427.

Peters, H.P.F. and A.F.J. van Raan (1993). Co-word based Science Maps of Chemical Engineering, Part I: Representations by Direct Multidimensional Scaling. *Research Policy* 22. 23-45.

Smith and Devine (1985). Storing and Retrieving Word Phrases. *Information Processing and Management* 21, 215-224.

Voutilainen, A. (1993). NPtool, a Detector of English Noun Phrases. In: *Proceedings of the Workshop on Very Large Corpora*.

Whittaker, J. (1989). Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis. *Social Studies of Science* 19. 473-496.

Zamir, O. and O. Etzioni (1998). Web Document Clustering: A Feasibility Demonstration. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 46-54.