Pattern Mining across Domain-specific Text Collections

Lee Gillam and Khurshid Ahmad

Department of Computing, School of Electronics and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom
{l.gillam, k.ahmad}@surrey.ac.uk

Abstract. This paper discusses a consistency in patterns of language use across domain-specific collections of text. We present a method for the automatic identification of domain-specific keywords – specialist terms – based on comparing language use in scientific domain-specific text collections with language use in texts intended for a more general audience. The method supports automatic production of *collocational networks*, and of networks of *concepts* – thesauri, or so-called ontologies. The method involves a novel combination of existing metrics from work in computational linguistics, which can enable extraction, or *learning*, of these kinds of networks. Creation of ontologies or thesauri is informed by international (ISO) standards in *terminology science*, and the resulting resource can be used to support a variety of work, including data-mining applications.

1 Introduction

A measurable difference appears to exist between language used in specialist communications and language used to communicate with a more general audience. The difference suggests that specialists are relatively disciplined in their language use, and provides the opportunity for automatic processing of natural language texts, for example for identification of the keywords of an *arbitrary* specialisation. On this premise we have explored a method that contrasts frequency lists obtained from predominantly scientific, domain-specific, text collections with lists from general language texts. The method is based on a novel combination of statistical measures originating in corpus linguistics, and is supported by developments in international (ISO) standards in relation to *Terminology Science*, and in the development of ontologies in the computing discipline of *Knowledge Engineering*; such ontologies may be construed as modern-day thesauri.

Approaches to the identification of domain-specific keywords – the *terminology* of the domain – generally rely on extensive prior linguistic knowledge, perhaps embodied in an initially linguistic extraction technique. Our method differs from these approaches in being a primarily statistical treatment that could significantly reduce the amount of prior linguistic knowledge needed. The results of this statistical analysis can be augmented using linguistic techniques such that the entire process bootstraps

itself. The analysis produces collocational networks suitable for use in visualizing sequences of texts [1]. Use of various interchange formats can make such networks available for use as a thesaurus, as a terminology, or as an ontology. In these forms, the results become useful for tasks such as document management or query expansion, and may also provide a means of feature selection for data mining applications involving text collections. The automatic identification of such patterns has long been a goal of terminologists. Approaches to identification of such patterns are generally considered from the perspective of a single collection (corpus) of texts with a single specialism (subject field) in mind. Our work has been undertaken using several such corpora from different specialisms to explore generalisation of the approach.

We consider that since these automatically identified domain-specific keywords – terms – are an artefact of notions, ideas or thoughts – *concepts* – then inter-relations between the terms provides evidence of the conceptual organisation (ontology or thesaurus) of that domain.

2 Automatically Extracting Terminology / Ontology

While for the Information Retrieval community a *term* seems to be any word that relates to a document or query, stop words aside, for the terminology community, a *term* is "a "verbal designation of a general concept in a specific subject field" (ISO 1087-1, ISO 12620). The phrase "verbal designation" may be misleading, and the phrase "general concept" is the subject of debate, however "specific subject field" indicates the treatment of particular specialisms. In both communities, the notion of statistical significance has been used to identify a term. Statistical significance, for IR purposes, is a function of rarity across a collection of documents: more occurrences of the keyword(s) in fewer documents. Statistical significance for terminology purposes, on the other hand, can be conceived of as a function of rarity in contrast with what is considered to be *general language*. By consideration of this statistical significance, a task variously referred to as *terminology extraction* / *terminology structuring* [2] or *ontology learning* [3] is possible; an intermediary of this activity may be a *collocational network* [1]. Papers on ontology learning and terminology extraction, and on information extraction, enumerate three techniques: (i) statistical [4], [5]; (ii) linguistic [6]; and (iii) hybrid [7], [8]. Typically, approaches to ontology learning from text employ syntactic parsing [9], [10], [11], [12], [13]. Some authors augment their approach using TF/IDF, word clustering, and coded linguistic relationships [10]. Hybrid techniques may include sophisticated classification systems (neural networks and other learning algorithms), and rely on the frequency counts of linguistic units. Statistical measures, such as log-likelihood and mutual information, may be used to rank information extracted linguistically [8], but predominantly statistical approaches to this task are scarce.

We have developed a statistical method to identify collections of these domain terms – the terminology – automatically, and we use the terms as the basis for thesauri, or the latter-day *ontologies*. These terms are extracted from collections of text in specific subject fields.

2.1 Method

We are interested in the Quirkian notion that frequency of use of words correlates with acceptability of those words as part of the vocabulary [14]:33. Our method uses the 100 million-word British National Corpus (BNC) [15] as reference collection of *general* language. We consider similarities between BNC and specialist corpora as can be derived from Zipf's Law [16], which produces a difference between general language and specialist language and suggests a similarity exists between language use in different specialisms: the approach may be generalisable to other specialist texts. Subsequently, we use a *weirdness* calculation [17], based on one of Malinowski's observations, that has been adapted by *smoothing* [18], to seed a *collocation extraction* technique [19] that results in the production of a network of terms/concepts. By reference to international standards (ISO) for terminology, we can facilitate the construction of terminological resources that have a potentially wider scope of use as thesauri or ontologies. Such a terminology/thesaurus/ontology is then suitable for validation by experts.

For our analysis we considered five specialist text corpora of various sizes: 4 collated at Surrey, consisting of full texts from automotive engineering, nuclear physics, finance and nanoscale science and design, and a fifth from the MuchMore Springer Bilingual Corpus consisting of abstracts from medical journals. Examples presented are generally from one of these corpora – concerned with nanoscale science and design – but similar results have been obtained from all of these corpora, and are the subject of ongoing expert evaluation, and further analysis, in other work. If different specialisms use language in similar ways, it may be possible to systematically extract terminology, or thesauri, or ontology, from *arbitrary* collections of specialist text. In our treatment, we make a distinction between **tokens** (words occurring at various locations in texts) and **types** (the different words used).

2.2 Comparing Text Corpora

The contrast between the general, everyday use of English with that of specialist use has been observed empirically and recently quantified using Zipf's Law. We consider the British National Corpus as an example of general language, alongside five specialist domain corpora, data for which is presented in Table 1.

Table 1. Type and token counts for the 6 corpora that are the subject of our analysis.

	Automotive	**Nuclear**	**Finance**	**Medical**	**Nanoscale**	**BNC**
Tokens	350920	393993	685037	1081124	1012096	100106029
Types	14252	14937	28793	35033	26861	669417

Counting words in collections of text provides a means by which to study properties of language use. The first 100 most frequently occurring types in a text corpus have been shown to comprise just under half the corpus, which is true of both specialist and non-specialist corpora [20]. The first 100 most frequent types within a specialist text collection of around half a million words comprises between 30 and 40 *open class words (types)* – predominantly nouns specific to the domain. The distinc-

tion between open-class words and *closed-class words* – grammatical words, for example determiners and conjunctions, that tend to appear in stop-lists – requires human judgement and prior (domain) knowledge: a word such as *keep* may occur in a stop-list for IR, but is important in documents relating to *medieval architecture*. Since we seek to automate our approach, we need to determine how language behaves by considering characteristics of language use.

George Kingsley Zipf's power-law function has been used to show that for a text corpus of N tokens, the rank of a word (type) multiplied by its frequency produces a constant of N/10 (e.g. for the Brown Corpus [21]:26-27). Zipf's law has been tested in analysis of large (sub-)corpora of newswire texts (the Wall Street Journal for 1987, 1988 and 1989: approximately 19 million, 16 million and 6 million tokens respectively) and shows similar behaviour between these collections, with little deviation due to corpus size [22]. In these analyses, Zipf's law holds only for a certain range of ranks: at high frequency and low frequency similar patterns of deviation from Zipf's law are encountered that may be accounted for by the Zipf-Mandelbrot law. Our chosen corpora have similar deviations from Zipf's law, however application of Zipf's law to words of low frequency - the *hapax legomena* – produced a clear difference between general language and specialist language (Fig.1.).

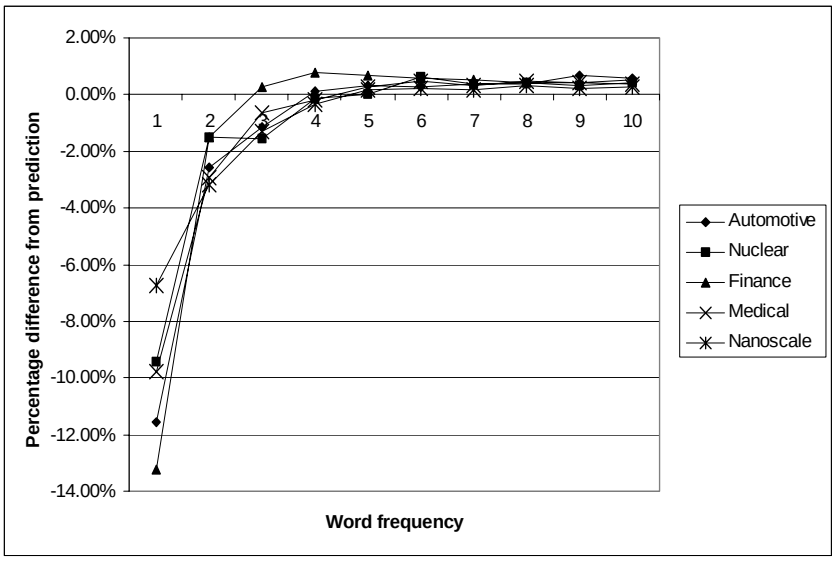

Fig. 1. Difference between percentage of types (y-axis) in the specialist corpora and predicted values derived from Zipf's law (0.00%) for frequencies between 1 and 10 (x-axis) for 5 specialist corpora. For frequency 1, specialisms have around 37-43% of types compared to expectation (50%) and BNC (53%).

By Zipf's law, 50% of words should occur at frequency 1. In the BNC, this figure is around 53%, whereas for the specialist corpora only 37-43% of words occur once. Additionally, within this *hapax legomena*, Zipf's law suggests that 90% of words should occur with frequency of 10 or less. For BNC, this is 84.45%, while it is around 80% for the specialist corpora. A first inference from such results is that spe-

cialist corpora, and the specialists who contribute to them, are disciplined in their language use since they deviate in the same manner from Zipf's law and from a general language resource such as BNC. A second inference is that the reduced values for specialist corpora fit better with Zipf's *principle of least effort* [16]: 22-23.

2.3 Automatic Single-word Term Extraction

The empirical observations above have led us to develop the notion of *weirdness* [17]. On the basis of similar properties of corpora, we compare frequencies of words (types) in the specialist corpora to frequencies in general language. This computation has been adapted using an additive smoothing technique to account for words not found in general language [18]. For the selected specialist corpora, between 13% and 38% of types are not found in the general language contained in the BNC (Table 2).

Table 2. Words (types) in the specialist corpora that do not occur in the reference corpus.

Corpus	Types	Tokens	Number of types not in BNC	% types not in BNC
Automotive	14252	350920	1794	13%
Nuclear	14937	393993	4190	28%
Finance	28793	685037	4718	16%
Medical	35033	1081124	11395	33%
Nanoscale	26861	1012096	10231	38%

Weirdness, smoothed, is calculated as

$$weirdness = \frac{N_{GL} f_{SL}}{(1+ f_{GL}) N_{SL}} \quad (1)$$

where f_{SL} is the frequency of word in the specialist corpus, f_{GL} is its frequency in BNC, and N_{SL} and N_{GL} are the token counts of the specialist corpus and the BNC respectively. For example, for the Nanoscale corpus, we can produce a list of words and a combination of frequencies and weirdness values which suggest their domain-specificity (Table 3).

Table 3. Words, their frequencies in a specialist corpus and the BNC, and the result of weirdness and its smoothing.

Word	Freq	BNC	Weirdness
nanowires	619	0	61225
nanoparticles	829	1	40998
nanowire	360	0	35607
nanotube	969	2	31948
nanoscale	268	0	26508
tunneling	514	1	25420
nanoparticle	232	0	22947

Highly frequent words include closed class words – *the* being most frequent in English; across corpora a weirdness value for *the* of around 1 is obtained. The com-

bination of high frequency and high weirdness is therefore of interest, and can automate removal of these closed class words (stop lists). This may be suitable for stop lists for other languages and purposes also. Resulting lists of frequency and weirdness values can be treated systematically by taking z-scores of each. Where z-score for **both** frequency and weirdness is above a given threshold, the words provided by this mechanism are used in subsequent analysis (Table 4).

Table 4. Number of words selected by z-score thresholds for both frequency and weirdness.

	Automotive	**Nuclear**	**Finance**	**Medical**	**Nanoscale**
z-score					
5	0	0	0	0	1
4	0	0	0	0	5
3	0	1	0	0	6
2	0	3	2	1	8
1	7	6	3	4	19
0	154	176	186	494	352

2.4 Automatic Multiword Term Extraction

While some single-words may be terms in their own right, terms in specialist domains tend to be formed from a number of single words to identify more specific concepts: for example as *multiword terms*. Multiword terms in English generally exclude closed class words, punctuation marks and numerals, although chemical, mathematical and, indeed, nomenclatures for logic use a variety of hyphenation, numerals and other symbols that are important in that subject field. The frequency with which words appear in close proximity to each other has been variously analysed. Magnusson and Vanharanta have created collocational networks for visualising sequences of texts [1] using the "information theoretic concept of mutual information" [23]. Elsewhere, Church's t-score has been used for calculating strength of association [24]:34. We have found both measures to be limited on three counts: first, the selection of words for treatment by both seems to be arbitrary; second, both metrics take no account of features of the neighbourhood of the selected word(s); third, both metrics consider only two words together. On the first issue, our weirdness-frequency combination seems to offer a solution; for the second and third we consider Smadja's work on collocations [19]. Smadja uses a neighbourhood of five words and records frequency of words occurring at each position. If the two words consistently appear together in the same relative position in contrast to other positions, this *collocation* is deemed significant. We refer to a process of using such significant collocates as inputs to a subsequent collocation phase as *re-collocation*. This expands a collocation network systematically, depending on the satisfaction of Smadja's constraints. Table 5 shows a sample of the words that collocate in the five positions either size of *carbon* (frequency of 1506 in about 1 million tokens), in the nanoscale science and design corpus, and that satisfy Smadja's constraints.

Table 5. Collocations with *carbon* (frequency of 1506) in the Nanoscale science corpus.

Collocate	Freq	-5	-4	-3	-2	-1	1	2	3	4	5
nanotubes	690	8	8	9	2	0	647	6	0	7	3
nanotube	252	3	2	2	0	0	229	2	1	5	8
single-walled	77	0	0	1	1	75	0	0	0	0	0
aligned	94	1	1	3	5	74	0	1	1	3	5
multiwalled	70	1	1	2	0	59	0	0	1	5	1
amorphous	58	1	1	6	0	46	0	1	1	0	2
atoms	51	1	2	0	1	0	42	0	1	3	1
nanotips	44	0	2	1	1	0	39	0	0	1	0

Re-collocation of carbon nanotubes produces collocating words such as those in Table 6.

Table 6. Collocations with *carbon nanotubes* (frequency of 647) in the Nanoscale science corpus.

Collocate	Frequency	-5	-4	-3	-2	1	1	2	3	4	5
single-walled	73	0	0	1	1	71	0	0	0	0	0
aligned	63	1	1	1	5	48	0	0	2	4	1
multiwalled	53	0	0	1	0	46	0	0	5	1	0
properties	60	1	4	15	32	0	0	0	6	2	0
multiwall	34	0	1	0	1	30	0	2	0	0	0
single-wall	26	0	0	1	0	24	0	0	0	1	0

2.5 Terminology/Ontology Construction

The *collocation network* in our case is a tree branching out from the originally selected list of words. Gillam has shown how this tree can be encoded in conformity with international standards for the production of terminology interchange formats using ISO 12620 and ISO 16642 and for use with so-called ontology exchange languages [25]. A fragment of such a tree, resulting from the above analysis, and suitable for such encoding, is shown in Fig. 2.

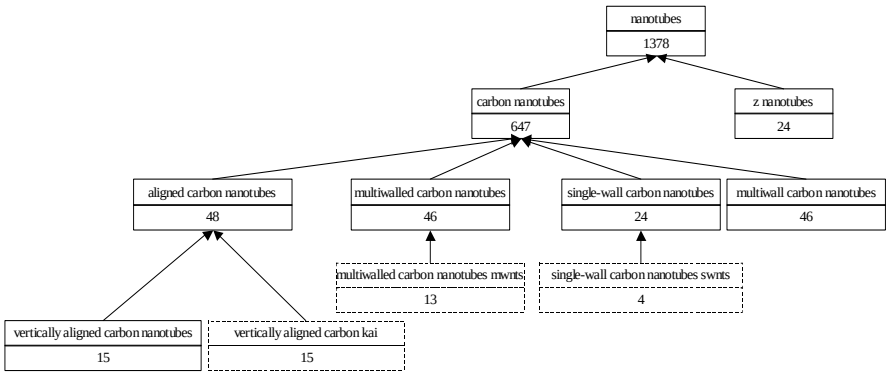

Fig. 2. Fragment of the resulting ontology from a corpus of Nanoscale science and design texts. Dotted outlines denote collocations deemed to be invalid.

3. Discussion

The method discussed relates to Minsky's "thesaurus problem" [26]:27 of building and maintaining a thesaurus useful for a specific task or set of tasks, of learning to build thesauri, and of finding "new ways [to] make machines first to use them, then to modify them, and eventually to build for themselves new and better ones". A thesaurus has been described as "one type of ontology, one specialized to information retrieval" [27], and is built to show the inter-relationships between words. Unlike a dictionary, which is usually alphabetically organized and seeks to elaborate meaning, a thesaurus is developed to elaborate the conceptual (hierarchical and part-whole) relationships between a word, or more specifically a word related to a concept, and one or more other words. Our method for constructing ontologies extends work on construction of *collocational networks*. These ontologies, or terminologies, or thesauri, can be extracted automatically from text collections, and show how knowledge in a specific subject field is organised. Such resources may be useful for the organization of this scientific information. Other applications include: generating buy/sell signals in financial trading [28], health policy communication [29], digital heritage [30] and query expansion for multimedia information retrieval [31], [32]. The resulting ontology has been exported to a *de facto* standard ontology editor – Protégé – for viewing and editing. Here it becomes the basis for developing intelligent (rule-based systems) using applications such as JESS [33]. The ontology also enables text-based *feature selection* to be made, which may be useful for systems such as WebSOM [34].

Initial discussions with domain experts have validated the first results with some degree of confidence, and we are studying the effects of increasing the length of multiword patterns being generated, against decrease in frequency. For example, at low

frequencies that are not statistically validated, we have *aligned single-walled carbon nanotubes* (2) and *large-diameter single-walled carbon nanotubes* (2). Subsequent effort is required to classify collocations into *facets*: in the examples presented, the types of carbon nanotubes appear to have "walledness" as a significant facet, and being *aligned* has importance also, though is perhaps a value of a different facet. Determining this distinction currently requires expert input. *properties* is not a positionally valid collocation – though we can infer that *properties* of *carbon nanotubes* are described in this collection. We have considered combined use with linguistic patterns elsewhere [35].

Acknowledgements. This work was supported in part by research projects sponsored by the EU (SALT: IST-1999-10951, GIDA: IST-2000-31123, LIRICS: eContent-22236) and by UK research councils: EPSRC (SOCIS: GR/M89041/01, REVEAL: GR/S98450/01) and ESRC (FINGRID: RES-149-25-0028).

References

1. Magnusson, C. and Vanharanta, H.: Visualizing Sequences of Texts Using Collocational Networks. In Perner, P. and Rosenfeld, A. (Eds): MLDM 2003, LNAI 2734 Springer-Verlag, Heidelberg. (2003) 276-283
2. Grabar, N. and Zweigenbaum, P.: Lexically-based terminology structuring. Terminology 10(1). John Benjamins, Amsterdam (2004) 23-53.
3. Maedche, A.:. Ontology Learning for the Semantic Web. The Kluwer International Series in Engineering and Computer Science, Vol. 665. Kluwer Academic Publishers (2002).
4. Salton, G.:. Experiments in Automatic Thesauri Construction for Information Retrieval. In Proceedings of the IFIP Congress, Ljubljana, Yugoslavia. Vol. TA-2. (1971) 43-49.
5. Jing, Y. and Croft, W.B.: An Association Thesaurus for Information Retrieval. In Bretano, F., Seitz, F. (eds.), Proc. of RIAO'94 Conference, CIS-CASSIS, Paris, France (1994) 146-160.
6. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Boston, USA: Kluwer Academic Publishers (1994)
7. Drouin, P.: Term extraction using non-technical corpora as a point of leverage. Terminology 9(1). John Benjamins, Amsterdam (2003) 99-115.
8. Vivaldi, J. and Rodríguez, H.: Improving term extraction by combining different techniques. Terminology 7(1). John Benjamins, Amsterdam (2001) 31-47.
9. Maedche, A. and Volz, R.: The Ontology Extraction and Maintenance Framework Text-To-Onto. Workshop on Integrating Data Mining and Knowledge Management. California, USA (2001)
10. Maedche, A. and Staab, S: Ontology Learning. In S. Staab & R. Studer (eds.): Handbook on Ontologies in Information Systems. Heidelberg: Springer (2003).

11. Faure, D. and Nédellec, C.: Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM. LNCS 1621. Springer-Verlag, Heidelberg. (1999) 329-334.

12. Faure, D. and Nédellec, C.: ASIUM: Learning subcategorization frames and restrictions of selection. In Y. Kodratoff, (Ed.), 10th Conference on Machine Learning (ECML 98), Workshop on Text Mining, Chemnitz, Germany. (1998).

13. Mikheev, A. and Finch, S.: A Workbench for Acquisition of Ontological Knowledge from Natural Text. In Proc. of the 7th conference of the European Chapter for Computational Linguistics (EACL'95), Dublin, Ireland. (1995) 194-201.

14. Quirk, R.: Grammatical and Lexical Variance in English. Longman, London & New York (1995)

15. Aston, G. and Burnard, L.: The BNC Handbook: Exploring the British National Corpus. Edinburgh University Press (1998).

16. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Hafner, New York. (1949).

17. Ahmad, K. and Davies, A.E.: Weirdness in Special-language Text: Welsh Radioactive Chemicals Texts as an Exemplar. Internationales Institut får Terminologieforschung Journal 5(2). (1994) 22-52.

18. Gale, W. and Church, K. W.: What's wrong with adding one? In Oostdijk, N. and de Haan, P. (eds.): Corpus-Based Research into Languge: In honour of Jan Aarts. Rodopi, Amsterdam (1994), 189-200

19. Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics, 19(1). Oxford University Press. (1993), 143-178

20. Ahmad, K.: Neologisms to Describe Neologisms: Philosophers of Science and Terminological Innovation. In (Ed.) Peter Sandrini: Proc. of Terminology and Knowledge Engineering (1999), 54-73.

21. Manning, C. and Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA. (1999)

22. Ha, L. Q., Sicilia, E. , Ming, J. and Smith, F. J.: Extension of Zipf's law to words and phrases. In Proceedings of International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan. (2002), 315-320

23. Church, K.W. and Hanks, P.: Word association norms, mutual information and lexicography. In Proceedings of the 27th Annual Conference of the Association of Computational Linguistics (1989), 76-82.

24. Jacquemin, C.: Spotting and Discovering Terms through Natural Language Processing. MIT Press. Cambridge, MA. (2001)

25. Gillam, L.: Systems of concepts and their extraction from text. Unpublished PhD thesis, University of Surrey. (2004).

26. Minsky, M.: Semantic Information Processing. MIT Press (1968)

27. Oard, D.W.: Alternative approaches for cross-language text retrieval. In AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence. (1997).

28. Gillam, L. (Ed): Terminology and Knowledge Engineering: making money in the financial services industry. Proceedings of workshop at 2002 conference on Terminology and Knowledge Engineering (2002).

29. Gillam, L. and Ahmad, K.: Sharing the knowledge of experts. Fachsprache 24(1-2). (2003), 2-19.

30. Gillam, L., Ahmad, K. Salway,.: Digital Heritage and the use of Terminology. Proceedings of Terminology and Knowledge Engineering. (2002)

31. Ahmad, K., Tariq, M., Vrusias, B. and Handy, C.:. Corpus-Based Thesaurus Construction for Image Retrieval in Specialist Domains. ECIR 2003, LNCS 2633. Springer Verlag, Heidelberg (2003), 502-510.

32. Vrusias, B. Tariq, M. and Gillam, L.: Scene of Crime Information System: Playing at St Andrews. CLEF 2003, LNCS 3273. Springer Verlag, Heidelberg (2004), 631-645.

33. Eriksson, H.: Using JessTab to Integrate Protégé and Jess. IEEE Intelligent Systems 18(2). (2003), 43-50

34. Kohonen, T., Kaski, S., Lagus, K. Salojärvi, J., Honkela, J., Paatero, V. and Saarela, A.: Self Organization of a Massive Document Collection. IEEE Transactions on Neural Networks 11(3). (2000), 574-585.

35. Gillam, L., Tariq, M. and Ahmad, K.: Terminology and the Construction of Ontology. Terminology. John Benjamins, Amsterdam. Terminology 11:1 (2005), 55–81.