Knowledge Sharing in a Community of Practice: a Text-Based Approach in Emergent Domains

Rafif Al-Sayed and Khurshid Ahmad
Department of Computing, University of Surrey, U.K.
r.al-sayed@surrey.ac.uk
k.ahmad@surrey.ac.uk

Abstract: The shared use of specialist terminology amongst the members of a community of practice is explored as evidence for the existence of the concept of *communal lexicon*. A computer-based method of investigating the extent of terminology is described – this method uses both univariate analysis, specifically frequency distribution of single and compound words, and multivariate statistical analysis, particularly factor analysis. The results show that terminology sharing may act as a metric for knowledge sharing and knowledge diffusion among different (sub-) communities. The case study chosen to demonstrate the efficacy of the terminology-sharing method is drawn from *cancer care* – especially breast cancer care, where texts produced for and by the three main components of the community are examined – namely the experts, the professionals and the patients.

Keywords: Knowledge diffusion and sharing, community of practice, communal lexicon, corpus linguistics, special language terminology, multivariate analysis.

1. Introduction

The fusion of knowledge, within and across domains, is critical for the sustenance of individual domains and for the well being of the society as a whole. Knowledge management literature shows that the application of knowledge and the feedback from end-users, in itself contributes as substantially to the knowledge of the domain experts as does the research output of the experts. The classical case studies of Japanese white-goods manufacturers, Matsushita and Cannon (Nonaka and Takeuchi, 1995) of the German conglomerate Siemens' recovery as major telecommunication enterprise during the 1990's (Davenport and Prusak, 2002), and of knowledge transfer within the photocopier division of Xerox Inc, clearly indicate the benefits the research laboratory derives from its interaction with the professionals and the end-users. The professionals include design engineers, marketing executives, accounts clerks, merger and acquisition lawyers – those who have to understand, critique and apply the knowledge developed in laboratory.

A large organization, with its different highly specialized interest groups may become a community – with shared values, common goals, a belief in the organization, and the community shares a tension as well. The tension is between a commitment to keep the identity of the organization whilst being prepared for change – both large and small-scale change. It appears that the vital force behind the successful organization is its willingness and ability to foresee change and adapt to it. Above all, a community of practice – a group of people who share a set of problems, or passion about a topic, and who deepen their knowledge and expertise in a given area, by self motivation interacting on an on-going basis (Wenger et al, 2002:4) – shares a common language that facilitates interaction amongst the members of the community. The interaction between people is very manifest when they use language – for linguists like Herbert Clark (1996) the use of language is a form of joint action that, in turn, is based on actions of each of the individuals involved in language based communication. Our work is an attempt to understand how knowledge communities emerge by observing their principal form of 'joint action' – the use of language within the community.

Like any other social community, members speak different dialects of the same common core language – working, middle and upper class English, Hindi, or Arabic. Each division of the community uses the common core for a different purpose: but despite the dialectal and pragmatic differences, and in some cases mutual unintelligibility, the community shares a *communal lexicon*. Each division may have their exclusive words, which may in time be shared or not shared across the community. The community sometimes accepts new words to denote new concepts, objects or events, and rejects some of its current stock of words related to obsolete concepts, objects or events.

Our investigation is based on the acceptance of the notion of a *common lexicon*, and we believe that an examination of documents produced by different members of a community of practice either synchronically, at the same time, or diachronically, over a fixed period of time, will indicate the extent of a community's cohesion or otherwise. We are interested in the dimensions of

Reference this paper as:
Al-Sayed R and Ahmad K (2006) "Knowledge Sharing in a Community of Practice: a Text-Based Approach in Emergent Domains" *The Electronic Journal of Knowledge Management* Volume 4 Issue 2, pp 99 - 108, available online at www.ejkm.org

variation at the level of word usage across a community of practice. The dimensions are an indicator of the extent to which words – or more accurately specialist terms – that are generally used to label concepts are shared across the community.

2. Motivation

The notion of a community of practice is a qualitative one and is underpinned by more abstract concepts like common ground. This term is used in language acquisition to suggest one can acquire language if there was to be a common ground between the larger linguistic community and the individual learners and that once the learner has accomplished the mastery of language, his or her contribution to the community will increase. Herbert Clark (1996) is one of the main proponents of that concept in applied linguistics, Davenport and Prusak (1998) and Wenger (1999) have stressed this concept in knowledge management: An organization facilitates exchange of knowledge through a common vocabulary that is used in formal and structured documents within the organization, for example, research reports and marketing brochures. Knowledge is equally well shared through informal and unstructured documents including organizational stories, interoffice memos and emails.

Our focus will be on the formal and structured documents and we will be analyzing the language used in the documents at the lexical level and in particular, our focus will be on the so-called lexical words rather than the so-called grammatical words. The fact, notwithstanding that this categorization of the words of a language in itself is controversial; the categorization has an intuitive appeal. Put simply, grammatical words are used much more frequently than all the other words and these words generally comprise fewer letters: e.g. in the English language the word “the” comprises over 5% of any written text followed by “an”, “of, “in”,”a”,“it”, “that”, “of”,”to”, “is” and “was”. These words make off 25% percent of most texts in English, and it is used for everyday or for general purposes.

The difference is in the usage of lexical words: then words, mainly nouns and adjectives, do distinguish between different specialist enterprises. The frequency of the word “telecommunication” will be very high in Siemens’ documents when compared, say, with documents produced by Xerox Inc; the converse will be true of the term “photocopier”. The choice of lexical words, incorporated in the terminology of the specialist domain changes over time in that the individual terms refer to the key concepts of the domain. The changes in the concepts, or changes in the preferential treatment of concepts, are reflected in terminology of a domain: for example, IBM Corp is no longer for its once world famous electric *typewriters*, and there is less emphasis on the term *mainframe* – *IBM Corp* now claims to be a *consultancy company*. The lexical content of *IBM’s* documents – research papers, inter-office memoranda, and marketing brochures- indicate the changes in the strategic outlook of the company. The choice of the certain lexical words, and the low-frequency usage of others, is a hallmark of a specialist domain (Ahmad, 1995). Does the communal lexicon of Clark, Wenger and others, manifest itself in the writings of the different members of the domain community? And more specifically, is there an idiosyncratic lexical signature of the domain community?

Our basic hypothesis is that much like the special language community shares aspects of a natural language with the broader linguistic community (in which the specialists are embedded), a community of practice shares aspects of its special language and preferentially use some constructs of the specialist language, coins its own terms, and avoids using terms used in the broader specialist community. One corollary of our hypothesis is that the changes in the lexical preferences of parts of a community of practice are an indication of knowledge diffusion. In this context it has been argued that one can ‘extract’ aspects of the conceptual system –or more ambitiously the ontology- of a domain using the lexical signature with some degree of success in domains as various as nuclear physics and forensic science, orthopedics and art criticism (see, for example Gillam, Ahmad & Tariq 2005 and Gillam & Ahmad 2005).

3. Method

The detection of the so-called lexical signature of a specialist domain has been of interest to researchers in Language for Special Purposes (LSP). An LSP is the variant of a natural language used in a specialist domain. A study of randomly selected collections of specialist texts, a specialist text corpus, is a good source for the terms and there are methods and techniques for automatically extracting terms (see, Ahmad and Al-Sayed 2005, for details and references therein). Typically, the distribution of individual single or compound words, or uni-variate analysis of individual random variables, is studied to study the nature and function of a given text (Manning and Schutze 1999). However, this uni-variate analysis is based on a number of simplifying assumptions about how a single lexical items

contributes to the 'make up' of a text at different levels of linguistic descriptions – lexical, syntactic, semantic, pragmatic- or different levels of conceptual descriptions – epistemological, ontological or logical. We use the distribution of single words to identify a specialist community – more preferentially used words show the ontological commitment of a community (Ahmad and Al-Sayed 2005).

In social and biological science the focus is usually two or more variables –many variables in fact- appear to affect the behaviour of a person or a system. Here techniques of multi-variate analysis are used to deal with the correlated behaviour of many variables; one of the techniques – with its own simplifying assumptions about how the variables may or may not be correlated- is *factor analysis*: According to Wikipedia, factor analysis seeks to 'explain most of the variability among a number of observable random variables in terms of a smaller number of unobservable random variables called *factors*. The observable random variables are modeled as linear combinations of the factors, plus "error" terms.'. We will be using factor analysis to look at the distribution of compound words with a view to *identifying* a smaller number of factors. The factors will help us to distinguish various sub-communities that may constitute a specialist community above the level of the community but below that of a specialism as a whole.

Specifically, we wish to examine patterns of correlation between a large number of (multi-word) or compound terms with a view to extract the main underlying *factors*. Each of the factors or *dimensions* is independent of the other factors that are extracted (automatically) from a study of the compound terms: Each dimension may be expressed as a linear combination of two or more compound terms; sometimes one term may explain most of the variation along a given dimension. The main intention here is to quantify the intuition that the authors of a specialist text use a number of terms to emphasize or de-emphasize a concept, to highlight aspects of a theory, or to report the results of an experiment. The inclusion of factor analysis as the basis of studying the influence of the compound terms (by computing the variances due to the terms) is, an addition to our reported method, based on Everitt and Dunn's algorithm of principal component analysis and factor analysis (2001: 50-51, 271).

Our method depends on the creation of a text corpus for the specialism including sub-corpora for different components of a community of practice. The corpus is then subjected to uni- and multi-variate algorithms:

Univariate Analysis: Single term and compound term detection

a. Single terms are extracted using the ratio of relative frequency of a term in a special language corpus and its corresponding relative frequency in a general language corpus, using averages and standard deviations for computing z-scores of frequency and frequency ratios – the z-score computation involves univariate analysis

b. Compound terms are detected by measuring the collocation of two or more words – joint frequency of distribution of the components of a compound terms within a window of 5 words and the computation of histograms and the z-score of the collocates, both involving the computation of univariate statistics.

Multi-variate Analysis Comparison of frequency distribution of the terms across the community of practice:

a. Let x={x1, x2, …xp} be a set of compound terms;

Let yj = aj * x, be the jth the principal components of the observations x, such that aj*aj-1= 1and aj*ai-1=0 when j ≠ i;

The variance of yj is given as Var (yj) = aj-1 S aj; and

The total variance of the p principal components is computed from the eigenvalues of S –called λ: [λ1+λ2+ … λp= Trace (S)]

The jth principal component accounts for a proportion Pj of the total shared variation on theoriginal data, where

[Pj = λj /(Trace (S))];

The correlation of the ith variable and the jth component is given as

[ri,j=√ λj . aij]

b. Include only those components j where the corresponding λj is greater than unity;

c. Compute the factor scores for each group of texts with respect to each factor where the factors are interpreted as exemplars of a particular group within a community of practice.

4. A case study: Breast cancer-care

Cancer care is one of the key planks of health care systems. The investment in cancer research is considerable both at the national and international level. Cancer care involves experts researching the domain and professional medics and support professionals applying the knowledge of the experts. The professionals provide feedback, extend or restrict the scope of the application of expert's knowledge, and make their own original contribution and establish best practice. Increasingly, patients are being involved directly in the cancer care loop – information is provided to the patients on an on-demand basis and the patients' feed back is also disseminated.

4.1 Input data

The texts used in this case study were drawn mainly from the US-based American Cancer Society and the National Institute of Cancer. The size of a corpus is usually determined empirically – for general language corpus the size is typically around 100 million words for capturing the massive variation in the different uses of the general language across economic and social classes, across the literacy divide and so on. The size of a special language corpus can be determined by arguing that there is an intrinsic limitation on the size of such corpora as the number of authors and readers of specialist text is limited when compared to the general language authors and writers; usually a 1 million word specialist corpora will suffice. The different text types are also smaller for a specialist language and contains mainly learned articles, highly formalized and structured documents like memoranda, research and marketing reports, and instructional texts including user manuals and best practice documentation – the corresponding choices in general language involve a whole raft of imaginative texts (novels, magazines etc).

Our domain of interest is *breast cancer care* and we have collected three kinds of texts – abstracts of journal articles, best practice documentation, and informative literature available on the American Cancer Society website and other websites dedicated to *patients*. Table 1 gives the details of our three sub-corpora in the breast-cancer domain:

Table 1. The composition of our three corpora

Corpus	No. of tokens.	No.of texts	Text Types
Expert	255,144	224	Journal abstracts and full text
Professional	431,856	638	Journal abstracts; Full text articles on best practice and clinical trials
Patient	497,625	420	Informative articles from cancer research charities - full text
Total	1,184,625	1282	

4.2 Distribution of key terms within the three corpora: A univariate analysis

A comparison of the 10 lexical single words most frequently used in all three corpora shows key differences amongst the distribution of single keywords but with the key signature terms of the domain – *breast* and *cancer* given equal preference (see Table 2).

Experts use new terms (Breast cancer gene 1 and 2 abbreviated as BRCA1, BRCA2) much more frequently than the professionals and the patient, the focus of the experts appears to have moved away from 'surgery' but remains that of patients' texts.

The rank correlation of frequently used words in the Expert and Professional corpus is 68% and between Professional and Patient corpus the rank correlation is 57%. There is a weak correlation between the ranks of the Patient and Expert frequent single words (0.28%)

A comparison of the distribution of compound terms reveals a similar picture. We have chosen 10 highest frequency compound terms (with a mutual information greater than or equal to 1): there is only one obvious term that has the same rank *breast cancer* but the other 9 are rather differently distributed (see Table 2b)

Reference this paper as:
Al-Sayed R and Ahmad K (2006) "Knowledge Sharing in a Community of Practice: a Text-Based Approach in Emergent Domains" *The Electronic Journal of Knowledge Management* Volume 4 Issue 2, pp 99 - 108, available online at www.ejkm.org

Table 2a. Sharing or otherwise of frequent single terms ranked according to frequency of all tokens in the three sub-corpora. In all three sub-corpora two key terms are shared (cancer, breast). The highlighted cells in the Table show the predominant use of the key terms in that particular sub-corpus.

RANK			
Single Words	Expert	Professional	Patient
cancer	5	6	6
breast	8	8	7
BRCA1	9	42	67
BRCA2	13	82	240
tamoxifen	221	31	64
chemotherapy	94	35	54
therapy	163	25	48
adjuvant	185	49	242
surgery	244	98	48
lymph	230	107	57

Table 2b. Only one compound term has the same rank in all three corpora *breast cancer*, *ovarian cancer* is shared between two corpora at the same rank, otherwise compound terms are used with different preferences in the three corpora. The term or term components in bold is those that were preferentially used as single words – indicating the lexical productivity in all the three corpora.

Compound Terms	Expert	Professional	Patient
breast cancer(s)	1	1	1
ovarian cancer(s)	2	2	9
mutation carriers	3	7	65
BRCA2 mutation(s)	4	26	58
BRCA1 mutation(s)	6	9	55
estrogen receptor(s)	13	5	35
endocrine therapy	50	8	60
metastatic breast cancer	51	3	24
lymph node(s)	42	27	2
radiation therapy	47	4	3

The rank correlation coefficient between the Expert and Professional corpus is positive (26%), but there is a stronger correlation between Expert and Patient (41%) and weaker between Professional and Patient (10%).

The above observations are based on a pre-knowledge of the language (English) and a working knowledge of the domain by the authors of the paper. Also, the use of the statistics is strictly based on a term-by-term basis. We have benefited from computation to the extent that over 1 million words of text were analyzed and both single- and compound terms were detected automatically.

4.3 Distribution of compound terms: A multivariate analysis

Table's 2a and 2b show that we have to deal with a large number of compound words. These terms when looked up by somebody who is a competent native speaker of English, and knows something about breast cancer, can tell us that these terms are interrelated lexically and semantically with each other. Furthermore, the knowledgeable person can tell us that these terms suggests a common theme throughout the three corpora (e.g. *breast cancer, ovarian cancer*) or that some of the terms are characteristic of each corpus (e.g. *BRCA1 mutations, BRCA2 gene* for experts; *endocrine therapy* and *estrogen receptors* for professionals; *lymph nodes*, and *breast reconstruction* for the patients). The commonalities, distinctions and the apparent relationships between the terms within and across corpora may indicate that these terms are different manifestations of one or more concepts. What is required is the ability to identify terms, categorise the terms, and make statistically well-grounded judgments about the individual and collective distributions of the terms.

Factor analysis provides a quantitative and statistically well-founded method for reducing the number of original variables to a smaller set of derived variables or *factors* (see, for example, Biber, 1988 for an application of factor analysis to the study of variation in spoken and written language); note that we prefer to use the term *dimension*. Each dimension is a linear combination of the individual terms derived from a

correlation matrix of all the terms: if a correlation matrix element is unity then factor analysis method tells us that the two correlating terms will always be found together; if the element is zero, then it is not possible for terms to co-occur.

Consider the correlation matrix of 10 compound terms that are most frequently used when we look at our three corpora collectively (Table 3). We have used the SPSS statistical package to compute the matrix and the rest of the calculations. One can see some terms correlate well with a few other terms whilst others either little or weakly negative correlation. But these judgments, like some made with univariate analysis, can be only made after a visual inspection of the results. Factor analysis helps us to make the statement with the help of multivariate statistics.

Table 3. Correlation matrix for the 10 most frequently used compound terms in our corpus. The term *breast cancer* does not appear to correlate with any of the other nine terms, indeed, it mildly anti-correlates with all others. But *ovarian cancer* correlates positively with *BRCA1* and *BRCA2 mutation,* and *mutation carriers* (BRCA stands for BReast CAncer gene/mutation and so on); *estrogen receptor* only correlates with *endocrine therapy.*

	breast cancer(s)	ovarian cancer(s)	Mutation carriers	BRCA2 mutations	metastatic breast cancer	estrogen receptor(s)	endocrine therapy	BRCA1 mutation	lymph node(s)	radiation therapy
breast cancer(s)	1									
ovarian cancer(s)	-0.12	1								
mutation carriers	-0.02	**0.34**	1							
BRCA2 mutation(s)	0.03	**0.35**	**0.49**	1						
Metastatic breast cancer	-0.05	-0.10	-0.04	-0.08	1					
estrogen receptor(s)	-0.02	-0.06	0.01	-0.01	0.12	1				
endocrine therapy	0.03	-0.09	-0.06	-0.07	0.18	**0.35**	1			
BRCA1 mutation	-0.03	**0.43**	**0.38**	**0.45**	-0.07	-0.01	-0.06	1		
lymph node(s)	-0.07	-0.16	-0.09	-0.09	0.01	0.10	0.01	-0.09	1	
radiation therapy	-0.08	-0.06	-0.01	0.01	0.00	0.07	0.00	-0.02	**0.36**	1

We have created a correlation matrix of all the compound terms whose **m**utual **i**nformation is greater than one; mutual information is given by Manning and Schutze (1999) as:

Mutual Information

$$(\mathrm{MI}) = \log_2 (f(a.b) / (f(a) \times f(b))$$

Where *f(a), f(b),* is the frequency of occurrence of the words *a, b*. and *f(a,b)* the frequency of occurrence of the compound word *ab.*

In Table 4, we can see a part of the factor matrix were 21 terms are shown as well as the final factor matrix including 7 factors or dimensions that were extracted. Each compound term makes its own contribution to the texts. In factor analysis, the loading of a variable, e.g a compound term on a factor reflects how the variation in the frequency of that compound term correlates with the overall variation of dimension. Indeed, it is considered as a good indicator of how strong or weak is the co-occurrence relationship between a given compound word term and the dimension as a whole; therefore, loadings less than 0.30 are generally considered not interesting for the interpretation of the dimensions. The important and salient loadings (loadings above the threshold) the should be interpreted as part of each dimension; whether negative or positive, which indicates that the sign does not really affect the importance of loading (See Biber,1988). The most frequent compound words that contribute significantly to one of the dimensions are: *BRCA1 mutation(s), ovarian cancer(s)*, *BRCA1 gene(s)*, these terms form the first dimension or (Factor 1), as they have loadings larger than 0.30 on this dimension. Note that *BRCA2 mutation(s)* and *mutation carriers* load also significantly on Dimension 4. However, they have their highest loadings on Dimension 4. While *DNA repair*, *BRCA1 protein* have loadings less than 0.30, so they don't show any significant relationship with Dimension 1, and so on for each of the factors. However, these loadings are not equal; hence, they are not representatives of the dimension. So, in the interpretation of each factor, the focus is on the variables with greatest loadings, regardless of its sign.

The positive and negative loadings show the groups of words that co-occur in the same texts systematically which indicates a specific subject that has been discussed in the text. Note that the compound words: *germline mutations* and *mutations carriers* load significantly on both Dimension Factor 1 and Factor 4, however, their greatest loadings are on Dimension 4: we consider their relationships with Factor 4 as more significant for interpretation with Factor 4. It should be noted however, that they load also on

Factor 1, and perhaps, these two compound words co-occur together with high frequency in texts and in a systematic way and they have a special relationship to each other. For example, when *radiation therapy*, *lymph node(s)*, and *hormone therapy* co-occur in texts, it is more likely to show the absence of *metatstatic breast cancer* where its loading on Factor 4 is negative and that should be taken into consideration.

We may conclude that, the results of the principal components of a total of 30 compound terms show the clear emergence of 7 dimensions. (Table 4).

Table 4. Part of the factor matrix of the analyzed terms, loadings in bold indicate significant relationship between dimension and term.

Term/Dimension	1	2	3	4	5	6	7
BRCA1 mutation(s)	0.77	0	-0.03	0.2	0	-0.03	-0.02
ovarian cancer(s)	0.72	0.01	-0.08	0.1	-0.05	-0.09	-0.09
BRCA1 gene(s)	0.56	0.04	0	0.07	-0.03	0.01	0
DNA damage	0	0.76	-0.03	-0.05	-0.03	0	-0.04
DNA repair	-0.04	0.72	-0.03	0.11	-0.02	-0.03	-0.05
BRCA1 protein	0.09	0.66	-0.01	-0.06	0.02	-0.02	-0.01
endocrine therapy	-0.08	-0.04	0.73	-0.03	-0.07	-0.15	0.28
estrogen receptor(s)	-0.05	-0.02	0.68	-0.03	0	0	0.01
progesterone receptor(s)	-0.03	-0.01	0.59	-0.02	0.21	0.09	0.02
brca2 gene(s)	0.13	0.06	-0.07	0.78	0.01	-0.1	-0.06
germline mutations	0.13	-0.06	-0.02	0.65	-0.03	-0.03	-0.01
brca2 mutation(s)	(0.42)	-0.05	-0.07	0.48	0.04	-0.11	-0.06
mutation carriers	(0.36)	-0.05	0	0.41	-0.06	0.01	-0.01
lobular carcinoma	-0.04	-0.02	0.04	-0.02	0.84	0.04	-0.05
ductal carcinoma	-0.03	0	0	-0.03	0.81	0.07	0.08
radiation therapy	-0.04	0	-0.03	-0.06	0.01	0.67	0.1
lymph node(s)	-0.1	-0.06	-0.02	-0.07	0.1	0.53	-0.14
hormone therapy	-0.05	-0.03	0.04	-0.03	-0.02	0.51	(0.45)
adjuvant tamoxifen	-0.04	-0.02	0.06	-0.03	0.04	-0.05	0.62
adjuvant therapy	-0.04	-0.03	0.07	-0.01	-0.02	0.26	0.58
aromatase inhibitors	-0.06	-0.04	(0.51)	-0.04	-0.11	-0.17	0.53

In order to characterize the texts with respect to each dimension, we computed the dimension value by summing, for each text, the number of occurrences of the compound terms that load saliently on that dimension. For ensuring the experimental independence of dimension values, each compound term was included in the computation only once, thus, each compound term is included in the dimension value of the one on which it has the highest loading.

In more concrete terms, the first dimension that accounts for 6.2% of the shared variance in the data consists of 30 compound terms that were included in the analysis. Just the three terms that have salient loadings (ovarian cancer, BRCA1 mutations and BRCA1 gene(s)), alone account for 4.7% of the total 6.2%, so if we did not include the loadings of these three compound terms in the computation of the total shared variance account for this dimension, then the account for the shared variance will be dramatically reduced to 1.4%. From here, we can see the importance of these three compound terms with respect to this dimension as they represent for 76% of the total shared variance that is accounted for in this dimension. The same applies for Dimension 2 which accounts for 5.4% of the shared variance in the data; the compound terms which have the salient loadings on this dimension are: DNA Damage, DNA repair, BRCA1 protein(s) which account for 5.1% of the shared variance while all the other compound terms account for the rest of 0.03% of the total shared variance for this dimension, and so on

The result of this constraint is that certain terms above the threshold (0.30) will not be included and they have been marked by the parentheses surrounded the value (in Table 4). For example, consider Dimension 1, we sum the frequency of occurrence of *BRCA1 mutation(s)*, *ovarian cancers*, *BRCA1 gene(s)*, for each text, then for each of the three corpora. The dimensions can be

expressed as linear combinations of these compound terms that were included for the computation of dimension values.(Table 5.)

Table 5. The dimensions expressed as linear combinations of the key compound terms

D1	BRCA1 mutation(s)	ovarian cancers	BRCA1 gene(s)	
D2	DNA damage	DNA repair	BRCA1 protein	
D3	endocrine therapy	estrogen receptor(s)	progesterone receptor(s)	
D4	BRCA2 gene(s)	germline mutations	BRCA2 mutation(s)	mutation carriers
D5	lobular carcinoma	ductal carcinoma		
D6	radiation therapy	lymph node(s)	hormone therapy	
D7	adjuvant tamoxifen	adjuvant therapy	aromatase inhibitors	

When we compute the principal components for each of the three corpora, we get a sense of how these un-correlated variables will help us in distinguishing the use of the terms used in the three corpora (Table 6)

Table 6. The values of each of the dimensions for our three corpora.

	Expert	**Professional**	**Patient**	**Identifies**
D1	**1.65**	0.05	-0.54	Experts
D2	**1.24**	0.00	-0.31	Experts
D3	-0.38	**0.28**	-0.29	Professionals
D4	**1.34**	-0.09	-0.38	Experts
D5	-0.28	-0.10	**0.13**	Patients
D6	-0.77	-0.36	**0.60**	Patients
D7	-0.32	**0.86**	-0.08	Professionals

From table 6. We can see the Expert corpus have high positive scores on D1,D2, D4 and negative on D6, where the Patient corpus accounts high on D5 and D6 and negative on D1, similarly for the Professional Corpus as we see in the following:

Experts	D1 (1.65)	D2 (1.24)	D4 (1.34)	D6(-0.77)
Professionals	D3 (0.28)	D7(0.86)		D6(-0.36)
Patients	D5 (0.13)	D6(0.60)		D1(-0.54)

4.4 An initial evaluation

The three dimensions of variation of the expert texts, that is, the terms that explain much of the variance amongst these texts focus on the novel concept of *breast cancer genes* and their mutations, and ovarian cancer (dimensions D1, D2 & D4)– the acronym BRCA is used frequently as an adjective to emphasize the novelty of the concept; the professionals focus on *therapies* of different types and *receptors* for *estrogen* and *progesterone* (D3 & D7); the variance in the texts produced for patients is explained by two different types of *carcinomas* and the *therapies* include *radiation* and *hormone*, and *lymph nodes* (D4 & D5). The dimensions score show strong differences between experts' texts and that written for the patients, with milder differences between professionals and patients.

The experts are focusing on novelty, the professionals are maintaining a balance between novelty and current knowledge, and patients' texts are oriented towards well established practices (*radiation* and *hormone therapy*) and well known after-effects of breast cancer on *lymph nodes*.

5. Conclusions and future work

A method for extracting key terms used in a specialist community of practice was described and factor analysis was used to compute the importance of some of the terms. This method is based on well-established methods in corpus linguistics, terminology, and multivariate analysis. The results show two interesting findings: First, the variance in the Expert corpus is accounted for by 10 compound terms, where the number that accounts for the variance in Professional corpus is 6 and in Patient corpus is 5 (see Table 4). Second, the dimension values show that one can discriminate between the dimensions of variations in Expert corpus where (D1, D2, D4) account for the highest sum, while in Professional corpus (D3, D7) account for the highest sum. Additionally, it was noted that the dimensions that have high positive values for Patient corpus have negative values for Expert corpus.

Our results support how the different parts of the community share some key terms and almost exclusively use others. We have used a more objective criterion for determining which of the terms the different parts of the community prefer.

The similarities and differences indicate the extent of knowledge sharing on the one hand and identify the emergence of new ideas on the other.

We are currently conducting a diachronic study where texts published at different times will be examined. The dimensions of variation across time perhaps will indicate the rate at which 'knowledge' is diffusing. Another strand of our work is to verify the results obtained in the breast cancer study in another domain. Initial work in the domain of *tunneling diodes* – a sub-branch of semiconductor devices and materials- shows we can similarly distinguish between research papers (written by experts) and patent applications (written by legal experts with a working knowledge of the domain).

References

Ahmad, K., and Al-Sayed, R. (2005). "Community of practice and the special language 'Ground'". In (Eds.) S. Clarke, & E. Coakes. *Encyclopedia of Knowledge Management and Community of Practice.* Hershey (PA): The Idea Group Reference. (In Press).

Ahmad, K. (1995). "Pragmatics of specialist terms and terminology management". In (Ed.) Petra Steffens. *Machine Translation and the Lexicon*. Heidelberg: Springer. pp.51-76.(*Lecture Notes on Artificial Intelligence*, Vol. 898).

Biber, D. (1988). "*Variations across speech and writing".* Cambridge: Cambridge University Press

Clark, H. (1996). "*Using language"*. Cambridge: Cambridge University Press.

Davenport, T. and Prusak, L. (1998). "*Working knowledge: how organizations manage what they know"*. Massachusetts: Harvard Business School Press.

Davenport, T., and Probst, G. (2002). "*Knowledge management case book siemens- best practices".* Second edition. Munich: Publicis Corporate Pub., and John Wiley & Sons.

Everitt, B.S. & Dunn, G. (2001). "*Applied multivariate data analysis (2nd Edition)".*London: Arnold.

Gillam, L., & Ahmad, K. (2005) "Pattern mining across domain-specific text collections". In (Eds.) P. Perner & A. Imiya. *International Conference on Machine Learning and Data Mining MLDM´2005*. (Lecture Notes on Artificial Intelligence). Berlin:Springer-Verlag. pp 570-579.

Gillam, L., Tariq, M., & Ahmad, K. (2005). **"**Terminology and the construction of ontology"**.** *Terminology* 11(1), pp55-81.

Manning C.D. and Schutze, H. (1999). "*Foundations of statistical natural language processing".* Cambridge, Massachusetts: The MIT Press.

Nonaka, I. and Takeuchi, H. (1995). "*The knowledge-creating company".* New York and Oxford: Oxford University Press, Inc.

Wenger, E. (1999). "*Communities of practice. learning, meaning and identity"*. Cambridge: Cambridge University Press, pp 72-85.

Wenger, E., McDermot, R., and Synder, W.M. (2002). "*Cultivating communities of practice: a guide for managing knowledge*." Boston: Harvard Business School Press.

