Proceedings of the ACL-2003 Workshop on
Patent Corpus Processing, pp. 46-55.

Can Text Analysis Tell us Something about Technology Progress?

Khurshid Ahmad
Department of Computing
University of Surrey, Guildford,
Surrey. GU2 7XH. UK
k.ahmad@surrey.ac.uk

AbdulMohsen Al-Thubaity
Department of Computing
University of Surrey, Guildford,
Surrey. GU2 7XH. UK
a.althubaity@surrey.ac.uk

Abstract

A corpus-based diachronic analysis of patent documents, based mainly on the morphologically productive use of certain terms can help in tracking the evolution of key developments in a rapidly evolving specialist field. The patent texts were obtained from the US Patent & Trade Marks Office's on-line service and the terms were extracted automatically from the texts. The chosen specialist field was that of fast-switching devices and systems. The method presented draws from literature on biblio- and sciento-metrics, information extraction, corpus linguistics, and on aspects of English morphology. This interdisciplinary framework shows that the evolution of word-formation closely shadows the developments in a field of technology.

Introduction

A patent document is written to persuade a techno-legal authority that the patentee should be allowed to manufacture, sell, or deal in an article to the exclusion of other persons. The article is typically based on an invention that the patentee(s) claim has been theirs. The term *article* is important in that it refers to a tangible object and its usage is to emphasise that *ideas*, intangibles essentially, cannot be patented. Patent documents are the repository of how technology advances and, more importantly, show how language supports the change.

The techno-legal authority requires the patent document to follow a template. This template is divided broadly into two parts: first, legal templates comprising patentee's details, jurisdictional scope, and related item; second, technical templates divided into a summary of the patentee's claims, relation of the article to previously patented articles – the so-called *prior art* – and the scientific/technical basis of the claim. The scientific claim is written in a language that is similar to the language of journal papers.

One important task that is slowly emerging is the extent to which the analysis of a patent document can be automated particularly to assess the overlap between the claims in the document about the article to be patented with that of related, relevant and even counter-claims about the article. The related and relevant claims and counter claims may be found in existing patent documents and may, more indirectly, exist in journal papers.

A patent document has to make references to all other relevant/related articles that have been patented prior to the invention of the article, which is yet to be patented and is the object of the patent document. The references are made primarily by citing the name of the prior art patentees and the titles of their patent documents. A patent document also has other linguistic descriptions of prior art; such descriptions are reminiscent of citations of journal papers in a journal paper. The overlap of a new patent document with a set of existing patent documents may suggest the impact of extant knowledge in patent documents on emerging knowledge in the new patent document. Such an overlap has been studied by the impact of US semiconductor technology on the rest of the world (Appleyard and Kalsow: 1999): this overlap relies largely on the frequency of citation of a US patent by the name of its author or the author's place of work. In computational linguistic (CL) terms this exercise relies on proper noun extraction.

The patent document relates to an explicit and exclusive right over an intellectual property. A journal article relates to an implicit and inclusive right over an intellectual property. The overlap between these two forms of claims is crucial not only in ascertaining the rights of the patentee, or the abuse of the rights of others by the patentee, but also for monitoring the effectiveness of research based on a specialism as a whole or that of its component groups.

The effect of one author or a group of authors working in an institution is indirectly measured by the so-called *impact factor*. This factor relates to the frequency of citation of one or more journal papers written by an author or by a group. The calculation of the impact factor relies mainly on computing the frequency of the authors' name(s) within a corpus of journal articles. Such an impact factor type calculation is used typically in bibliometrics (Garfield 1995). Again, as in intra-patent impact studies mentioned above, in CL terms this is an exercise in proper noun identification and extraction.

The analysis of a patent document, together with the analysis of the related corpora of other patent documents and intellectual property documents, should be based on a framework which provides methods and techniques for analysing the contents of the document and of the corpora. For us the source of a framework still lies in linguistic and language studies. Here we are particularly interested in word formation and terminology usage in highly specialised disciplines particularly those disciplines that deal with intangible articles coupling the word formation and terminology usage with the citation patterns of proper nouns brings us closer to analysing the *contents* of a patent document and its siblings distributed over corpora.

Information scientists usually use the referencing data of research documents to analyse knowledge evolution in scientific fields as well as to identify the key authors, institutes, and journals in specific domains, using tools such as publication counts, citation analysis, co-citation analysis, and co-term analysis to do so. In recent years, patent documents have gained considerable attention as a valuable resource that can be used to analyse technology advances using the same tools.

Gupta and Pangannaya (2000) have applied bibliometric analysis to carbon nanotube patents to measure the growth of activity of carbon nanotube industries and their links with science. They have also used patents data to study the country-wise distribution of patenting activity for the USA, Japan, and other countries. Sector-wise performances of industry, academia and government, and the active players of carbon nanotubes were also studied. They describe the nature of inventions taking place in this particular field of technology, and the authors claim to have identified the emerging research directions, and the active companies and research groups involved.

Meyer (2001) has used citation analysis and co-word analysis of patent documents and scientific literature to explore the interrelationship between nano-science and nano-technology. Meyer investigated patent citation relations at the organizational levels along with geographical locations and affiliations of inventors and authors. The term *co-occurrence* is used by Meyer to find the relationship between the patent documents and the two scientific literature databases SCI and INSPEC. He has noticed that '…the terms that occur frequently in the document titles of all databases are related to […] instrumentalities and/or are located in fields that are generally associated with substantial industrial research activity' (2001:177). Meyer has argued that 'Our data suggests that nano-technology and nano-science are essentially separate and heterogeneous, yet interrelated cumulative structures' (2001:164).

The study of word formation through neologisms within the special language of science and technology has led some authors to argue that it is the scientists as technologists who attempt to rationalise our experience of the world around us in written language by using new words or forms or by relexicalising the existing stock (see Ahmad 2000 for relevant references). Some lexicographers (see for example Quirk et al. 1985) have suggested that neologisms can be formed by two processes: First, the addition or combination of elements such as compounding: *Resonant Tunneling Diodes* and *Scanning tunneling microscopy* are examples for this type of neologism (compounding as a neologism formation is used extensively in science and technology literature); Second, the reduction of elements into abbreviated forms. The abbreviations FET (Field Effect Transistor) and MOSFET (Metallic Oxide Semiconductor FET) are examples of this type.

Neologisms appear to signal the emergence of new concepts or artefacts and the frequency of this new word might indicate the scientific community's acceptance of this new concept or artefact. Effenberger (1995) has argued that '… the faster a subject field is developing, the more novelties are constructed, discovered or created. And these novelties are talked and written about. In order to make this technical communication as efficient as possible, provision should be made for avoiding misunderstanding. One crucial point in this process is the vocabulary that is being used' (1995:131, emphasis added).

In this paper we discuss the idiosyncratic language used in patent documents. The language is replete with terms and there are instances within a patent document that suggest that the authors not only use the specialist terms but use a *local syntax* as well. We look specifically at the structure of the US Patents and suggest how with existing techniques used in information extraction and NLP, including term extraction and proper noun identification, one can perform fairly complex tasks in patent analysis – some of which are performed by patent experts by hand currently (Section 2). This examination suggests to us a model of development in computer and semi-conductor technology: an incremental model where each subsequent patent helps in the development of ever-complex artifacts – starting from devices onto circuits and onto systems. We will look at one of the key inventions in the field of semiconductors physics – the *electron tunneling device*. These devices combine technical elegance, experimental complexity and manufacturing challenge. Due to its strategic importance, a number of patents have been obtained by the US government and also by a number of US and Japanese companies (Section 3). Section 4 concludes this paper.

The Structure of US PTO Documents and a Local Grammar for the Documents

The USPTO database is a representative sample of patent documents. The USPTO has documents related to most branches of science and technology. It includes information about all US patent documents since the first patent issued in 1970 to the most recent. The USPTO database allows the user to search the full text of the patent documents for a certain word or a combination of words. It also provides a field search for specific information such as *inventor* or *assignee*. The search can also be conducted for a specific year or range of years. The US Patents are written partly as a legal text and partly as a scientific document. Over the last 50 years or so, it appears that US Patent documents have been structured in terms of layout and have a superficial resemblance to Marvin Minsky's *frame*-like knowledge representation schema.

The patent document can be divided into three main parts for the present discussion: The first part comprises the biographical details of the inventors (and their employers) together with the title of the invention and a brief free-text abstract, dates when the patent was applied for and when the patent was granted and so on. The free text is essentially a summary of the claims of the patentee; The second part contains external references of three sorts: the first sort is the specialist domain of the invention – the subject class indicating the super-ordinate class and instances; the second sort are other cited patents organised as a 4-tuple: (i) patent number, (ii) date of approval, (iii) first inventor and (iv) classification number; and, the third sort is a bibliographic reference to publications that may have contributed to the patent; The third part of a current US Patent document comprises 'claims' related to the patent and the description of the 'invention' (there are diagrams of the invention attached to the document and the diagrams described in the text). Table 1 on the next page shows the template of the current (c. 1980 and after) USPTO's.

The 'claims' of the patentees are clearly itemised and initialised by the number of the claim; the first claim is the basis of the patent abstract generally. The 'background to the invention' is written in an idiosyncratic fashion as well – the invention is first contextualised in a broader group of other inventions to date and then the specific nature of the invention is exemplified. The broader and the specific are usually marked by phrases like 'The (present) invention relates to' and the specificity is phrased as '(more) specifically.' or '(more) particularly'. These phrases are followed by one or more noun phrases connected with, for example, conjunctions or qualifiers. The first noun phrase names the article invented, for instance, a name of a new device, circuit or a fabricating or testing process.

FIELD	VALUE
United States Patent Number	NUMBER
First Inventor	PROPER NOUN ET AL.
Date Patent Approved	DATE
Title:	FREE TEXT
Abstract:	FREE TEXT
Inventors:	PROPER NOUNS
Assignee:	PROPER NOUNS
Application No.:	NUMBER
Filed:	DATE
Patent Classification Data:	NUMBER
References Cited [Referenced By]:	[PATENT NUMBER, DATE, FIRST INVENTOR, CLASS NO.]
Parent Case Text: CROSS REFERENCE TO RELATED APPLICATION	FREE TEXT
Claims:	**'What is claimed is: '**
CLAIM 1: CLAIM 2:	FORMULAIC FREE TEXT FORMULAIC FREE TEXT
Description	
BACKGROUND OF THE INVENTION	
1. Field of the Invention:	FORMULAIC FREE TEXT
2. Related Background Art:	FORMULAIC FREE TEXT
SUMMARY OF THE INVENTION:	SEMI FORMULAIC FREE TEXT
BRIEF DESCRIPTION OF THE DRAWINGS:	FREE TEXT
DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS:	FREE TEXT.

Table 1: A slot-filler template of the US PTO approved patent documents.

The NP comprises determiners and modal verbs together with (compound) nouns. The first NP is optionally followed by a qualification that restricts or extends the scope of the discovery – the enlargement or restriction is named and another NP is used for the naming and so on. This simple grammar can be verified by examining a corpus of patent documents. To illustrate this point we have looked at a recent randomly selected patent on memory devices – a patent filed by *Kabushiki Kaisha Toshiba* of Japan (or Toshiba for short), and approved by USPTO on 20th May 2003, on a semiconductor memory device which uses the emergent notion of *memory cells* (a memory cell is a tiny area within the memory array that actually stores the bit in the form of an electrical charge [1]). An analysis of the title and that of the 'Background of the Invention: Field of Invention' fields shows the use of this restricted syntax (Table 2). In much the same as the 'claims' and 'the 'background', the 'summary of the invention' is also phrased in a formulaic manner (see Table 1 for the structure of the patent document).

The analysis of the other slots governed by a simpler grammar yields interesting results and suggests that the names of assignees and the manner in which patents are being cited can be easily inter-related (Table 3). Toshiba's USPTO 6567330 refers to **8** other patents. The details of the referenced patents are in a 4-tuple, which can be unambiguously interpreted. Each of the referenced patents refers to about 10 patents in turn. An examination of 82 such patents may help to initiate, perhaps, a discussion of the 'invention life cycle' or 'licensing potential of a patent' (Mogee 1997), or even a discussion of 'micro foundations of innovation systems' (Andersen 2000).

[1] Definition form http://rel.intersil.com/docs/lexicon/M.html, site visited 29 May 2003)

Title of the Patent	US PTO Number	Field of Invention	
Semiconductor memory device	6567330	The present invention relates to a *semi-conductor memory device* with a current-read-type memory cell […]	More specifically, the present invention relates to **a data sense circuit** for the semiconductor memory device.
Patents cited by USPTO 6567330			
Nonvolatile semiconductor memory device	6407946	The present invention generally relates to a *nonvolatile semiconductor memory device*,	and more particularly relates to an **electrically erasable and programmable read only memory**
Semiconductor memory device	6337825	This invention relates to a *semiconductor memory device,*	and more particularly to a **sense amplifier** of a nonvolatile semiconductor memory using **current read-out type memory cells.**
Memory cell sense amplifier	6219290	The present invention relates to **memory arrays,**	and in particular, the **sensing of data** from a **non-volatile memory cell**.
Current conveyor and method for readout of MTJ memories	6205073	This invention relates to *M[agnetic] T[unneling] J[unction] memories*	and more particularly, to **apparatus and a method for reading data stored** in MTJ memories.
Read reference scheme for flash memory	6038169	This invention relates to ***flash memory***	and in particular to creating a **reference** by which to **read the state of flash memory cells**.
Sensing circuit for a floating gate memory device having multiple levels of storage in a cell	5910914	The present invention relates to a *sensing circuit* for use with a memory array comprised of floating gate devices, [..].	More particularly, the present invention relates to the use of a **plurality of inverters** to **compare the current from a reference cell** […]
Flash memory device having a page mode of operation	5742543	The present invention relates generally to *memory devices*	and more particularly to a **nonvolatile memory device** having a **page mode of operation**.
Single cell reference scheme for flash memory sensing and program state verification	5386388	The invention relates to *the field of metal-oxide semiconductor (MOS) [..]EPROMs [..]*	particularly to the field of **"flash" EPROMs [..]**

Table 2: The use of restricted syntax in the d escription of the generic and specific fields of inventi on. The higher patent number shows that it was filed at a later date than a lower patent number. So, the above figure shows a time o rder as well.

Assignee	Country	Patent Number	USPTO Class	Approval Date (a)	Earliest Reference (b)	Latest Reference (c)	Invention Cycle Time' (a) – (c)	Invention Cycle Time'' (b) – (c)
Toshiba	Japan	6567330	365/210	May-03	Jan-95	Jun-02	1.0	6.5
	Patents	**cited**	**by**	**USPTO**	**Number**	**6567300**		
Matshushita	Japan	6407946	365/185	Jun-02	Jun-93	Nov-99	2.5	6.3
Toshiba	Japan	6337825	365/185	Jan-02	Nov-92	Aug-00	1.5	7.3
Macronix	Taiwan	6219290	365/185	Apr-01	Aug-93	May-98	3.0	4.8
Motorola	US	6205073	365/171	Mar-01	Jun-98	Aug-00	0.5	2.1
Halo LSI	US	6038169	365/180	Mar-00	Dec-92	Aug-99	0.8	6.8
Silicon Storage	US	5910914	365/185	Jun-99	Sep-80	Jun-97	2.0	17.0
Intel	US	5742543	365/185	Apr-98	Nov-96	May-80	1.5	19.5
Intel	US	5386388	365/185	Jan-95	May-72	Dec-92	2	19.5

Table 3: A glimpse of the technology transfer in the Toshiba patent for 'data sensing circuits' for sem iconductor memory de vices. The US Patent Classification ***365*** refers to 'Static Information Storage and Retrival, and the subclassifcations ***185*** & ***171*** refer to 'Floating Gate Memories' & 'Magnetic Thin Films'

A finer grained analysis to show which 'country' is more influential can also be performed fairly readily and indicates the extent to which patents that are held by assignees domiciled in the USA have over half the cited patents (Table 4).

Assignee Country	#	%	Assignee Country	#	%
US	45	54.9%	Korea	2	2.4%
Japan	18	22.0%	France	1	
Independent	7	8.5%	Germany,	1	1.2%
Italy	5	6.1%	UK	1	
Taiwan	2	2.4%	TOTAL	**82**	**100**

Table 4: An analysis of USPTO No. 6567330 (Toshiba Japan) shows the major influence of US-based assignees, followed by Japan. A significant number of patents (8.5%) are held by individuals and not assigned specifically to a country.

A semi-automatic analysis of terms used in the Abstracts and Titles of the patents (Toshiba 6567330 and patents referenced in the Toshiba patents) shows the co-citation pattern of terms. This may help in the clustering of patents on the basis of terms extracted from the patent documents as well as novel terms (terms not included in the USPTO Patent Classification terminology data base) found in the document. We show the co-citation of the two key terms *memory cell* and *memory device* in the nine patents discussed above. The use of the two terms individually and as roots and stems of other compounds is also shown. The more frequent citation is to the newer term *memory cell* and it is cited in all but one of the 9 related patents. The related *memory devices* – newer devices now incorporate *memory cells* – is less frequently used and it is only found in the abstracts of 5 out of the 9 patents. Both terms are co-cited in 6 out of the 9 patents (see Table 5 for details).

The interrelationship between the different patents can be explored further by examining closely as to what is being patented within the patent and what is being patented in the referenced patents. Again, we use the example of the Toshiba patent No. 6567330 which refers to 8 other patents. The patent itself relates to the invention of a *system*. The referred patents relate to other *systems* and *circuits*. Let us look at the earliest patent cited in Toshiba's patent: this is US PTO No. 5386388 filed by Intel Corporation (USA) approved in January 1995. The title of Intel's patent is '*Single cell reference scheme for flash memory sensing and program state verification*'. Flash memory is defined as 'A nonvolatile programmable semiconductor memory product [2]. This patent relates to the invention of a circuit. Intel's patent comprises references to another 15 patents: 5 refer to other systems, 8 to circuits, and one each to a device and a software program (see Figure 1 on the next page). The information whether a patent is related to any of the four classes can be gleaned from the Patent Classification Number. Further analysis of the referenced patents shows a similar pattern – references to circuits, devices, systems and software. This appears to be a basis of the inventions within the semiconductor industry, especially those related to the development of computer systems based on these systems, devices and circuits. This is the basis of our more speculative investigations related to the *resonant tunneling systems.*

Patent No.	**Freq.**	**Compound Term**	**Freq.**	**Compound Term**
	Memory	**Cell (m.c.)**	**Memory**	**Device (m.d.)**
6567330	*4*		*3*	*semicond.* +m.d.(3)
6407946	*2*	m.c. +*transistor(2)*	*1*	*non-volatile semicond.* +m.d.(1)
6337825			*2*	*semicond.* +m.d.(2)
6219290	*3*	m.c. +*sense amplifier (1)*		
6205073				
6038169	*3*	*flash* +m.c. (1); m.c. *current (2)*		
5910914	*2*		*2*	*Floating gate* + m.d. (2)
5742543	*3*		*1*	*flash* +m.d. (1)
5386388	*1*			
Total	18		9	

Table 5. Distribution of the two co-cited terms in the nine patents. The frequency of the compound terms is included in the frequency count.

[2]From http://www.micron.com/, site visited 29 May 2003

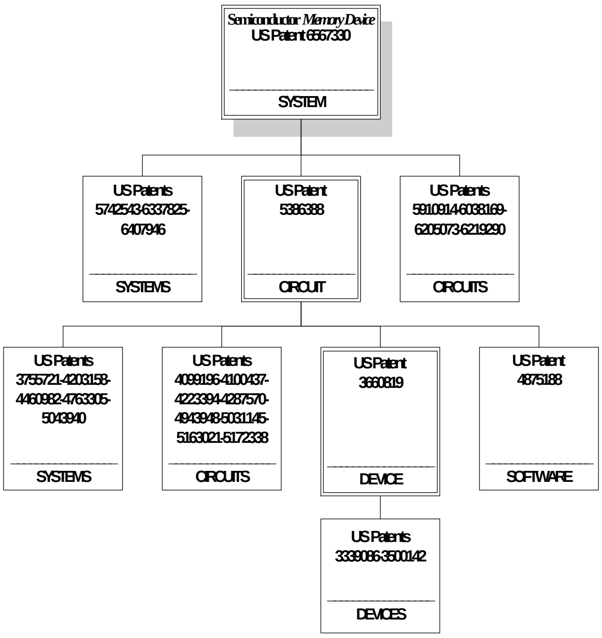

Figure 1: A hierarchical citation -based ordering of patents and the distribution of patents into three categories – *systems*, *circuits* and *devices*.

3 The Evolution of the Resonant Tunneling Devices

We will now focus on how terminology u sage may help in tracking the evolution of *resonant tunnel-ing devices*. These are ultra high-speed devices, which perhaps will be used in the compute rs of the next decade or so. In order to study how one can track technology progress we have adopted an i n-tuitive, but realistic, framework. For us, all co m-plex systems comprise subsystems and subsystems are made up of much smaller (and simpler) *de-vices*. A computer system is made up of i ntegrated circuits and the circuits made up of transistors and transistors come in di fferent types. One model of growth can be thought of as follows: First, devices are patented, then su bsystems, and finally the complex systems (remember only tangible articles can be patented). So fo llowing this intuitive framework we will first see a number of devices being patented then subsystems and finally the sy s-tems themselves. Tunnel diodes are supposed to empower faster switching devices, which in turn have to be incorporated into subsystems with tu n-neling transistors and into complex systems with circuits. Our hypothesis is that an analysis of a diachronically organized text corpus will show the working of the above-mentioned framework.

A corpus was built containing more than 2.2 million words of patent documents. The co r-pus contains all patent documents that co ntain the term *tunneling* in the title. USPTO search r esults showed that there are 372 titles, approved from 1975 to 1999 in semiconductor physics. We have analysed frequency of compound word in the USPTO patent documents published b etween 1975-1999 (Table 6).

	75-79	80-84	85-89	90-94	95-99
No.of Texts	7	8	68	133	156
Total No. of tokens	43812	43262	378272	771525	995894

Table 6. The diachronic breakdown of patents comprising at least one instance of the token *tun-neling* over 5 year intervals between 1975 -1999.

The compound word analysis was co n-ducted using System Quirk and no compounds were pre specified (System Quirk a text analysis system, is available on www.computing.surrey.ac.uk/ai/SystemQ). The system extracts compound words based on a si m-ple heuristic: a set of word that does not co ntain closed class words (i.e. determiners, conjun ctions, prepositions, and moderators) or the orthographic signs (including pun ctuation, numbers, currency and other symbols) is considered by Sy stem Quirk to be a compound word (see Ahmad and Rogers, 2001, for details). The va lidation of compound words can also be carried out by statistical tests, for instance described by Smajda (1994).

To investigate the progress of resonant tunneling devices and circuits, the multi-word terms were extracted from the USPTO full text corpus using System Quirk. The extracted terms that relate to resonant tu nneling diodes, resonant tunneling transistors and resonant tunneling ci r-cuits were arranged in a five year interval starting from the first emergence of the term *resonant tun-neling* in USPTO abstract documents in 1985.

Tracking the frequency usage of the terms associated with resonant tunneling artefacts in the USPTO full text corpus shows a considerable i ncrease of frequency usage interval by interval. The frequency of the term *resonant tunneling diode* (and its plural form *resonant tunneling diode**s***, both denoted as the lemma *resonant tunneling diode~* subsequently) increased significantly from 45 in 1985-1989 to 446 in 1990-1994 by about a factor of 19 and then in the next time interval 1995-1999 the frequency dropped by about half to 240. The frequency usage of the term *resonant tunneling transistor~* in the USPTO full text corpus increased from 23 in the period 1985-1989 by about a factor of 10 to 225 in 1990-1994. The increase of frequency usage of the term in the time period 1995-1999 increased by a factor of 1.3 to become 293. The term *resonant tunneling circuit~* appears in the USPTO full text corpus 45 times in the time interval 1990-1994. Frequency usage of resonant tunneling circuits increased by a factor of 1.3 in the next interval (1995-1999) to 57.

Word formation is not restricted to the inflection of a compound word. Rather, we see further instances of compounding where an existing compound, say, *resonant tunneling diode/transistor* is used as a head of other compounds (Table 7).

1990-1994	1995-1999
barrier resonant tunneling diode	triple *barrier resonant tunneling diode*
band resonant tunneling transistor~	bipolar quantum *resonant tunneling transistor*

Table 7. The specialization, through prefixation, of the term *resonant tunneling diode & transistor* over a 10 year period in our patent corpus

We note the very productive use of compounding and inflection in our corpus. Note, however, that the size of the corpus for the three different periods, 1985-89, 90-94 and 95-99, are different: 378272, 771525 and 995894 respectively. The size of the corpus perhaps for the later two periods is roughly the same but the earlier corpus (85-89) is three times smaller. In order to present a better comparison we will look at the relative frequency of the compounds in that we will sum up the frequency of all the extracted compounds related to *resonant tunneling* diodes, transistors and circuits, as per our intuitive framework, and assign relative frequency to each of the three relative to the sum.

Consider the result of analysis of 133 texts of patents published in 1990-1994 for tunnel diode related patents. The total number of terms comprising the lemma *resonant tunneling diode~* is 490, which includes the lemma on its own and two terms containing the lemma as the headword; these are *multiple peak resonant tunneling diode*, *barrier resonant tunneling diode*. The total containing the lemma *resonant tunneling transistor* is 225, which is made up of 188 for the lemma on its own and the rest for the two other terms. The lemma *RT circuit* also includes hyponyms of the term, e.g. *RT oscillator (circuit)*, *RT logic gate (circuit)* and *RT memory (circuit)*; note that the term *circuit* is shown in parentheses as it is ellipsed in the text – the reader of the patents, an expert in the discipline, is expected to know that an *oscillator* is a *circuit*. The two terms occur 24 and 12 times together with 4 other terms that collectively occur 9 times making a total of 45. The three lemmas *RT diode, transistor* and *circuit* occur for a total of 490 + 225 + 45 (= 760) times, hence the relative frequency of the three lemmas is 64.4% (490/760), 29.6% (225/760) and 6% (45/760) respectively (Table 8 shows a breakdown of the distribution).

This relative frequency computation was conducted over the periods 1985-1989 and 1995-1999. Table 9 (on the next page) shows that over 64% of the terms belong to the lemma *resonant tunneling diode~*, about 30% to *resonant tunneling transistor* and just about 6% to *resonant tunneling circuit ~*. This situation changes quite dramatically in the next quinquennium (1995-1999).

Artefact	1990-1994	Freq	%
Resonant tunneling diodes	resonant tunneling diode~	446	
	multiple peak resonant tunneling diode	24	
	barrier resonant tunneling diode	20	
Total		490	64.4%
Resonant tunneling transistors	resonant tunneling transistor~	188	
	band resonant tunneling transistor~	35	
	bipolar quantum resonant tunneling transistor~	2	
Total		225	29.6%
Resonant tunneling 'Circuit~'	resonant tunneling oscillator~	24	
	resonant tunneling logic gate~	12	
	resonant tunneling diode memory	3	
	resonant tunneling diode oscillator	3	
	multiple resonant tunneling circuits	2	
	resonant tunneling photodetector	1	
Total		45	6%

Table 8: *Resonant tunneling* artefacts in the USPTO full text corpus in the time period 1990 - 1994.

Compound term	**Period**		
	85-89	90-94	95-99
RT diode~	66.2%	64.4%	41.2%
RT transistor~	33.8%	29.6%	49.1%
RT circuit~	0	6%	9.7%
Total	100%	100%	100%

Table 9. The growth of compound terms comprising the headwords *diode & diodes* denoted collectively as *diode~*, *transistor~*, and *circuit~*, together with the stem *resonant tunneling (RT).*

4 Afterword

It appears that there is a *local grammar*, comprising vocabulary of the specialist domain and a syntax that appears different from the general (universal?) syntax, used in framing the claims, background and summary of the invention in a US Patent document. A number of slots in the US PTO document are reserved for proper names – patentees, assignees, places of work, and other slots hold dates and all these slots show the extremes of the local grammar – essentially a grammar for a one-word language. The document comprises 'references to (other patents) and also citations to an extant by other later patents – this information is encoded in another local grammar of one or more 4-tuples referring to a referenced patent – the 4-tuple has a clearly defined sequence and allows expressions only in terms of four noun-phrases. The referenced patent number is an active hyperlink through which the details of the referenced patent can be accessed and subsequently a chain of references can be established in a (semi-) automatic manner. The existence of a local grammar and the hyperlinks suggests to us that one can create a historic (diachronic) description of an invention together with the crucial account of the influence of other inventions.

Restricted syntax is used, for example, in describing time (hours, minutes, seconds, days, years, months), in financial news wire as well as mission-critical communication. The specialist vocabulary, and more so the productive use of the vocabulary (see below for details), as well as the restricted syntax emerges initially for assuring ambiguity-free communication in an inherent noisy medium of communication – natural language.

Complementary to the emergence of the present US patent document, there has been an accumulation of terminological knowledge in terms of the repositories usually referred to as *patent classification*. The Patent Offices around the world classify all manners of 'articles' ranging from micro-electronics to kitchen utensils and from software systems to heavy excavation machinery, for example. Much like a number of other utilitarian classification systems, including the Dewey Decimal Classification on the one hand and the US National Library of Medicine's Disease Classification system on the other, the US PTO classification system is detailed, complex, full of cross references, and occasionally confusing. The fact remains, however, that like all utilitarian systems, the US PTO classification system is a rich repository that can be used, with some alterations, as the lexical/terminological resource for information extraction in particular and NLP in general. The repository states the ontological commitment of the US PTO and its advisers, and can be used for building knowledge representation schema or semantic processing systems.

The appearance of a local grammar, or perhaps local grammars, used to frame a patent document together with an extensive terminology database of patent classification, is good news for the patent processing community. There is some hope that the information extraction and NLP systems will be able to extract the terminology and identify the idiosyncratic syntax that governs the different parts of the patent document with the help of techniques pioneered in corpus linguistics. Terminology extraction can be facilitated by referring to the patent classification terminology base and facilitated by various statistical and linguistic techniques used to identify complex noun-phrases in specialist texts. Once the local grammar is identified it will be able to meaningfully process the documents for inferring the import of a given invention in relation to other inventions and to assess the impact of journal publications of inventions. And, indeed all manner of new ways of examining a patent document may open up once the investigator overcomes the burden of sifting through an overgrowing lexical mountain of new patents, revisions to existing patents and the scientific and technical publication juggernaut that adds more to the mountain on almost daily basis. The automatic extraction of compounds from a corpus of patent documents appears to show the introduction of new artifacts through the use of morphological processes like word formations. Currently, our work in progress is to 'chart' a transfer of such terms in journal papers onto patents, in addition to the exercise reported which charts the transfer of terms within a diachronically organised corpus of patent documents.

References

Ahmad, K. 2000. Neologisms, Nonces and Word Formation. *Proceedings of the Ninth EURALEX International Congress* (Munich August 2000).pp 711-729.

Ahmad, K. and Rogers, M. 2001. Corpus Linguistics and Terminology Extraction. *Handbook of Terminology Management*. Amsterdam: John Benjamins Publishing Co. pp725-760.

Andersen, B. (2000). *Technological change and the evolution of corporate patenting: The structure of patenting 1890-1990.* Cheltenham: Edward Elgar.

Appleyard, M.M. and G.A. Kalsow. 1999. 'Knowledge diffusion in semiconductor industry'. *Journal of Knowledge Management*. Volume 3 (No. 4). pp 288-295.

Effenberger, D. 1995. Fundamentals of Terminology Work. *Computer Standards & Interfaces*, Vol. 17, 131-137.

Garfield, E.1995 The Impact of Cumulative Impact Factors. *Proceedings of the 8th IFSE Conference, Barcelona,* pp58-81.

Gupta, V. K. and Pangannaya, N. B. 2000. Carbon Nanotubes: Bibliometric Analysis of Patents. *World Patent Information* 22: 185-189.

Meyer, M. 2001. Patent Citation Analysis in a Novel Field of Technology: An Exploration of Nano-Science and Nano-Technology. *Scientometrics* 51.1:163-183.

Mogee, Mary E. (1997). 'Patent Analysis Methods in Support of Licensing'. Paper presented at the *Technology Transfer Society Annual Conference (Denver, USA).* (http://www.mogee.com/services/tl-methods.html, site visited 20 May 2003).

Quirk, R, S Greenbaum, G Leech, J Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London and New York: Longman

Smajda, F. 1994. 'Retrieving Collocations from Text: Xtract.'. In (Ed.) Susan Armstropng, Using Large Corpora Cambridge, MA/London/England: MIT Press. pp 143-177.