Extracting Complexity Metrics of Technological Artifacts and Systems from Patents using Patent Document Structure

D. Nel

orcid.org 0000-0001-6939-1503

Dissertation submitted in fulfilment of the requirements for the degree *Master of Engineering in Electrical and Electronic Engineering* at the North-West University

Supervisor: Mr. A.J. Alberts

Graduation October 2018

Student number: 21172587

To my wife, Sarina -

Without you this study would have been completed a year ago,

but never attempted in the first place.

I dedicate this work to you.

Completed in the spirit of these words -

"Do not merely practice your art, but force your way into its secrets;

It deserves that, for only art and science can exalt man to divinity."

- Beethoven, 1812

Abstract

Keywords: Patent; Complexity; Complex Systems; Innovation; Technology Trends

This study investigates the relationship between different complexity metrics from the literature and metadata fields in patents. Patent citations, claims and other fields are compared against complexity metrics derived from the number of parts and part interactions, as extracted from over 100 000 patents. A method is proposed to extract part descriptions from patents by leveraging the XML structure of the patent publication. The part names are used in a normalisation process to gain an accurate count of unique parts in a patent. A commonly used metric for complexity, specifically derived from patents, is analysed and tested against complexity measures derived from the amount of parts and interactions in a patent. The study underlines several shortcomings in the measurement of complexity in patents, especially the literature's propensity to use sub-class classifications from patent classification hierarchies as proxies for the complexity of different technology types. Furthermore, it is demonstrated that derivations from data fields of patents, including citation and claim counts, are positively correlated with the complexity of the patented invention.

Opsomming

Sleutelwoorde: Patente; Kompleksiteit; Komplekse Sisteme; Innovasie; Tegnologiese Tendense

In hiedie studie word die ooreenkoms tussen maatstawwe van tegniese kompleksiteit vergelyk met datavelde verkry vanuit patente. Patent bronverwysings, patenteregeise en ander datavelde word vergelyk met kompleksiteitsmaatstwe afgelei vanuit die hoeveelheid parte en wisselwerking tussen parte. Die data is ontgin uit ongeveer 100 000 patente. ‘n Metode om partbeskrywings te verkry vanuit die XML-struture in patentdokumente word beskryf. Partname word gebruik in ‘n kontroleproses om unieke parte te verkry vanuit patentbeskrywings. ‘n Maatsaaf van kompleksiteit word afgelei vanuit die hoeveelheid parte en wisselwerkinge tussen parte in patente. Dié word dan vergelyk met ‘n kompleksitieitsmaatstaaf wat in algemene gebruik is in die literatuur en ook self uit patendata gegenereer word. Die studie wys verskeie swakshede uit in die metodes wat gebruik word om kompleksiteit te meet. Hier word veral klem gelê op die algemene gebruik van klassifikasiekodes om verskillende tegnologieë voor te stel en die gevare hiervan. Daar word ook getoon dat dat daar ‘n positiewe verwantskap bestaan tussen die kompleksiteit van ‘n gepatenteerde innovasie en die datavelde in die patentdokumentasie.

Acknowledgements

I would firstly like to thank my wife for her unending patience while I was chained to my desk for the duration of this project. Her support was pivotal.

I would like to thank my study leader, Andreas Alberts, for his guidance, time and insights.

Lastly, I would like to thank the NWU THRIP program for financial contributions during the first phase of this study.

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 OVERVIEW AND CONTEXT

1.1.1 A SHORT HISTORY OF INTELECTUAL PROPERTY

The idea of a patent, or similar notions to protect an inventor's intellectual property, is quite old. One of the earliest examples of intellectual property rights was penned by the Greek compiler Athenaeus in the third century A.D. According to his *Deipnosophistae* cooks who created new dishes were granted executive rights to prepare it for one year. Several examples of industry privilege were documented between the third century and the advent of the Middle Ages, but it was not until the 15th century that the first real patent appeared. In 1421 the state of Florence, Italy granted the architect Filippo Brunelleschi the executive right to use a device of his invention to transport heavy loads on rivers. In the spirit of medieval thinking, it was stipulated that any imitation of his work should be burned. After 1450 the granting of patents became systematic in Venice. One of the main Venetian industries of the time was glassmaking. The trade secrets of the industry were fervently guarded, to the point where practicing glassmaking abroad was punishable by death. Many Venetian artisans did however defect to other parts of Europe and sought the same monopoly that was granted them under Venetian law. In this way the idea of a patent disseminated throughout Europe. Patenting became more nuanced and at the dawn of the 16th century, Venice was granting "registered designs". An excellent example of this is a patent granted to a Venetian printer named Aldus Manutius in 1501 for his design of a new slanted printing typeset. To this day it is known as *italic*. [1]

Around 1555, king Henry II of France introduced the concept that a patent should fully disclose an invention, so that others may benefit from it after its protection period has ended. This remains a pivotal principal in modern patent law. The Canadian Patent act of 1985 encapsulates this principle well – "*The specification of an invention must ... set out clearly the ... method of constructing ... of a machine, ... in such full, clear, concise and exact terms as to enable any person skilled in the art or science to which it pertains, or with which it is most closely connected, to make, construct, compound or use it.*" [2] [1]

In 17th century England, the monarch bestowed industry monopolies on courtiers as rewards. This created industry instability and in 1602 Francis Bacon started lobbying the House of Commons to pass into law that only new inventions may be given a market monopoly. The struggle between the crown and parliament lasted over 20 years. It was only in 1624 that the

Statute of Monopolies was enacted and the crown's power to bestow monopolies ceased. The principal of patenting only novel inventions is also pivotal in modern patent law. [1]

On the 31st of July 1790 the United States Patent and Trademark Office issued its first ever patent to a Mr. Samuel Hopkins for his process of making potash, an ingredient in fertilizer. The granting certificate was signed by George Washington [3].

From these early steps in identifying and protecting ideas grew the notion of modern patenting. The system of protecting ideas has become more nuanced, but the basic principles remain unchanged. In summary, patented technologies should be novel and adequately disclosed, and only then will the patentee receive the reward of having a temporary monopoly on the monetisation of the idea.

A side effect of needing to protect ideas was that humanity started keeping a very accurate and descriptive record on how technology was developed and utilised. This study aims to investigate aspects of this record, how the technology recorded therein changes and integrates to become more complex.

1.1.2 PATENTS AS A DATA SOURCE

From its humble beginnings patenting has grown into a massive global phenomenon where over 1.2 million patents are granted annually [4]. World Bank data shows that 998,572 patents were filed in 2006, and the application rate steadily rose to 1,624,969 patents in 2013 [5]. This translates into an average of just over 100 000 more patens being filed every consecutive year.

The primary purpose of patents has always been to provide the patentee with a temporary monopoly in the market, but more recently researchers started using it for other purposes. It is not hard to see why – the global corpus of patent documentation is a mostly standardised library of human inventiveness, and is also the most complete publicly available description of many inventions. It is estimated that about 80% of all invention related technical information can only be found in patents [6].

Economics and management science studies are some of the main consumers of patent information. R&D investment, the resulting intangible capital and its effect on market valuation of the firm is one example of an economic question that requires patent information to answer [7]. Patent citation information has also been used to study knowledge spillovers and as a method of information dissemination in the innovation process [8]. The patent system is, at

least partially, a policy tool designed to incentivise firms to invest in R&D. A whole stream of research on the interaction between law, economics and industrial productivity makes use of patent information [9].

Patent information can also be used to help describe the process of innovation. Patented inventions are fully disclosed and cite previous links in the chain of technological development. It is therefore a good source to use when developing or validating a hypothesis. This bibliometric approach can help craft general trends in the innovation process. Consider, for example, Moore's Law [10]. It observes that the number of transistors in an integrated circuit doubles approximately every two years and the cost halves. Note that this observation elegantly combines the nature of change in a technological artefact with external factors such as the cost of production.

This study aims to elaborate on this theme – where data is extracted from patents and applied to highlight aspects of the innovative process.

1.1.3 JUSTIFICATION

Several good reasons exist to study technological complexity. This includes how this complexity is encoded in patent documents. Reasons for studying technological complexity and change include:

1. Future labour force requirements
   One of the side effects of the industrial revolution, or any large technological change, is that certain professions become redundant and new professions and skill sets are created to manage and utilise the superseding technology. Insight into trends of change is therefore very applicable in any form of labour planning or labour related economic forecast.
2. Market advantage and relevance
   A continual assessment of current and possible future changes in sector specific technologies is required for a firm to remain competitive. IBM and Kodak are some modern examples of this. Both of these companies failed to innovate and adopt new innovations and both are now mere shadows of their former glory. IBM thought home computers were a craze that will pass and Kodak could not imagine that digital photography could replace “the warmth and grain of good old film”.

3. Building a theoretical framework

   Chapters 2 and 3 deal with the theory of complexity and complex systems. It will be shown there that there is little consensus on the nature and definition of complexity. A more complete theory of complexity will be able to add value to almost all fields of study and business. For example, if a company has a firm grasp of the effects of complexity on the production requirements of some product it will be much easier to optimise the production cost and process.

1.2 PROBLEM BACKGROUND

1.2.1 THE FLEMING-SORENSON STUDY [11]

This research problem was loosely derived from a prominent study by Fleming and Sorenson [11]. Their study develops a theory of innovation based on complex adaptive systems theory and elements from work done in evolutionary biology.

They frame technological evolution as a recombination of new or existing components. In this analogy the parts of a technological artefact are analogous to the set of genes that that drive natural evolution.

Parts/genes are conceptualised by making use of a landscape. In this view a landscape is built up from a unique set of genes or components. Every point on such a landscape corresponds to a particular configuration of the components making up the landscape. The height of a particular point presents the fitness of a particular configuration of parts. In the analogy, fitness is equivalent to the usefulness of an invention.

Two variables are used to map the topography of the landscape, and thus the usefulness of the invention. These are the number of components (N) and their interdependence (K). Components are defined as any part or constituent technology that are recombined by the inventor. Interdependence is defined as the functional sensitivity of the invention to changes in its constituent parts. This is illustrated by the dopant concentrations used in the production of semiconductors. If the concentration of the dopant in a silicone-based semiconductor changes by one part in 10^8, its resistance can fluctuate by a factor of 24100 [12]. This is an extreme case, but it does demonstrate the idea of interdependence well.

To produce this technology landscape, data from ~17 000 patents are used. The value of an invention is measured by counting the amount of citations that a patent receives within a set period after publication.

A metric is constructed form patent classification codes to measure interdependence. These codes consist a hierarchal classification structure. All patents have at least one classification. Interdependence is measured as -

$$Interdependence\ patent\ l\ \equiv K_l = \frac{Count\ sub-classes\ on\ patent\ l}{\sum_{l \in i} \frac{sub\ classes\ previously\ combined\ with\ sub-class\ i}{patents\ in\ sub-class\ i}}$$

The number of components (N) are measured by counting the number of sub-classes on a patent.

1.2.2 SHORTCOMINGS OF THE FLEMING-SORENSON STUDY [11]

Several assumptions are prevalent in the choice of proxies for the number of parts and their interdependence. It should be noted that the critique of Fleming and Sorenson's methodology is presented speculative in this section. The critique will be re-evaluated once the research questions set out in the following section has been addressed.

1.2.2.1 ANALYSIS

To aid in this critique an analysis was done on a sample (n = 28 747) of patents. The presentation of this analysis will also act as an introduction to patent classification schemes, on which the Fleming-Sorenson methodology is based.

Table 1 lists example sub-class classification codes as well as the sub-class descriptions. Sub class codes form part of a hierarchal classifications system of technology types[1]. At least one such code is added to every patent.

Sub-class	Description
G01C	MEASURING DISTANCES, LEVELS OR BEARINGS; SURVEYING; NAVIGATION; GYROSCOPIC INSTRUMENTS; PHOTOGRAMMETRY OR VIDEOGRAMMETRY
G09F	DISPLAYING; ADVERTISING; SIGNS; LABELS OR NAME-PLATES; SEALS
Y10T	TECHNICAL SUBJECTS COVERED BY FORMER US CLASSIFICATION
H01C	RESISTORS
H01L	SEMICONDUCTOR DEVICES; ELECTRIC SOLID STATE DEVICES NOT OTHERWISE PROVIDED FOR

Table 1: Sub-class classification codes from the Cooperative Patent Classification system.

[1] The hierarchy levels, in order of granularity, are Section, Class, Sub Class, Main Group and Sub Group. Patents are tagged at Sub Group level. See Literature section for more detail.

From the analysis sample 11 patents are classified under the sub-class code “H01C”, or “Resistors”. These are shown in Table 2 along with the patent titles. Note that this classification is based on function, though this is not always the case. Many other classifications are solely based on form.

Patent Number	Title	Application Date
8933775	Surface mountable over-current protection device	05/06/2013
8934205	ESD protection device	27/09/2011
8935122	Alignment detection device	05/12/2011
8937525	Surface mountable over-current protection device	23/05/2013
8940193	Electronic device for voltage switchable dielectric material having high aspect ratio particles	10/06/2011
8941462	Over-current protection device and method of making the same	19/04/2013
8942552	Plastic tubular connecting sleeve for a pipe with internal liner	25/07/2011
8947193	Resistance component and method for producing a resistance component	31/08/2011
8947852	Integrated EMI filter and surge protection component	30/05/2013
8952492	High-precision resistor and trimming method thereof	30/06/2011
8957756	Sulfuration resistant chip resistor and method for making same	19/08/2013

Table 2: Sample patents classified as Resistors

It is quite rare for a patent to have only a single classification code. Table 3 shows the full list of classifications for some of the patents listed in Table 2. Note that a patent can have multiple granular classifications that may or may not fall under the same sub-class classification. Classifications as “Resistor” are highlighted in blue.

Patent No	Classifications						
08933775	H01C 7/021	H01C 1/08	H01C 1/1406	H01C 7/027	H01C 17/0652	H01C 17/06526	H01C 17/06566
08935122	H01Q 1/125	H01Q 3/02	H01Q 3/005	H01C 1/125			
08940193	H01C 7/1006	B82Y 10/00	H01B 1/22	H01B 1/24	H01C 7/105	H01C 17/0652	H05K 1/0257
08942552	H05B 3/02	F16L 1/15	F16L 1/206	F16L 13/0263	F16L 47/03	F16L 58/181	H01C 17/00
08947193	H01C 7/18	H01C 1/1413	H01C 1/146	H01C 7/041	H01C 7/008	H01C 17/00	
08947852	H02H 7/00	H01G 4/30	H01C 1/14	H01C 7/12	H01G 2/14	H01G 4/005	
08957756	H01C 1/034	H01C 17/288					

Table 3: Full classifications of Sample patents classified as resistors

Table 4 lists the sub class level classifications that co-occur with the H01C classification, as shown in Table 3. These include form descriptions, such as F16L ("Pipes"), as well as functional classifications, such as H05B ("Electrical Heating").

Sub-Class	Description
B82Y	SPECIFIC USES OR APPLICATIONS OF NANOSTRUCTURES; MEASUREMENT OR ANALYSIS OF NANOSTRUCTURES; MANUFACTURE OR TREATMENT OF NANOSTRUCTURES
F16L	PIPES; JOINTS OR FITTINGS FOR PIPES; SUPPORTS FOR PIPES, CABLES OR PROTECTIVE TUBING; MEANS FOR THERMAL INSULATION IN GENERAL
H01B	CABLES; CONDUCTORS; INSULATORS; SELECTION OF MATERIALS FOR THEIR CONDUCTIVE, INSULATING OR DIELECTRIC PROPERTIES
H01G	CAPACITORS; CAPACITORS, RECTIFIERS, DETECTORS, SWITCHING DEVICES OR LIGHT-SENSITIVE DEVICES, OF THE ELECTROLYTIC TYPE
H01L	SEMICONDUCTOR DEVICES; ELECTRIC SOLID STATE DEVICES NOT OTHERWISE PROVIDED FOR
H01Q	AERIALS
H01T	SPARK GAPS; OVERVOLTAGE ARRESTERS USING SPARK GAPS; SPARKING PLUGS; CORONA DEVICES; GENERATING IONS TO BE INTRODUCED INTO NON-ENCLOSED GASES
H02H	EMERGENCY PROTECTIVE CIRCUIT ARRANGEMENTS
H05B	ELECTRIC HEATING; ELECTRIC LIGHTING NOT OTHERWISE PROVIDED FOR
H05K	PRINTED CIRCUITS; CASINGS OR CONSTRUCTIONAL DETAILS OF ELECTRIC APPARATUS; MANUFACTURE OF ASSEMBLAGES OF ELECTRICAL COMPONENTS
Y10S	TECHNICAL SUBJECTS COVERED BY FORMER USPC CROSS-REFERENCE ART COLLECTIONS [XRACs] AND DIGESTS

Table 4: Sub-classes co-occurring with H01C (Resistors)

This short analysis aims to demonstrate several aspects of patent classification. Apart from providing an introduction to patent classification it is also meant to provide some insight into the different technologies that are clumped under single sub class classifications. The examples above are purposefully focused on the classification of resistors (H01C) and its usage in co-classification with other sub class codes. The classification has been chosen deliberately to demonstrate some of the shortcomings in the Fleming-Sorenson methodology. These will be described in the remainder of section 1.2.2.

1.2.2.2 COVERAGE OF SUB-CLASSES

Table 1 lists example sub-class classification codes as well as the sub-class descriptions. It is clear from the descriptions alone that several technologies, or even branches of technologies, can fit under any of the listed sub-classes.

The metric for interdependence assumes that all the technologies within a sub-class will have a similar amount of interdependence. Inspection of the classification code descriptions in Table 1 should intuitively discount this notion, or at the very least instil a healthy dose of

scepticism. The scope of sub-class classifications is clearly broader than a single technology stream or field of innovation.

A second assumption is that patent classification scheme structure is determined only by the nature of technology being classified. Consider sub-class Y10T in the Table 1. This classification is meant to absorb classifications from the deprecated US classification system and does not differentiate between different branches of technology. This should, at the very least, cast serious doubt on whether sub-class classification is the right level of classification to be used as an indicator of interdependence within a technological instance.

1.2.2.3 SUITABILITY OF PROXY VARIABLES

As a corollary to the point above, it is not clear how classification codes relate to the interdependence of the constituent technologies within a patent. The validity of its use as a proxy must therefore be brought into question. Consider the following given definition, the calculation of the ease of recombination (inverse of interdependence).

$$Ease\ of\ recombination\ of\ sub-class\ i\ \equiv E_i = \frac{Count\ of\ sub{-}classes\ previously\ combined\ with\ \ i}{count\ of\ previous\ patents\ in\ sub{-}class\ i}$$

Patents are not classified at sub-class level. The equation above is a measure of the number of sub-classes co-occurring with a specific sub-class *i*, normalised to the number of patents in that subclass. This indicates that two very broad classes of technology interact, but it does not measure the ease with which they recombine. Furthermore, the number of patents published under a sub-class is not necessarily an indicator of its recombinative potential, which draws into question the normalisation method used.

Another consideration is that not all the constituent technologies that make up a patented technology is explicitly mentioned in the classifications. One aspect of Table 3 that is immediately apparent is the small number of sub-classes that co-occur on patents classified as resistors (Table 4). Note that it only contains other sub-groups that define the function or form of the patent directly. There is no co-occurrence with sub-classes that represent technologies that use resistors as a component. This is problematic when trying to measure the ease with which resistors integrate into other technologies. It is clear that it is easy to integrate resistors into most applications – they are designed with this purpose in mind. It is therefore probable that the ease of recombination with other technologies should be high. This is obvious at a sub-class level, but it does not capture the intricacies and interdependencies implicit in the production of the resistor itself.

It should also be mentioned that the definition for ease of recombination does not take into account the ease with which technologies within a sub-class recombine or interact. Table 3 show the classifications given to a set of patents. Some of these have classifications from a diverse set of sub classes while others only have classifications from a single sub class.

1.2.2.4 PATENT SAMPLING

Fleming and Sorenson's sample size was n = 17 264. This would seem sufficient for most studies, but a problem occurs when their methodology is contingent on a patent classification system with more than a quarter of a million classification codes and with at least 2 000 sub class classifications.

Random sampling also poses a threat in that some sub classes are assigned more frequently than others. This is clearly demonstrated in the analysis provided at the onset of this section: Out of a sample of 28 747 patents only 11 are classified as resistors.

In summary, sub-classes are not necessarily based on a branch of technology, does not represent a single configuration of parts and their co-occurrence on a patent does not define the ease of recombination between the subclasses. The use of these metrics as proxies for complexity measures is thus brought into question.

1.3 RESEARCH QUESTIONS

A critical analysis of the Fleming-Sorenson study [11] invokes several questions on the use of complexity metrics derived from patents. From this appraisal it is reasonable to ask -

Research Question 1
What is the relationship between patent classification codes and the amount of parts, part interactions and part interdependencies in patented inventions?

Patent classification codes are only one dimension of patent data. Several other fields, such as citations and inventor information, are also available. It would therefore be prudent to expand the parameter of patent classification codes to patent metadata. This will allow for a more contextualised analysis of the problem.

The parameter of part count, interactions and interdependence all stem from complexity metrics. To allow for a more general approach these are considered under the umbrella term of “complexity”.

Further refinements to this question will be presented in the synthesis of the literature. With all the current considerations the first research question can therefore be refined to –

Research Question 1
What is the relationship between patent metadata and invention complexity in patented inventions?

To answer the question above requires that several sub-questions should be considered first:

Research Question 1.1
What is the range of data fields and metadata available in patents?

Research Question 1.2
How is patent information useful in the study of technology, innovation and complexity?

Research Question 1.3
What measures for complexity exist or can be defined that apply to patentable inventions in general?

Research Question 1.4
How can identified measures of complexity be extracted from patent documentation?

Once all of the questions above have been addressed, their output can be used to review the Fleming and Sorenson study [11] according to the following criteria:

Research Question 2

Given a set of parameterised relationships between complexity metrics and patent data (Research question 1), would the conclusions of the Fleming and Sorenson study remain valid?

The last question aims to measure the impact on the validity of the original study. It is quite possible that their metrics and the chosen proxies are valid. It is also possible that insights into these would shed light on weaknesses within their methodology.

1.4 LITERATURE SUMMARY

The literature in the following section is arranged according to the research questions. The first section investigates the data fields and metadata available in patents. Kim & Lee [13] investigated different patent sources and concluded that the source choice would have a direct impact on the outcome of any innovation study. They recommend using data from the United States Patents and Trademarks Office (USPTO) to be most fitting for innovation studies. The South African patent office was also investigated, but found wanting on several accounts [14]. Different data fields from patents are explored. It was found that patent titles, abstracts, claims, descriptions, citations and classifications could all be used in innovation related studies [15] [16]. It was shown that especially the USPTO patents have very well defined and granular XML structures that can be leveraged to extract data from patent texts. No prior publication that uses this approach could be found.

The second research question looks at how patent information is useful in the study of innovation, technology and complexity. The nature of innovation and technology was scrutinised first. It was shown that evolutionary analogies abound in the descriptions of innovation [17] [18], innovation can be modelled in various ways [11] and innovation is interpreted differently according to the field it being studied in [19]. The TRIZ design methods were explored to enrich the background on innovation theory. TRIZ design is a set of principles used to abstract a design problem to help find creative solutions. It is accompanied by a set of observations on the nature of technological growth and trends [20]. The nature of complexity was investigated by looking at Baldwin & Clark's [21] analysis of complexity in modular design. They posit that complexity is not limited to an artefact, but rather a systemic phenomenon. They furthermore posit that the key to managing complexity is modular design. The properties of complex adaptive systems are also explored [22] [23] [24]. It was found that there is very little consensus around what constitutes a complex system. Most authors agree

that a complex system consist of many actors interacting in varied ways that gives rise to unexpected aggregate behaviour.

The next literature section covers the quantitative modelling of technology and innovation, with a specific focus on the use of patent. A more detailed overview of the Fleming and Sorenson [11] study is provided. The literature also demonstrated that patent classifications can be used to extend keyword based patent searches to concept based searches [25]. Other search methods use natural language processing techniques to impose a function-behaviour structure on patent text. This structure is used to classify interactions within patents into TRIZ solution sets (mechanical, thermal, wave based etc.). This allows solutions to be grouped according to these types [26]. Other modelling approaches demonstrated that the value of a patent can be estimated by looking at the number of citations it receives [27]. On average every citation ads 3% more value to a patent. One approach aims to identify valuable patents by using TRIZ evolution trends [28]. These predict that every branch of technology goes through a measurable lifecycle. Patents with a trend phase higher than the average of the technological domain is deemed more valuable. A great contributor to this study is Luo and Wood's investigation into changes in the complexity of the innovative process [29]. They traced several patent data fields over a period of 30 years. They showed that the innovation process was getting more complex over time and that this was measurable from patent data.

The last section of the literature deals with natural language processing principles and tools for natural language processing. The section analyses the concepts of tokenisation [30], stemming and lemmatisation [31], part of speech tagging [32] and sentence parsing [30]. Tools for language processing such as the Stanford Parser [33] and WordNet were reviewed [34].

1.5 OVERVIEW OF SYNTHESIS

In chapter 3 several concepts set forth in the literature is explored and developed. It also contains the working definition of complexity that will be used for the remainder of the study and explores the difficulties and possibilities of measuring patent data.

The literature is divided on a definition for complexity. This is, at least in part, due to the vast differences in approaches used in fields of study that consider complexity, as well as the nuanced nature of the subject. There is, however, a general convention when complexity is applied to patent data and this will also be adopted in this study. In most cases the number of parts and part interactions are modelled as proxies for the complicatedness of patented inventions.

The plurality and similarity of parts in inventions makes it difficult to simply count the number of parts. To overcome this two measures for part count is introduced. The first is simply a brute count of every part mentioned within a patent. The second aims to normalise the part count by extracting the lexical similarity between parts are use this measure to extract a set of unique parts. As a matter of convention any reference to the “normalised count” refers to this approach.

Two approaches to measuring the interaction between parts is introduced. The first is to use statistical language processing methods to extract the relationships between parts, and then assign complexity weightings to these relationships. An alternate, more simplistic, approach is to assume that the co-occurrence of parts within a phrase indicates some mode of interaction between them.

The synthesis concludes with a manual analysis of patent XML mark-up and how it can be leveraged to extract parts. It was found that some very simple tagging conventions can be utilised to extract part names. As a general rule patent part references are accompanied by the part number shown in the patent drawings. These are emboldened with a XML <b> tag. These tags can therefore be used as entry points to extract part mentions from the free text description of a patent.

1.6 OVERVIEW OF METHODOLOGY

1.6.1.1 INTRODUCTION

From the Literature and synthesis a methodology was constructed to test the research questions. To this end several metrics were defined and tests for their validity were set forth. Filters were included on the sample set to ensure that patents adhered to the required standard for extracting parts and part interactions.

1.6.1.2 PATENT FILTERS

Not all patents are testable through the methods used in this study. Most notably, patents relating to chemical formulae and processes rarely use part names or the accompanying tags used to extract them. These are therefore stripped out of the sample set. In exceptional cases patents deviate from the prescribed format. Patents with part counts smaller than two times the patent’s figure count are also removed.

1.6.1.3 METRICS AND MEASUREMENTS

Three types of metrics are defined – base complexity metrics, meta-data metrics and segmentation metrics. The base complexity metrics includes the different measurements of part and part interaction counts. The meta-data metrics aim to quantify several other fields of patent information. This includes classification code quantities and variance, citation counts and claim counts. The complexity metrics are measured against the meta-data metrics to test if any correlation exists between these data fields and the invention's complexity. The last metrics category, segmentation metrics, is a set of binary patent classifications. These aim to classify patents as systemic or standalone inventions and are also compared against the complexity metrics.

The complexity metrics are measured against Fleming and Sorenson's method to test the accuracy of their method.

1.7 OVERVIEW OF IMPLEMENTATION

A large sample of patents was acquired and analysed according to the methodology. The change of average patent complexity metrics were measured and demonstrated a strong linear increase in part count and interaction over time. This, in combination with findings in the literature, validated the idea of using patent parts and interactions as complexity proxies.

During the implementation of the methodology it became apparent that the use of statistical language processing techniques were not suited. The preliminary results were promising, but the computational complexity made it impractical to apply to a statistically significant sample set.

1.8 OVERVIEW OF RESULTS

The following highlights are presented from the study results –

1. The number of parts in a patent and their measured interaction can be used as a proxy for complexity in patented innovations. This observation was validated by demonstrating that the increase in the complexity of the innovation system was reflected in the number of parts and part interactions within patents.
2. The distributions of parts vary for different sub-classes within the patent classification hierarchy. The distribution is determined by the scope of the sub-class and the nature of the technological streams it covers.

3. The distribution of parts varies significantly for lower level classifications within the same sub-class. This indicates that classification sub-classes are not representative of the complexity of a particular technology stream. A more granular grouping of patents under a specific technology stream is required.
4. Patents with the word “system” in the title tend to be more complex than the average patent in the sample. This demonstrates that keyword based filters can be used to find more complex inventions.
5. The amount of citations on a patent correlates positively with the complexity of the patented invention. This follows intuitively as it is probable that more complex inventions with more parts need to cite more works.
6. Patents with citations to other documents beside earlier patents tend to be more complex than patents only citing earlier patents.
7. The number of claims on a patent is positively correlated to the number of parts in the patent. However, a claim count distribution in the sample also demonstrated the influence of other factors on patent data. A disproportional number of patents have exactly 20 claims. Patentees want the maximum amount of protection for their intellectual property, but need to pay additional costs for every claim in access of 20. This is only one example of a systemic influence that introduces noise into the data.
8. Fleming and Sorenson’s metric of interdependence is flawed in several ways. The metric is built on the premise that sub-class classifications encapsulate the complexity of a patented invention. This might be true for a small subset of sub-classes, but as a general rule it is a fallacy. The normalisation process also does not consider that some sub-classes are assigned only once or twice in the sample. This results in additional skewing. It is not surprising that no direct relationship could be found between this metric and other complexity proxies.
9. With the high noise levels in the data it was concluded that complexity could be gauged from patent data, but only in aggregate. The sample size in this study contained over 100 000 patents. Even with the massive sample statistical significance degraded in areas where the measured variable has a sparse distribution.

2 LITERATURE STUDY

2.1 OVERVIEW

The literature reviewed and presented here follows the general flow and order of the research questions posed in chapter 1. The research questions are disaggregated into several themes before being addressed.

1. Patent Information
    a. Patent data sources and their suitability in innovation studies.
    b. The data fields available in patent documentation.
2. Patents, Technology and Innovation (Qualitative)
    a. The nature of technology.
    b. Views on innovation and design.
    c. Views on complexity.
3. Patent data in innovation studies (Quantitative)
    a. The modelling of innovation and technology
    b. Patent and technology discovery
    c. Patent impact and value modelling
    d. Metrics of innovation and technology
4. Measurement of Complexity in Innovation
5. Language Tools
    a. Information Retrieval
    b. Natural Language Processing Methods and Capabilities
    c. Language Resources and Tools

2.2 PATENT INFORMATION

2.2.1 REVIEW SCOPE

Research question 1.1 reads as follows - *What is the full range of data fields and metadata available in patents?* This question is addressed under the following themes:

1. Patent data sources and their suitability in innovation studies.
2. The data fields available in patent documentation.

2.2.2 PATENT DATA SOURCES

Several institutions and patent offices offer patent search and download services. Kim & Lee [13] posited that the selection of a patent database will have a direct effect on the outcome of any study based on patent search results. They compared the databases of the United States Patents and Trademarks Office (USPTO), the European Patent Office (EPO), the Japanese Patent Office (JPO) and the Korean Intellectual Property Office (KIPO). The authors concluded that the USPTO is the most fitting for innovation studies and is representative of global innovation patterns. The USPTO also has the most applications from foreign applicants and the widest variety of applications. Another advantage of the USPTO database is the presence of citations of prior art within the documents, but it could be argued that this is an outflow of the legal regime governing IP in the USA and not a reflection on the quality of the database itself. Furthermore, it is found that the EPO database and JPO database contain less information than the USPTO, but can still provide relevant information for the study of global innovation trends. The authors conclude that using any of the three sources can provide adequate information for an innovation based study.

The format of available information differs from source to source. The EPO provides an API form where the bibliographic patent information of about 90 000 000 patent documents from 334 countries can be accessed. This includes 239 791 partial patent documents from South Africa. The API also provides access to full-text (description and claims) patent documents published through the EPO, United Kingdom, World Intellectual Property Organisation (WIPO), Austria, Canada, Switzerland and Spain. Query responses can be returned in both XML and JSON formats. [35] [36]

The USPTO provides a bulk-download function. A file is compiled for every week of the year and contains all the patent information of documents published in that time. The files contain

full text information on all patents. All documents are available in XML with an accompanying schema file. [37]

The drawbacks to the USPTO system is that the files contain patents filed in the US alone and all information will need to be downloaded and processed before a meaningful search can be achieved. The entire collection is approximately 8 TB in size, making it impractical to first filter patents when building a sample set.

Unlike the United States, South Africa has a non-examining patent system. It is therefore possible that a patent can be filed without proof of novelty. A study on prior art is only done in the event that a case of infringement is laid before the court. Pouris & Pouris [14] demonstrated that the current intellectual property rights regime *"...not only fails to support the objectives of the national innovation system but also facilitates exploitation by foreign interests and creates substantial social costs."*

The same research shows that more than 80% of patents granted in South Africa would not have been granted if they were examined with the same level of scrutiny applied in countries such as Canada, Australia and the United States.

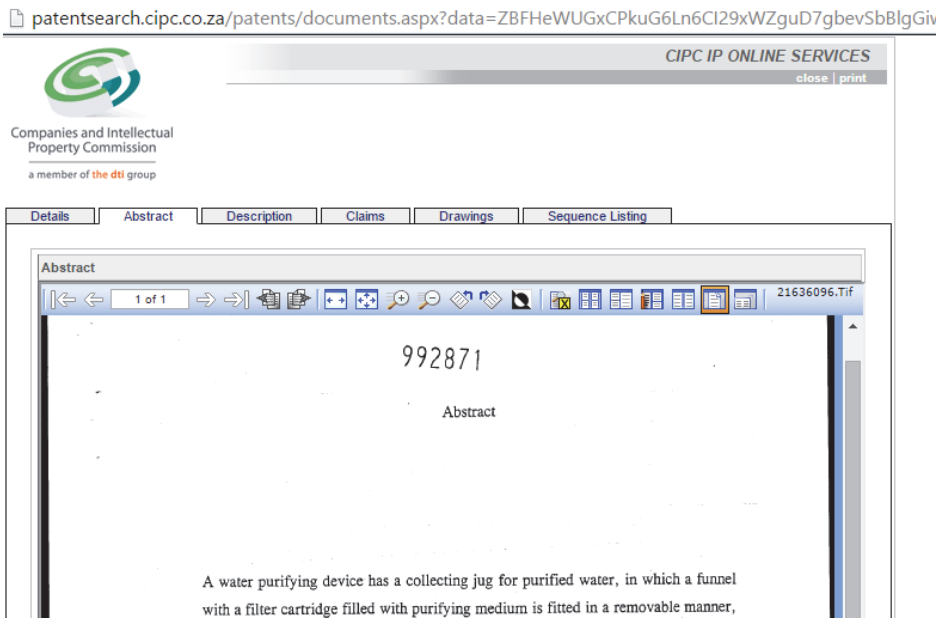


Figure 1: CIPC Patent Search

Not only is the quality of South African patents a point of concern, but also the availability. South African patents can be searched on the Companies and Intellectual Property

Commission's website [38], but only the title and several biographical fields are properly digitised. Most documents are scanned in as images, making textual manipulation and extraction impractical. Figure 1 shows an example search result from CIPC. The lack of quality in both the nature and availability of South African patents disqualify them as a viable source of technological knowledge.

2.2.3 A CATALOGUE OF PATENT DATA

2.2.3.1 OVERVIEW OF A PATENT

The World Intellectual Property Organisation (WIPO) defines a patent as *"an executive right granted for an invention, which is a product or a process that provides a new way of doing something, or offers a new technical solution to a problem"*. [16]

An invention must fulfil several criterions to be considered patentable. It must contain an element of novelty, in other words it must contain some element not known in the existing body of knowledge. It should also show an inventive step that would be unobvious to a person with an average knowledge of the field. Several types of knowledge are generally not patentable. These vary according to region, but generally include scientific theories, mathematical methods, plant or animal varieties, natural substances, methods for medical treatment and creative endeavours such as musical composition. [16]

The use of patents extends beyond the exclusive right to an invention, even though the structure of patents is geared towards this purpose. Several prominent examples are available of patents' usefulness outside of their intended legal framework. Griliches [7] used patent data to model the effects of research and development on the market value of firms. Jaffe et al. [8] used patents citations to trace the geographical extent of knowledge dissemination. The patenting environment itself is also a prominent source of data. Sakakibara and Branstetter [9] showed that the 1988 patent reforms in Japan, which significantly expanded the scope of patent rights, had no measurable effect on either R&D spending or the innovative output from affected firms. Patents are also used for forecasting technological developments in a particular domain, finding input to R&D tasks, technological road mapping, strategic technology planning and the identification of competitors. [39]

2.2.3.2 PATENT SECTIONS

This section explains the general structure and components of a patent. Although slight variations exist between the patent publication formats of various regions the main concepts remain unchanged. A patent (US 4741121 A) is used to illustrate these. [15]

Patent Section Description	**Example from Patent US4741121**
The Invention title – A general description of the invention.	*"Gas chamber animal trap"*
As with academic publications the abstract provides a synopsis of the invention.	*"A compact industrial animal trap is provided with a gas injector to effectively, continually and rapidly exterminate rodents and other animal pests with carbon dioxide or other gases in a reliable, efficient and safe manner. The animal trap has a disposal chamber and an elongated entrance chamber. In the preferred form..."*
The claims form the legally operative part of the patent and are therefore central to the document. These provide a detailed list of components and functions of the technology being patented.	*"1. An animal trap, comprising:* • *an elongated entrance chamber for attracting and an animal, said entrance chamber providing bait-dispensing means for dispensing the aroma of bait and having at least one access opening defining an entrance;* • *carbon dioxide powered door means in said access opening being... "*
The patent description encapsulates the background of the invention, gives context to the claims, describes any cited prior art and provides a preferred embodiment of the invention. The description	*This invention pertains to industrial animal traps, and more particularly, to devices for killing rodents and other animal pests.* *In the farming, harvesting, and storing of food grains, it has been estimated that as much as 30% of the food products are lost to rodents (rats, mice, etc.) whether the food be in the field, in a silo, or in transportation. The worldwide loss*

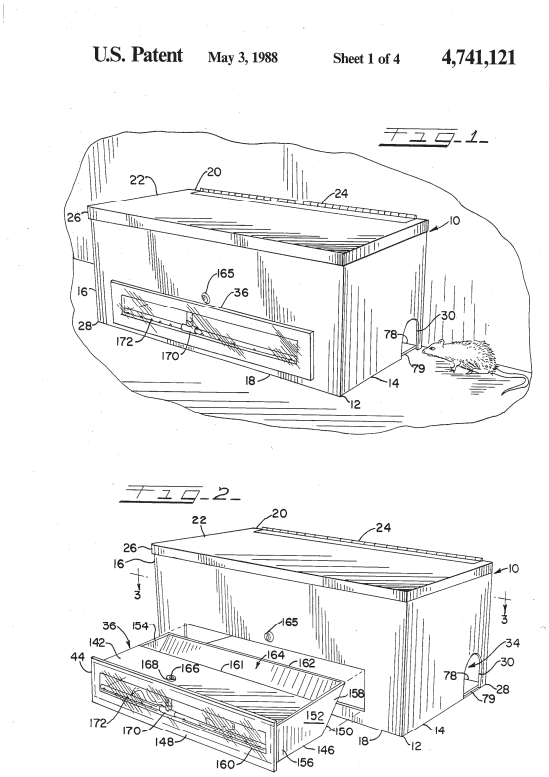
can reach several pages in length.	*due to rodent consumption has been estimated to run into billions of dollars…*
The components in the drawings are numbered and referred to in both the claims and description sections of the patent. An example is shown in Figure 2.	

Figure 2: Drawing of Patent - US 4741121

Table 5: Patent Sections with examples

Several other data fields are also available that is not used to describe the patent directly. These include the applicants, inventors, citations, classifications and publication date.

2.2.4 Patent Classification

Several schemes are currently used to classify the different classes of technology present within a patent. Any given patent is manually assigned one or more of these classifications.

2.2.4.1 INTERNATIONAL PATENT CLASSIFICATION [40]

The International Patent Classification (IPC) scheme was first published in 1968 and had undergone several revisions since then. It forms the basis of several other patent classification schemes currently in use. An example of the classification hierarchy is shown below.

- Section e.g. A - HUMAN NECESSITIES
- Class e.g. A01 - Agriculture; Forestry; Animal Husbandry; Hunting; Trapping
- Subclass e.g. A01B - Soil working in agriculture or forestry ...
- Main Group e.g. A01B3/00 - Ploughs with fixed plough-shares
- Subgroup e.g. A01B3/04 - animal-drawn ploughs

This classification system is used by over a hundred patent-issuing bodies, making it the most widely used patent classification system in use today. [41]

2.2.4.2 COOPERATIVE PATENT CLASSIFICATION [42]

The Cooperative Patent Classification (CPC) was developed by the United States Patent Office and the European Patent Office in an attempt to create a unified global classification system. It is based on the IPC, but offers additional detail in around 200 000 subdivisions. Several major patent offices, including China and Korea have made their intent clear to adopt the system over the next few years. The classification scheme is updated on more regular bases than many of the other systems and is therefore more suited for classifying fast developing technologies.

Several other classifications schemes exist, but a full review of them is unnecessary as most major patent offices use either the IPC or CPC schemes. Table 6 shows a list of countries and organisations whish adopted the CPC system.

Country	Implemented CPC in Patents published from -
ARIPO[2]	3/7/1985
Austria	15/1/1971
Australia	18/1/1973
Belgium	1892
Canada	4/8/1970
Switzerland	31/1/1939
Germany	04/1/1973
EPO	20/12/1978
France	20/12/1968
United Kingdom	27/1/1910
Luxembourg	1920
The Netherlands	1913
OAPI[3]	15/01/1966
The United States	04/10/1855
The World	19/10/1978

Table 6: Global CPC Implementation [43]

2.2.5 Justification as a Data Source

Patents are a central part of any intellectual property rights (IPR) system. The primary theoretical objective of IPRs is to supplement market forces which on their own do not lead to

[2] African Regional Intellectual Property Organisation. ARIPO has 19 member states of which South Africa is not one.
[3] African Intellectual Property Organisation. OAPI has 17 member states, all in North Africa.

desired levels of research and innovation. IPRs aim to encourage innovation and the diffusion of technology which in turn fuels economic growth. [14]

Patents contain a unique description of technology, and in many cases, it is the only public record of an invention or its composite parts. There are various estimates for the uniqueness of information found in patents. A 1977 study on US patents showed that about 70% of patents published between 1967 and 1972 were not disclosed in any other form of publication [44]. In 2011 it was reported that more than two-thirds of patents relating to vascular health and risk management had no parallel publication in a scientific journal [45] and according to a 2007 joint report by the USPTO and EPO *"...up to 80% of current technical knowledge can only be found in patent documents."* [6]

Even though patents contain a wealth of information, as demonstrated above, several other reasons exist to justify their use in this study.

Firstly, the data in patent publications are generally well structured. Both the USPTO and EPO employ XML schemas towards this purpose. An example of the USPTO XML schema is shown in Figure 3. Secondly, patent information is available more freely than many other viable data sources, such as academic publications. Lastly, it should be noted that a large amount of research is available on the extraction of knowledge from patents.

```
<?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet type="text/xsl" href="/3.0/style/exchange.xsl"?>
<ops:world-patent-data xmlns="http://www.epo.org/exchange" xmlns:ops="http://ops.epo.org" xmlns:xlink="http://www.w3.org/1999/xlink">
    <ops:meta name="elapsed-time" value="834"/>
    <ops:biblio-search total-result-count="100000">
        <ops:query syntax="CQL">pn = US and cpc =/low Y</ops:query>
        <ops:range begin="1" end="10"/>
        <ops:search-result>
            <exchange-documents>
                <exchange-document system="ops.epo.org" family-id="51870016" country="US" doc-number="RE45853" kind="E1">
                    <bibliographic-data>
                        <publication-reference>
                            <document-id document-id-type="docdb">
                                <country>US</country>
                                <doc-number>RE45853</doc-number>
                                <kind>E1</kind>
                                <date>20160119</date>
                            </document-id>
                            <document-id document-id-type="epodoc">
                                <doc-number>USRE45853E</doc-number>
                                <date>20160119</date>
                            </document-id>
```

Figure 3: XML Encoded Patent Data

2.2.6 DISCUSSION

To demonstrate the different data fields available in patent documentation, the different sources and formats were firstly considered. It is apparent that the only two practical data sources are the USPTO and EPO. The data fields available in patent documents from these offices are remarkably similar. One difference between the offices is the way in which data

can be accessed. The EPO provides a web interface that can filter data according to most data fields in the document. The USPTO only provides the information in bulk downloadable archives. It was also noted that the USPTO XML schema is more granular than that of the EPO. The source of patent data thus becomes a trade-off question that needs to be addressed in the methodology.

It has been shown that patents contain fields for a title, abstract, claims, detailed descriptions, drawings, applicants, inventors, citations and dates. These provide many dimensions from which to analyse and dissect patent information. The mentioned fields are all at patent level. Many other useful metrics emerge when patenting is viewed systemically. The publication rate of inventions under certain classifications is one example.

2.3 PATENTS, TECHNOLOGY AND INNOVATION

2.3.1 REVIEW SCOPE

The second research question reads - *How is patent information useful in the study of technology, innovation and complexity?*

This question can only be answered once an understanding of innovation, technology and complexity is gained. The response to this question will be split over two sections of the literature study. This section will deal with the underlying concept of technology, innovation and complexity. The next section aims to show how these concepts are used in combination with patent data.

This section will be addressed under the following themes:

1. The nature of technology.
2. Views on innovation and design.
3. Views on complexity.

2.3.2 DEFINING TECHNOLOGY

To understand how technology changes this section firstly aims to elicit the nature of technology and thereafter the natural evolution thereof. Many descriptions and perspectives can be used to describe the nature of technology. Some studies treat technology as an economic concept, while others focus on engineering design. In this section it will be explored from both perspectives.

The Oxford dictionary defines technology as *the application of scientific knowledge for practical purposes, especially in industry* [46]. The Webster Dictionary is a bit more liberal and states that technology *is a capability given by the practical application of knowledge* [47].

These definitions capture the heart of what aims to be measured in this study. It is wide enough that it not only looks at physical artefacts and how they evolve, but also at ideas and how they manifest in the physical due to human action. This implies that any measurement of technological change should account for changes in ideas and thought. Language itself should then be considered a technology and changes in lexicon a progress or regress.

2.3.3 INNOVATION

Innovation is another popular description of technological change. Several fields of study are dedicated to its progress, engineering included. This section explores the various literature branches relating to innovation.

2.3.3.1 EVOLUTIONARY ANALOGY

It is a long tradition to borrow biological terms and frames of reference to describe the changes in the technological environment. Fleming & Sorenson [11] ascribe some of the earliest instances of this phenomenon to Gilfillan [17] who in 1935 wrote "*The nature of invention ... is an evolution, rather than a series of creations, and much resembles a biologic process*". Later in 1939 Schumpeter [18] proposed that the effects of inventions "*... illustrate the same process of industrial mutation - if I may use that biological term - that revolutionises the economic structure from within incessantly destroying the old, incessantly creating a new one.*"

The analogy is extended in that technology is described as having a lifecycle, starting with its invention, going through growth, adoption and spread phases and finally becoming obsolete as a new paradigm replaces it. These new paradigms are termed disruptive technologies and can be defined as an emerging technology whose arrival signifies the eventual displacement of the dominant technology in that sector [48].

2.3.3.2 RECOMBINANT INNOVATION

Recombinant innovation is the principle that many inventions use prior inventions as components. In this view an invention can be classified as either a synthesis of existing and/or new technologies or a refinement of a previous combination of technologies. [11]

This phenomenon is visible in both products and processes. For example, the automobile can be thought of as a recombination of the wheel, bicycle, horse carriage and combustion engine. An example of recombinant process innovation is the use of Formula 1 pit-stop and aviation models in patient handover from surgery to intensive care. The incorporation of these models improves the safety and quality of care received by the patient. [49]

2.3.3.3 OPEN INNOVATION

Henry Chesbrough defines Open Innovation (OI) as the purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation, respectively. It is the paradigm that assumes that firms can and should use

external ideas as well as internal ideas, and internal and external paths to market, as they look to advance their technology. [50]

At its heart, OI is a model for innovation management and the study of its finer nuances falls outside the scope of this enquiry. The paradigm was popularised by Chesbrough in 2003 [19]. His original work currently has gathered more than 11 000 citations over the last 12 years [51]. It has been shown that the bulk of the research can be classified into four streams [52]. Broadly stated these are strategic planning and external sourcing, user-centric innovation, technology and innovation management and resource and knowledge-based view of the firm.

2.3.4 DESIGN METHODS

2.3.4.1 INTRODUCTION

The design process is an intricate part of the innovative process. There are many methods available that aim to facilitate the design of a process or artefact. One such method, TRIZ, is presented here for context.

2.3.4.2 THE TRIZ CONCEPT [20]

TRIZ is the Russian acronym for Theory of Inventive Problem Solving. The methodology was developed by G.S. Altchuller and his colleagues in the former USSR between 1946 and 1985. They analysed more than 3 million patents to discover patterns that predict breakthrough solutions to problems. The hypothesis behind the theory is summarised as *"Somebody someplace has already solved this problem (or one very similar to it). Creativity is now finding that solution and adapting it to this particular problem."*

The main findings of the research conducted in this field are:

1. Problems and solutions are repeated across industries and sciences.
2. The nature of the contradictions within a problem determines the possible solution set.
3. Patterns of technical evolution are repeated across industries and sciences.
4. Creative innovations tend to use scientific effects outside of the field where they were developed.

2.3.4.3 TRIZ PROBLEM SOLVING [20]

TRIZ differentiates between *general* and *specific* solutions. Figure 4 illustrates the problem solving method.

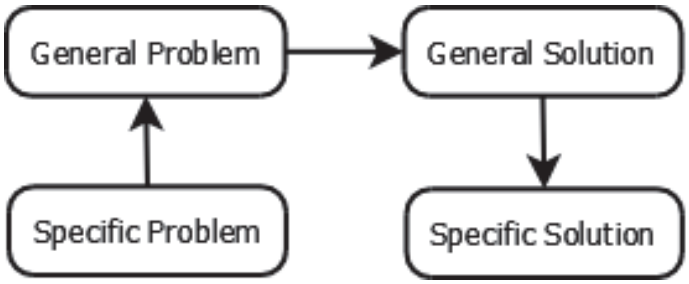


Figure 4: TRIZ Problem Solving Method

Consider the following example problem: Bacteria is utilised to produce a hormone for pharmaceutical use. In order to extract the hormone, the bacteria cell structure first needs to be broken down. A mechanical grinding process is used to achieve this, but the yield is variable between 60% and 80%.

In generalised terms the ideal solution is a process with no waste, that has 100% yield and that is less complex than the previous solution. One principle in TRIZ is that chemical solutions replace mechanical objects and chemical solutions are replaced by field-based solutions in the process of technological advancement. As an example, energy storage for operating small device such as a watch has evolved from spring storage (mechanical), to battery power (chemical) to capacitors (field).

The same set of solutions is applicable to the sample problem. The current solution is in the mechanical phase. Enzymes can be added to dissolve the cell structure, which would constitute a chemical solution or it can possibly be broken up with the application of ultrasound which will be a field based solution.

2.3.4.4 TRIZ TECHNOLOGY TRENDS [53]

TRIZ posits that several trends are present in the macro-evolution of technology. These are presented in this section.

The first trend is the **increase of ideality**. A technology becomes more suited to its intended function in its evolutionary process. In other words, functionality is increased while the negative effects of the technology (high cost, environmental impact, etc.) diminish.

The **Law of Transformation of Quantity into Quality** states that quantitative changes in a system take place continuously according to the S-curve of evolution. When a certain limit within the system is reached, it undergoes qualitative changes. A new S-curve is created and

the cycle repeats itself. In summary, quantitative changes in technology is a continuous process whereas qualitative changes manifest at discrete intervals.

The **Law of the Negation of the Negation** states that progressive technological evolution consists of a series of repetitions. The same phases are repeated, but every repetition occurs at a higher level of evolution which makes use of new materials and technologies.

2.3.5 COMPLEXITY

Complexity presents itself as a natural measure for technological progress, especially as a general measure. Perspectives of what complexity is, varies vastly and therefore the definitions of complexity contains a lot of variation [23]. If we should look at the Google dictionary definitions of complex two stand out [54]:

a. Consisting of many different and connected parts

b. A group or system of things that are linked in a close or complicated way; a network.

Both of these embody the standard use of the phrase as well as most of the definitions found in more formal and field specific publications.

A more intuitive way to look at the complexity of an artefact is penned by Baldwin & Clark [21]. Consider a line of artefacts organised from extremely simple to extremely complex. On the one end might be a toothpick and on the other an F-22 fighter jet. Two conceptual points would emerge as one would move from simple to complex. The first is the where one person is no longer sufficient to produce the artefact, the second where one person can no longer comprehend the artefact in its entirety. Crossing the first threshold would require a division of labour. Crossing the second would require a division of knowledge. This might not be the most advanced analogy in terms of measurable outcomes, but it does tease out a rather potent truth – the complexity of a technological artefact is not only a function of the object's intrinsic properties, but also of the knowledge and process that produce and distribute it. It can therefore be said that complex artefacts are embedded in complex systems. These systems aim to divide labour and knowledge, and provide tools for task coordination and decision making. These coordination systems must become more complex as artefacts themselves become more complex. The capacity and efficiency of these systems will have a direct influence on the cost of designing and producing an artefact. [21]

The key to economically viable mass production of complex artefacts is *modular design*. This is the central theme of Baldwin & Clarke's study on the evolution of design in computers. The success and phenomenal growth of the industry would not have been possible without modular thinking.

2.3.6 COMPLEX ADAPTIVE SYSTEMS

Technological complexity as a term only makes sense under a systemic view of the technological world. It is an interaction rich environment where a multitude of intertwined factors create the space for new development. In this light an overview of complex adaptive systems is provided.

To understand the nature of complex adaptive systems, their properties should first be investigated – in other words what makes them adaptive and complex. Hollard [22], one of the earlier writers on the subject, provides three of these properties – evolution, aggregate behaviour and anticipation. Evolution in this context constitutes two concepts: firstly, it refers to the optimisation and replacement of parts in a modular technological artefact and secondly to the technological artefact's ability to adjust to variant external stimulus. A simple example of the latter is the feedback loop between a thermostat and furnace to keep the furnace at a constant temperature. Aggregate behaviour is described as the behaviour arising from the interaction of many components that cannot be directly attributed to their individual behaviour. Anticipation refers to the ability to anticipate the likelihood of an event based on the occurrence of a preceding event.

Other observations of note include:

a. As a result of aggregate behaviour, the system is usually not in an optimal state, if indeed an optimal state can be defined for the system as a whole.

b. Complex adaptive system behaviour is not governed by a single equation or rule. Each component in the system is controlled by a local set of rules that cumulatively generates the system's behaviour.

It is clear that there is a lot of room for interpretation, even if only Hollard's properties [22] are considered. Complex systems are studied firstly as concrete examples and then abstracted. The nature of the studied examples varies broadly, and as a consequence so does the definitions in the abstract form. To illustrate this point, Ladyman et el. [23] chronicled the definitions for complexity and complex systems in a single issue of the renowned journal

Science. The issue in question was dedicated to complex systems [24] and contains the following definitions for complexity and complex systems:

1. *"To us, complexity means that we have structure with variations."* [55]
2. *"In one characterization, a complex system is one whose evolution is very sensitive to initial conditions or to small perturbations, one in which the number of independent interacting components is large, or one in which there are multiple pathways by which the system can evolve. Analytical descriptions of such systems typically require nonlinear differential equations. A second characterization is more informal; that is, the system is "complicated" by some subjective judgment and is not amenable to exact description, analytical or otherwise."* [56]
3. "*In a general sense, the adjective "complex" describes a system or component that by design or function or both is difficult to understand and verify. [...] complexity is determined by such factors as the number of components and the intricacy of the interfaces between them, the number and intricacy of conditional branches, the degree of nesting, and the types of data structures*." [57]
4. "*Complexity theory indicates that large populations of units can self-organize into aggregations that generate pattern, store information, and engage in collective decision-making."* [58]
5. "*A complex system is literally one in which there are multiple interactions between many different components.*" [59]
6. "*Common to all studies on complexity are systems with multiple elements adapting or reacting to the pattern these elements create*." [60]

The most eloquent definition found to describe a complex system was penned by Simon in 1962 –

"*Roughly, by a complex system I mean one made up of a large number of parts that interact in a nonsimple way. In such systems, the whole is more than the sum of the parts, not in an ultimate, metaphysical sense, but in the important pragmatic sense that, given the properties of the parts and the laws of their interaction, it is not a trivial matter to infer the properties of the whole.* " [61]

Physicist Nigel Goldenfeld of the University of Illinois is a lot more liberal in his interpretation of complexity - "*Complexity starts when causality breaks down*" [62] This might be true under some extreme interpretations of quantum mechanics, but a system without causality is

logically impossible. The irksome truth of the matter is that the validity of this definition is contingent on what is meant by complexity, and without a broadly accepted definition of complexity, in the general sense, it may indeed be valid.

2.3.7 DISCUSSION

The variance in definitions for technology, innovation and complexity is noteworthy. Consider the aspects of innovation that were presented. The first aspect is the use of analogies to describe progress. These can be useful, but effort should be made to keep the context of the subject matter and the analogy aligned – just because natural evolution can be used as an analogy of technological evolution, does not imply that the underlying workings are the same. Another aspect of innovation that was presented was the recombinant mode of innovation. Recombinant innovation has explanatory power in terms the complexity of inventions and should be kept in mind when designing the methodology. Open Innovation was added to this section to demonstrate that the definition of innovation reaches into the domain of management science.

To limit the scope of interpretation in this study, innovation can be defined as the change in technology over time, including all processes of knowledge flow and actors required to materialise a technological artefact.

This section briefly reviewed TRIZ as a method of problem solving. The idea of abstracting a problem and using template solutions is a viable method of design. The technology trends presented is less convincing. No mention of how these trends were measures could be found in the TRIZ journal [63] – discounting the incessant repetition that they were deduced from “the study of patents” by G.S. Altchuller.

It is worth noting that complexity only makes sense from a systemic viewpoint. It is also true that very few inventions, if any, can be considered as complex systems in and of themselves. They lack several of the key properties of a complex adaptive system. Consider the complex adaptive system property of emergent behaviour. Individual inventions are designed for a specific purpose; any unforeseen or emergent behaviour will most likely decrease the value of the invention. Anticipation of events based on the frequency of prior events is another property of complex adaptive systems. Anticipation by design will be present for a set of inventions, but for the most part it is absent. The use of metrics should distinguish between metrics with the purpose of measuring the complexity of a complex system and the complexity of a single artefact.

2.4 QUANTITATIVE USE OF PATENT DATA IN INNOVATION STUDIES

2.4.1 REVIEW SCOPE

The second research question reads - *How is patent information useful in the study of technology, innovation and complexity?*

The previous section of the literature dealt with the nature of technology, innovation and complexity. In this section the different uses of patent information are explored. This will be addressed under the following themes:

1. The modelling of innovation and technology
2. Patent and technology discovery
3. Patent impact and value modelling
4. Metrics of innovation and technology

2.4.2 THE MODELLING OF INNOVATION AND TECHNOLOGY

2.4.2.1 OVERVIEW

The Fleming and Sorenson [11] study shaped a large part of this research endeavour. An overview of their work was provided in the introductory chapter. More detail is provided in this section.

2.4.2.2 FLEMING & SORENSON'S METHODS AND FINDINGS

Fleming & Sorenson [11] developed a theory of invention by modelling technology as a complex adaptive system. The effect of the number of components (N) making up an invention and their interconnectedness (K) is measured against the usefulness of the invention. Several concepts are incorporated from evolutionary biology and applied to patent information. The concept of fitness in biology describes an organism's success in reproduction and is a measure of its contribution to the species gene pool. In terms of patent information, the amount of citations generated by a patent is interpreted as its fitness. The number of components is measured as the number of sub-classes the invention is classified under. Interdependence, in this study, is defined as the functional sensitivity of an invention in its constituent components. It is measured as the observed ease of recombination for each patent's sub-classes. K is calculated in two parts:

$$Ease\ of\ recombination\ of\ sub-class\ i\ \equiv E_i = \frac{\sum sub-classes\ co-occuring\ with\ i}{\sum previous\ patents\ in\ sub-class\ i}$$

$$Interdependence\ of\ patent\ l\ \equiv\ K_l = \frac{\sum sub-classes\ on\ patent\ l}{\sum_{l \in i} E_i}$$

The propensity to provide citations differs between technology classifications. To account for this the authors introduced statistical control variables, technology mean control (M_i) and technology variance control (V_i). These are calculated as:

$$Average\ citations\ in\ patent\ class\ i\ \equiv\ \mu_i = \frac{\sum_{j \in i} citations_j}{\sum patents\ j\ in\ sub-class\ i}$$

$$Technology\ mean\ control\ patent\ l\ \equiv\ M_l = \frac{p_i}{\mu_i}$$

where p_i is the proportion of patent *l*'s classification members that fall under class *i*.

$$Citation\ variance\ in\ class\ i\ \equiv\ \sigma_i^2 = \frac{\sum_{j \in i}\ (citations_j\ - \mu_i)^2}{\sum\ patents\ j\ in\ sub-class\ i}$$

$$Technology\ variance\ control\ \equiv\ V_l = \frac{p_i}{\sigma_i^2}$$

These variables are calculated and analysed in a Poisson model. The following results were obtained:

1. An intermediate amount of interdependence optimises invention usefulness. At the mean (N = 4.2) this was measured as K = 1.37.

2. The fitness curve has a massive range. It shows that the number and interdependence of the elements recombined can make the difference between an invention being average versus being in the top 6% of successful patents.

3. An increase in the number of components in a recombinant technology leads to increased usefulness, but not to a degree of practical relevance.

4. Increased numbers of components lead to increased citation counts, but this effect diminishes as the number of components increase.

5. An increase in components decreases the certainty of outcomes with highly interdependent components.

6. The effect of interdependence far outweighs the effect of the number of components.

2.4.3 PATENT AND TECHNOLOGY DISCOVERY

2.4.3.1 OVERVIEW

This section aims to demonstrate that patent data can be used to build more advanced search methods to facilitate the discovery of technologies.

2.4.3.2 CONCEPT BASED SEARCH

Montecchi, Russo & Liu [25] provide a comparison between keyword and concept-based search methods as applied to the Cooperative Patent Classification (CPC) scheme. The majority of classification based patent searches are keyword based. Keyword based search methods have the following drawbacks:

1. The levels of detail in patent descriptions vary - There are several reasons for this. Inventors have different writing styles and use different levels of abstraction in their descriptions. Different technical backgrounds and cultural references also influence the language use within a patent. In many cases the content of the patent is purposefully vague to maximise its claims validity.

2. Inaccurate terminology - The meanings assigned to terms and phrases by the patentee do not necessarily subscribe to the dictionary definition of the concept. Non-precise, erroneous or in some instances new terms are used to describe a technology.

3. Lack of standardisation - Developing technologies can have an array of different descriptions with the same underlying meaning. For example, *the use of an electric current carried through ionised air to simultaneously cause a phase change in the electrode and base material causing the base materials to fuse with the deposits from the electrode* could also be described as *arc welding*.

4. Different official languages - Patents are published in multiple languages. In many instances parts of a patent may be translated into English, but neither the existence nor the quality of such translations are certain.

To overcome these drawbacks the use of patent classifications in patent searches is proposed. The Knowledge Organizing Module (KOM) is used throughout the proposed method. KOM is a concept-based and semantic tool used to find patents with relevant

classifications. Figure 5 (Adapted from [25]) presents the generic outline of concept and keyword-based search methods.

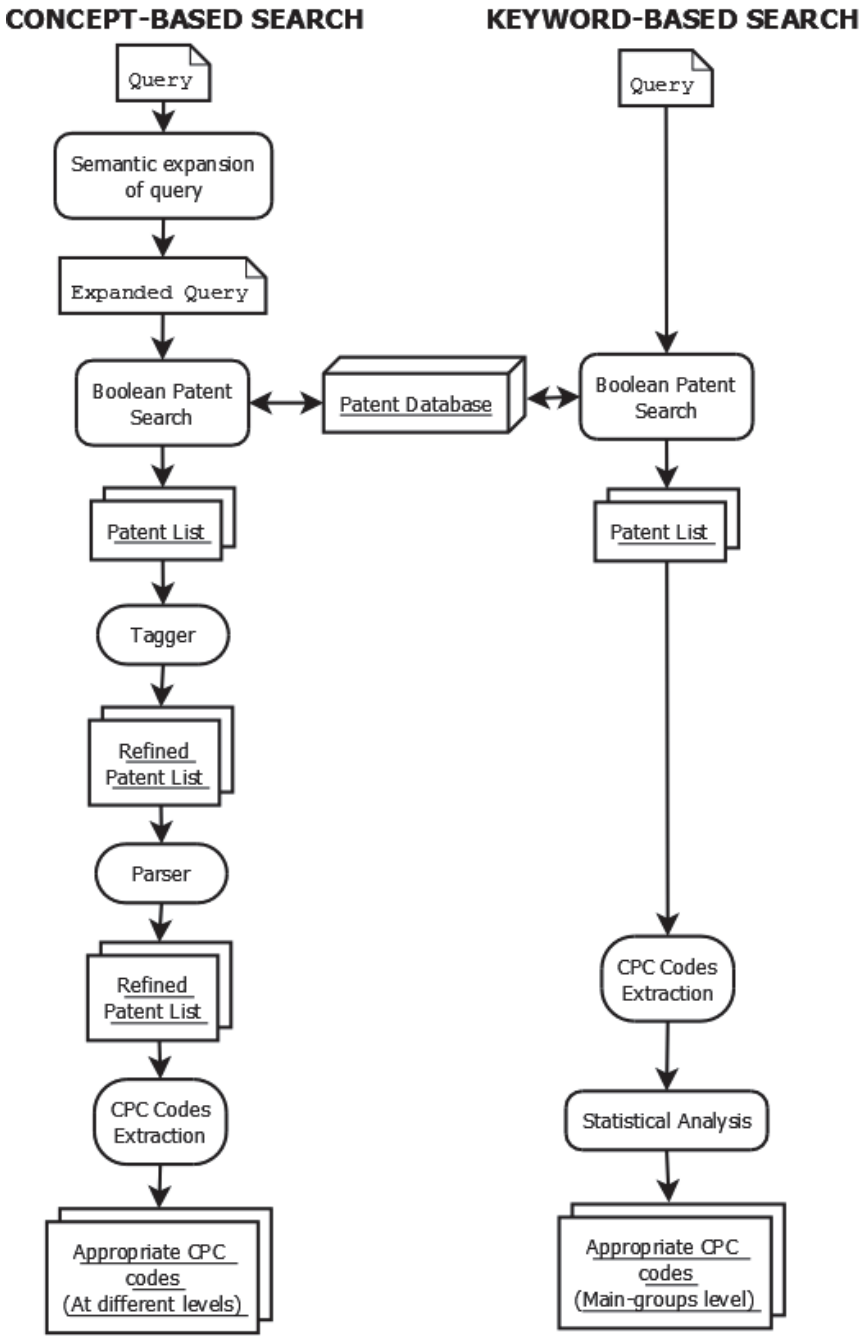


Figure 5: Concept vs. Keyword-Based Search

The concepts and steps listed in Figure **5** are as follows:

1. Semantic expansion of a query - Patent language use differs greatly. To increase the recall of a query knowledge bases such as dictionaries and thesauri are used to generate a list of valid synonyms and coordinate terms of the base query. For example, *lamp* can be expanded as *LED*, *light*, *light bulb* or *incandescent*.

2. Boolean search - The search matches the terms in the query to those in the patent database and retrieves the ones that match. A compelling argument is made for the use of the entire patent in this step, as opposed to only using the title and abstract. In a case study the authors found almost double the amount of relevant patent families when full text was used.

3. POS Tagger - A part of speech tagger is used to disambiguate the results found during the Boolean search. Consider the search term LED. This can either be interpreted as the past tense verb of lead or as the noun acronym of *light-emitting diode*. A POS tagger is useful way to discard all patents found by the Boolean search where led is used as a verb. In the accompanying case study this step increased the accuracy of the search from by 20%.

4. Parser - A parser is used to identify the semantic relationships between words within the search results. Patents that contain the correct keywords, but not the desired relationship between them, are eliminated from the answer set. In the accompanying case study this step was able to eliminate all false positive cases. It should be noted that a parser can misinterpret some of the relationships within valid results and discard them. It is thus a risk to the recall of the system.

5. CPC codes extraction - The classifications of the remaining patents are extracted and presented at the predetermined hierarchal level. The resulting list of classifications are compared to both a manually derived list and classification searches produced by Classification Search (EPO), PatentScope(WIPO) and IPCCAT(WIPO). Concept based searches are shown to have a significantly higher accuracy and recall than keyword-based approaches.

This study sets forth several useful concepts for patent searching and technology discovery. Patent classification codes can be seen as an index of existing technology. The study mentions several knowledge bases that were used in the semantic expansion of the query, but did not elaborate on them. Another shortcoming is the lack of information regarding

system performance. POS Tagging and parsing algorithms are generally resource intensive and applying them on a per search basis can possibly lead to the development of impractical system requirements.

2.4.3.3 FUNCTION/BEHAVIOUR-ORIENTED SEARCH

Montecchi & Russo [26] proposed the Function/Behaviour-Orientated (FBS) system framework to assist in technology discovery and transfer. The system is used to search for technologies within patents at various levels of abstraction.

In the FSB structure a system is decomposed into several abstract parts. These are:

1. Function (F) - The function of a technical system is the motivation/purpose for its existence. In general a system has main and peripheral functions.

2. Behaviour (B) - The behaviour of a function is a sequential change of states in a system to achieve the purpose expressed in its function.

3. Structure (S) - A description of a system's components and their relationships to one another.

4. Physical Effect (PE) - Physical Effect is an intermediary level between behaviour and structure. A physical phenomenon is described as the cause of a transition between two states. Physical Effects are the natural laws governing these transitions.

A static list of nouns, verbs, adverbs and adjectives is compiled for every class of PE. At the highest level interactions are classified as mechanical, acoustical, thermal, chemical, electrical, electromagnetic or biological. Subclasses are assigned to each main class of PE. For example, *compression* is an instance of a mechanical PE and will contain keywords such as *pressure, compression, press* and *compressible* as well as measurement units related to pressure such as Pa, bar, atm and psi.

IPC/CPC codes are used as a constraint in the first part of the search methodology, firstly incorporating the full length code and then moving up the classification hierarchy to include adjacent technologies.

Objects are abstracted by replacing them with their associated functions. For example, when looking for the state of the art in nut crackers the device itself is omitted from the search and replaced by the function *to crack*.

To assist in the abstraction of objects Hypernymy is employed to find more general terms and Meronymy is used to find more general descriptions of the function. This, along with other semantic relations, is extracted from the WordNet database. Figure 6 shows a partial result for a FSB search on nut cracking methods.

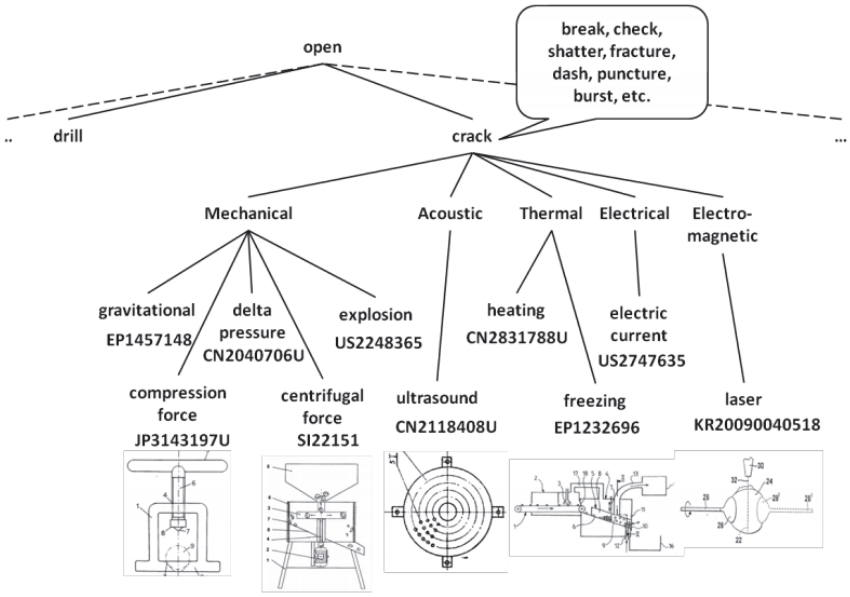


Figure 6: FSB search result for 'nut cracking' [64]

The system proposed in this study is still under development. As a result its description is quite abstract, but the methodology shows promise.

2.4.4 PATENT IMPACT AND VALUE MODELLING

2.4.4.1 OVERVIEW

Patent documents do not contain any direct measures of their value. As a result, the valuation of patents is difficult. Several researchers have attempted to overcome this by trying to infer the value of the patent from its data. These are presented in this section.

2.4.4.2 THE VALUE OF CITATION

Hall et al. [27] studied citations in patents published between 1963 and 1999. They found that each citation raises a patents market value by an average of 3%. Further findings indicate that unpredictable or counterintuitive citations are more valuable and that self-citation (i.e., patents owned by the same company) is more valuable than external citation.

2.4.4.3 TRIZ BASED FUTURE VALUE ESTIMATION

Park, Ree and Kim [28] proposed Subject-Object-Action (SOA) based text mining as a method to identify promising patents for technology transfers using TRIZ evolution trends. TRIZ evolution trends (described earlier) is a TRIZ tool that represents a specific sequence of technological transitions which indicate how a system evolves over time. The study differs from several similar studies by giving different weights to the TRIZ evolution rules based on the life-cycle stage of the technology. Their method is as follows:

1. Download an entire technology section, based on IPC/CPC classification codes.
2. Identify the life-cycle stage of the technology by analysing the cumulative number of patents filed pertaining to the selected technology. The number of patent filings generally increases over time in an S-shaped curve.
3. SAO language structures are extracted from the raw patent data. Their methodology section suggests using the Stanford Parser, MiniPar or Knowledgist.
4. The TRIZ trend and trend phase are identified by analysing the overlap between words contained in the subject of a sentence and action-object couplings.
5. Two rule bases are employed to identify TRIZ trends. The first is a *Reason for Jump* (RFJ) rule base. A jump in this instance refers to a simultaneous increase in benefits and reduction of drawbacks in a technology. The rule base consists of a set of Verb-

Noun or Verb-Noun phrase combinations indicative of technological evolution. The second rule base is a set of hints (Noun phrases) that indicate a technology forms part of a TRIZ trend. For example, *shape-memory alloy* is indicative of the *smart materials* trend.

6. The rule bases and the extracted SAO structures are measured with Simpson & Dao's [65] measure for sentence similarity.

7. Promising patents are identified. A patent is deemed a future high value patent if it is related to future important TRIZ trends in the technology domain and is in a trend phase higher than the average of the domain. One point is allocated to a patent if it is related to a future important TRIZ trend and one if one point is allocated if the patent's trend phase is one level higher than the domain average.

The method was applied to 92 patents on wind turbine technology. Several patents with potential future value were identified and presented.

2.4.5 METRICS OF TECHNOLOGY AND INNOVATION

2.4.5.1 OVERVIEW

Several studies present methods of measuring technology and innovation metrics without necessarily prescribing a use for the measured information. This section aims to present an understanding of these metrics.

It also catalogues several measures used to describe technology and innovation.

2.4.5.2 TECHNOLOGICAL SIMILARITY

Aharonson & Schilling [66] describes 11 methods with which the distance between technologies is measured. These methods are summarised below:

Technology Distance - This measurement is a geodesic measurement between technology vectors based on subclass classification vectors. It is suited to measure the technology readiness of firms as well as the proximity of an institution to a technology. This form of measurement can be used to create visual representations of a technology landscape. Drawbacks of this method are its computational intensity and propensity to inaccuracy when working with newly defined classification classes. The methodology of determining the technology distance is fairly simplistic. Two patents that share a classification is deemed

connected. On an industry level the brute amount of connections determines the distance between two sectors.

Technological Footprint - This method is an aggregation of technological positions a firm is in. It is used to characterise a firm's technology profile, mostly in the context of finding opportunities or assessing risks from competitors. As this method gives insight into the nature and placement of the firm, as opposed to its technology, it is disqualified as a usable method in this study. Several other measures in Aharonson & Schilling's study provide insight into the position of a firm and not of its technology. These are precluded from this literature review.

Technology Lifecycles - A technology lifecycle, as a metric, is a measurement of a number of patents within a technology position over time. This method can be used to observe the growth of a technological field and the ease with which spillovers occur within that sector. A visual interpretation of the technological landscape can be produced with this method.

Overlap in Citations - This method compares the amount of shared citations in patents from two arbitrary fields. It is one method used to trace the evolution of a specific technological entity. One problem with this measure is the influence of firm's patenting strategy [67]. These patent strategies make citations less indicative of knowledge flows.

Count of overlapping patent classes - This method is used to capture the proximity and similarity of technologies present within patents. Although it is widely used its output is course as it does not take into account the interdependencies of technology components.

Text Analysis - This method, or rather collection of methods, uses various levels of text processing to determine the similarity between technologies. It provides a far richer analysis than patent classification metrics, but requires a more advanced system to implement.

2.4.6 DISCUSSION

The ideas presented in the methodology, especially those regarding the extraction of data from patents, hold some merit and can possibly be used in this study. Some questions around their methodology persist. The first of these is the general vagueness around the TRIZ rule bases. The idea of using a set of phrases as rules can be a valid patent processing approach, but here the actual parameters and nature of this central part of their study is not divulged. The second major shortfall is the lack of testable results. Several patents are presented with possible future value, but there is no way of confirming this. This brings into question the actual value of the method and study and a retrospective study would have been more

prudent. Thirdly, the choice of technology rigs the study by design. All instances of wind turbine technology is classifiable under the TRIZ trend of renewable energy, therefore the technology domain is bound to deliver more positive results.

2.5 MEASURING COMPLEXITY IN INVENTIONS

Luo and Wood [29] investigated the changes in the complexity of the inventive process. They devised a set of 7 metrics from patent data and used it to characterise the inventive process. These were used to analyse trends in the innovation process between 1975 and 2011. The metrics and the results from the study are listed below.

1. The number of inventors per patent. This metric is used as an indicator of the extent of collaboration and coordination effects. It was found that the number of inventors per patent steadily increased from an average of 1.66 to 2.61 over the measurement period. This represents an increase of 57%.

2. The number of regions of co-inventors per patent. This metric is used as an indicator for the sophistication of coordination in remote collaboration. The rational for this metric is that the geographical dispersion of inventors would increase the complexity of interactions among them. Regions were defined as either a state in the USA or a country outside the USA. It was found that the number of patents with collaborators from more than one region gradually increased from 4.5% in 1975 to 15% in 2011.

3. Patents granted to individuals versus organisations. This metric is used as an indicator of the sophistication of organisational methods to integrate resources, teams and expertise. The rational for this metric is that individual inventors do not have the same amount of resources, both capital and knowledge, that an organisation can attain. It follows that the complexity of inventions attributed to individual inventors would be less complex than those assigned to companies. It was found that the amount of patents granted to individuals were almost stagnant over the measurement period. Dramatic growth was measured in patents granted to organisations.

4. The number of references to earlier patents. This metric is used to determine the knowledge base of existing technologies and known solutions. It is used in tandem with the amount of normalised technology classes in a patent. Measurements indicated growth at an increasing rate of the average number of references per patent over the measurement period. This supports the assertion that new inventions are based on an increasing amount of prior knowledge and solutions.

5. The normalised count of technology classes assigned to a patent. This metric is used to indicate the sophistication of the coupling of knowledge domains which an invention integrates. Normalisation is done in the same way as in the Fleming and Sorenson study [11]. It was found that the coupling between classes assigned to patents increased steadily from an average of 0.4 to 0.95 between 1975 and 2011. This suggests an increase in knowledge coupling in inventions over time.

6. The percentage of "system" patents. This metric is used to indicate the systemic nature in innovation. Patents are classified by analysing their titles. If the word "system[s]" appears in the title, it is seen as a systematic combination of technologies. Remaining patents are classified as stand-alone units of technology. Patents counted as "system" patents increased from about 6% in 1975 to about 18% in 2011. Most of the growth occurred after 1990. In addition to this it was found that "system" patents cite a greater number of references than stand-alone inventions.

7. Patents per inventor per year. This metric is used to indicate individual invention productivity. The number of patents per inventor decreased steadily during the measurement period.

The authors note the limitations of these metrics. Patents are seen as proxies for inventions, and measurements based on patent data can be skewed by the patenting behaviour and strategies of firms. Synthesising the results indicates an increase in the complexity of the invention process.

2.6 PATENT DOCUMENT PROCESSING

2.6.1 REVIEW SCOPE

Research question 1.4 reads - *how can identified measures of complexity be extracted from patent documentation?* Previous sections had shown that virtually all the methods used to extract data from patents use some sort of textual processing. This section will present the different methods and tools available for this task under the following themes:

3. Information Retrieval
4. Natural Language Processing Methods and Capabilities
5. Language Resources and Tools

2.6.2 **Information Retrieval** [68]

Information Retrieval (IR) can be defined as finding material (usually documents) of an unstructured nature (usually text) that specifies an information need from within large collections. Figure **7** illustrates the process behind the generation of a query for IR and possible sources of noise in its generation. A user task leads to a specific information need. This is expressed in the form of a query that is then compared to a large data set. During this process ignorance on a subject can cause a user to misunderstand his/her information need. Another problem that can occur is the misformulation of a query.

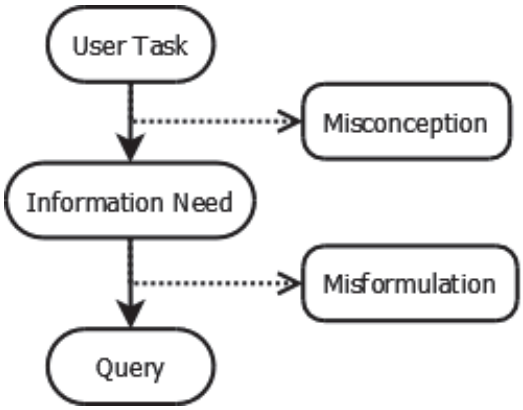


Figure 7 : Information Retrieval Noise Sources

Two metrics are used to measure the success of an information retrieval system:

Precision - The fraction of documents retrieved that are relevant to the user's information need.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{\{retrieved\ documents\}}$$

Recall - The number of documents that are relevant to the query that were successfully retrieved.

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{\{relevant\ \ documents\}}$$

F-score - Precision and recall represented in a single unit. The F-score is a weighted harmonic mean of precision and recall.

$$F_\beta = \frac{(1+\beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$

The simplest form of IR can be achieved through an exhaustive search, where the query is matched directly to every part of every document in the corpora. This approach has several drawbacks:

6. It is slow, especially on large corpora.
7. NOT queries become non-trivial and are badly handled.
8. Other operations (like finding two words in close proximity) become unfeasible.
9. Ranked retrieval is not possible with this method.

A more practical approach is the use of Term-Document Incidence Matrices. This is a matrix with terms as rows and documents as columns. Table 7 shows an example. These structures provide functionality that an exhaustive search cannot, but there is a drawback. Consider a very large corpus with N = 10^6 documents each containing about 10^3 words using an average 6 bytes/word. Storing the corpus would then require about 6 GB storage capacity. If the corpus contains M = $5 \cdot 10^5$ distinct terms the matrix would consist of $5 \cdot 10^{11}$ Boolean entries requiring at least 62.5 GB of storage space (not counting any database overheads). Another problem with this approach is that word order and occurrence frequency is not considered in the retrieval process.

	Doc1	Doc2	Doc3
Term1	1	0	1
Term2	1	1	0
Term3	0	0	1

Table 7: Term-Document Incidence Matrix

To solve the problem of the massive storage space requirement term-document relations are stored in an Inverted Index. This data structure is used in many modern applications, including search engines.

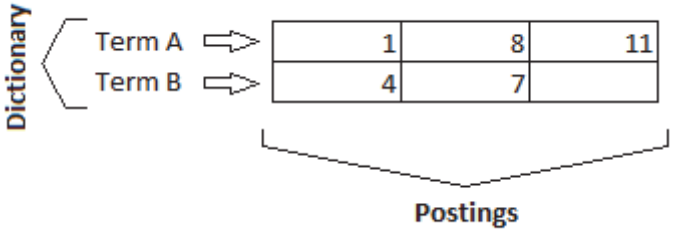


Figure 8: Inverted Index

Although this solves the storage requirement problem it is still a Boolean retrieval method and does not consider word order or frequency. To address this several Ranked Retrieval Methods have been developed. The premise is to score the relation between a query and a document and to present the query results in the scored order. The score is usually a number between 0 and 1.

One method used to rank document-query pairs is the Jaccard coefficient. This is given by:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

A problem with this measure is that it does not take into account term frequency. In many cases rare items in a collection are more informative than common ones. Furthermore, the Jaccard coefficient is bias toward short documents. A better normalisation of document length is thus required.

Term Frequency Weighting is based on the premise that a document's relevance increases with the amount of term mentions, but not linearly. When this method is used the term count in every document is indexed. This is also known as a Bag of Words model. Scoring query-document pairs in this way can be done as follows:

$$\omega_{t,d} = \begin{cases} 1+\log_{10} tf_{t,d} & if\ \ tf_{td}>0 \\ 0 & otherwise \end{cases} \quad \text{and} \quad Score = \sum_{t \in q \cap d} \omega_{t,d}$$

Where $\omega_{t,d}$ - Log-term frequency of term *t* in document *d*

$tf_{t,d}$ - Term frequency of term *t* in document *d*

Inverse Document Frequency Weighting is based on the premise that rare terms are more useful than common ones when retrieving information. If a rare term such as *reciprocornous* or *idioticon* occurs only once in a document and query it is quite probable that the document will be relevant. The document frequency of a term is a the total amount of documents in a corpus that contains that term. The inverse document frequency of a term can be calculated as follows:

$$idf_t = \log_{10} \frac{N}{df_t}$$

where idf_t - Inverse document frequency of a term

N - Total amount of documents

df_t - documents containing term *t*

One of the best methods for term weighting is a combination of term frequency and inverse document frequency weighting. This is known as TF-IDF weighting and can be calculated as:

$$\omega_{t,d} = (1 + \log_{10} tf_{t,d}) \cdot \log_{10} \frac{N}{df_t}$$

A query- document pair score can thus be calculated as follows:

$$Score(q,d) = \sum_{t \in q \cap d} tf \cdot idf_{t,d}$$

The different forms of indexing presented here can all be classified as Latent Semantic Analysis methods as semantic aspects within the text corpus is presented in the indexing scheme.

2.6.3 NATURAL LANGUAGE PROCESSING METHODS

2.6.3.1 TOKENISATION

Tokenisation is the process of segmenting running text into useful linguistic parts such as sentences, words, punctuation and numbers. This task can be more complex than it initially appears. An example of this is the use of full stops which can either indicate the end of a sentence or an abbreviation. Although these hurdles complicate the process adequate results can be achieved through standard white space tokenisation. [30]

Different tokenisation schemes exist. It has been shown that the scheme selection can have an impact on following processing tasks. This was demonstrated with text classification where alphanumeric and alphabetic tokenisation yielded different results in different knowledge domains. With alphabetic tokenisation terms such as '1st' and '2nd' will be presented as 'st' and 'nd' while they will stay intact with an alphanumeric scheme. Alphabetic tokenisation yields better classification results with e-mail document analysis while alphanumeric tokenisation works better with news articles. [31]

No literature could be found on optimal tokenisation schemes for the processing of patent data.

2.6.3.2 STEMMING AND LEMMATISATION

Stemming is the process of presenting a word (or token) in a standardised morphological form. It is done by stripping away affixes from word stems. One of the most popular methods of achieving this is Porter stemmer. As an example a stemmer will convert *automaton, automotive* and *automatic* to *automat*. In some instances a simplified morphological query can increase accuracy and recall in information retrieval activities. In most, however, a simplified morphological structure diminishes the system capabilities. There are three reasons for this. Firstly, stemming results in information loss. Analysis on relations between words become impossible. For example in *operating a tractor* the word *operating* can only be classified as a verb. In *operating system* it is a noun and in *operating costs* it is an adjective. Secondly, meaningful multi-token language structures are lost. Many statistical NLP methods can be made more efficient by regarding frequently occurring chunks of words as single

distinctive tokens. Thirdly, stemming only has potential benefit in morphologically poor languages like English. [30]

A second option for presenting words in a standardised morphological form is lemmatisation. With lemmatisation a full morphological analysis of the word is required. It is then matched to its corresponding dictionary root word or lemma. A lemmatiser will only remove inflection. As an example *automating* and *automation* will be transformed into *automate*. Lemmatising and stemming show little difference in performance when used in applications such as Information Retrieval. [30]

In some instances the inclusion of stemming in the pre-processing of patent information might be beneficial. Chen and Chang [69] used stemming in their patent classification algorithm. Patents were automatically classified into IPC patent subgroups (approximately 72 000 subgroups exists) with an accuracy of 36.07%. Eisinger et al. [70] criticized their approach, at least in part, for its choice of pre-processing methods. Gomez and Moens [71] surveyed 16 automated patent classifiers. Seven of these used stemming in their pre-processing.

In conclusion the shortcomings of stemming and lemmatisation are abundantly clear. There are however cases in this study where lemmatisation might be useful. One such would be to simplify component names that occur in plural form within the patent text.

2.6.3.3 MAXIMUM ENTROPY MODELS [72]

Many problems in NLP are at heart statistical classification problems. In many tasks the probability that a class *a* and context *b* co-occur or are related needs to be calculated. Large text corpuses give incomplete insight into p(*a,b*) because words/tokens in b is sparse compared to the corpus size and no practical corpus can give insight into all possible configurations of (a,b).

Maximum entropy modelling provides a reliable estimate for p(a,b) with sparse evidence for the relation between a and b. The principle is demonstrated in the formula below where *A* denotes the set of possible classes and *B* denotes the set of possible contexts.

$$H(p) = -\sum_{x \in E} p(x) \log p(x)$$

where $x = (a, b), a \in A, b \in B, E = A \times B$

The probability p should maximise the entropy H and should remain consistent with the parameters of the partial information provided.

The aim of this study is not to directly implement this method, but rather to use tools such as the Stanford Parser (discussed later) that incorporates it. It is provided here to demonstrate that many of the methods employed in NLP is statistical in nature and as such fallible in at least some portion of use-cases.

2.6.3.4 PARTS OF SPEECH TAGGING [32]

A Part of Speech tagger (POS tagger) is a piece of software that assigns a part of speech (syntactic class) to every token in a phrase or sentence. As an example consider the following sentence with tagged parts of speech:

The-DT glass-NN shattered-VBD loudly-RB.

Several methods have been implemented and refined from the early 1960s onward. Modern approaches include statistical methods, classification based methods, transformation based methods, manually built rule sets and combinations of the forenamed. Several of the statistical approaches are shown below.

2.6.3.5 HIDDEN MARKOV MODELS (HMMS) IN POS TAGGERS

The application of HMMs on POS tagging is generally implemented as follows. Given a sequence of words $W = w_1 \dots w_k$, the task is to find the most likely sequence of tags $T = t_1 \dots t_k$ such that $\underset{T}{argmax}\ P(T|W)$. Different variations exist in the application of HMMs with some requiring supervised training. In 2000 Brants [73] demonstrated the application of this model on Wall Street Journal data with a correct classification rate of 96.7%

2.6.3.6 MAXIMUM ENTROPY MODELS IN POS TAGGERS

As discussed in a previous section the general assumption of a maximum entropy model is that the probability distribution should be uniform if no knowledge exists about it with uniformity being equivalent to having the maximum possible entropy. In a practical POS tagger this idea is implemented as shown below.

Consider the probability distribution p(t|c) where c presents a context and t presents a tag. These are presented in binary indicator functions called feature functions, with

$$f(c,t) = \begin{cases} 1 & if\ t = T\ and\ c\ matches\ C \\ 0 & Otherwise \end{cases}$$

The value assigned to each feature is constrained through the application of training data, such that

$$\hat{p}(f) = \sum_{c,t} \hat{p}\,(c)p(t|c)f(c,t)$$

where $\hat{p}(f)$ and $\hat{p}(c)$ are the observed probabilities from the training corpus and $p(t|c)$ is the parameter that needs to be estimated.

The training cost for this model is higher than that of HMMs, but the expression of dependencies become easier. In 2003 Toutanova et al. [74] reported accuracies of up to 97.24%.

2.6.3.7 DISCUSSION

Although several other methods have made contributions to POS tagging it is not crucial to the scope of this project to investigate all of them as a suited software tool will be used and the project does not require a POS tagger to be built from first principles.

It should be noted that even though high accuracies were reported by many researches the sample data they worked on differed vastly. This implies that the same results might not be attainable for patent data. The main argument for this is that patent language is complex, purposefully vague at times and sentences are longer than those found in an average corpus.

2.6.3.8 PARSING [30]

Parsing is the practice of inducing a structure of higher level units on an arbitrary sentence. These structures encapsulate the interpretation of a chunk of words. Most forms of parsing depend on a training set as most sentences can be parsed in multiple ways. These are mostly nonsensical to human readers and require a statistical example set to remove ambiguity.

Consider the sentence parse trees shown in Figure 9. The sentence *"Fruit flies like bananas"* is intended to be interpreted as fruit flies (insect species) showing a particular affinity towards bananas. The Stanford parser deduced another possible meaning - that fruit and bananas share the property of being able to fly.

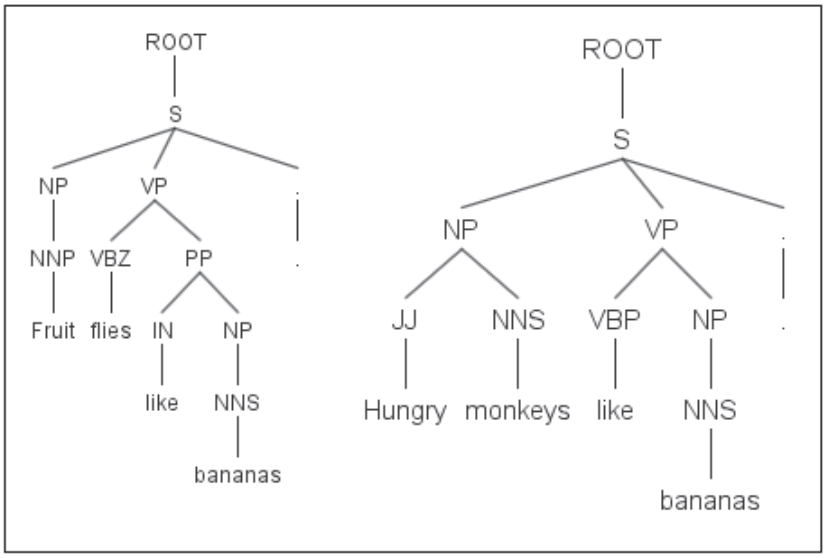


Figure 9: Stanford Parser Sentence Trees

Multiple meanings hold various implications for patent processing. As seen in previous sections many methods use at least some aspect of parsing to extract technological properties from patent text. These are extracted either directly from a sequence of tagged words or from parsed phrases. In many cases the relations are stored as object-action pairs for technology searches. Figure 10 gives a crude interpretation of the ambiguous sentence in the previous figure.

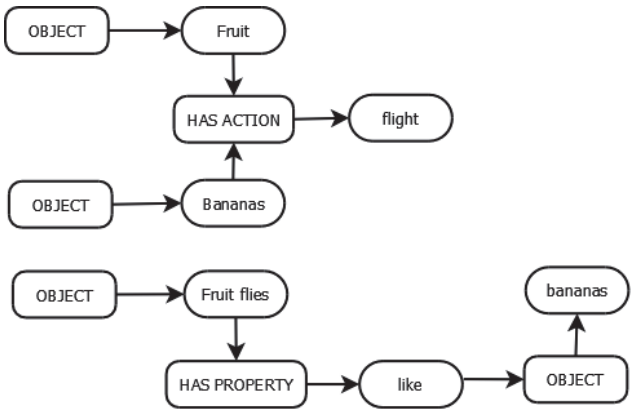


Figure 10: Object-Action Interpretation of parsed sentence

It is clear that this method needs to be implemented with care and that a test protocol needs to be set up to verify the validity of results produced by patent parsing.

2.6.4 LANGUAGE RESOURCES AND TOOLS

2.6.4.1 SYNTACTIC AND SEMANTIC TREEBANKS [75]

A Treebank is annotated corpus of text that specifies the syntactic and/or semantic relationships present therein. The corpuses are applied as training data for a variety of NLP related tasks.

The PENN Treebank and its associated annotation methodology have been adopted as the industry standard for most modern applications, including the Stanford parser (described later). It contains approximately 7 million words of part-of-speech tagged text. The annotated texts are diverse and include IBM computer manuals, nursing notes, Wall Street Journal articles and transcribed telephone calls.

Table 8 and Table 9 show the syntactic and semantic tags used in the PENN Treebank, respectively.

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Table 8: Penn Treebank part-of-speech tags (Semantic features)

Tag	Description
ADJP	Adjective phrase
ADVP	Adverb phrase
NP	Noun phrase
PP	Prepositional phrase
S	Simple declarative clause
SBAR	Subordinate clause
SBARQ	Direct question introduced by wh-element
SINV	Declarative sentence with subject-aux inversion
SQ	Yes/no questions and subconstituent of SBARQ excluding wh-element
VP	Verb phrase
WHADVP	Wh-adverb phrase
WHNP	Wh-noun phrase
WHPP	Wh-prepositional phrase
X	Constituent of unknown or uncertain category
*	“Understood” subject of infinitive or imperative
0	Zero variant of that in subordinate clauses
T	Trace of wh-Constituent

Table 9: Penn Treebank phrase tags (syntactic features)

This study makes use of tools trained on the annotated texts in a variety of Treebanks. It is not necessary to investigate the complete nature of these structures to use the tools built on it. The previous section focused on the syntax used within most of the Treebanks as it is the same standard that is used in the output of tools used in this study. The syntax is provided in the preceding two tables and an understanding thereof is crucial for the interpretation of tool output.

2.6.4.2 STANFORD CORENLP [33]

The Stanford CoreNLP toolkit is an extensive pipeline with broad natural language analysis capabilities. The tool is one of the most comprehensive analytics tools freely available. It is fed a piece of text and produces the annotated equivalent. The output can be set to XML format if needed. Figure 11 shows a simplified workflow of the tool.

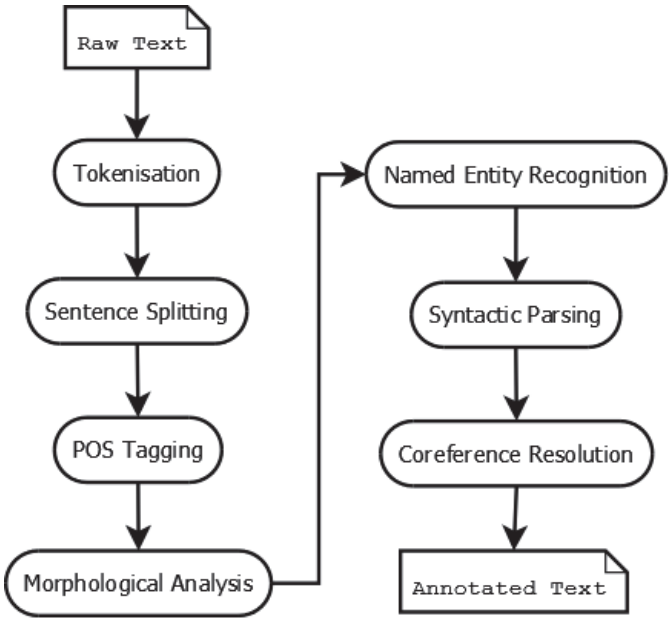


Figure 11: Stanford CoreNLP Workflow

The tasks done by the CoreNLP tool is discussed in previous sections and so for brevity it is not repeated here. The imperfections of the various methods have also been discussed in earlier sections and will not be repeated here.

It should be noted that the tool can either be used via a stand-alone interface, or it can be executed via command line. As such it can be fully integrated with another platform.

2.6.4.3 WORDNET [34]

WordNet is a lexical database for the English language. It contains entries for most nouns, adjectives, adverbs and verbs. The tool also contains a fair amount of relationships between words. Consider the output from the WordNet interface [76] in Figure 12. It shows the different used for the word “complex”. Clear distinction is made between noun and adjective use cases.

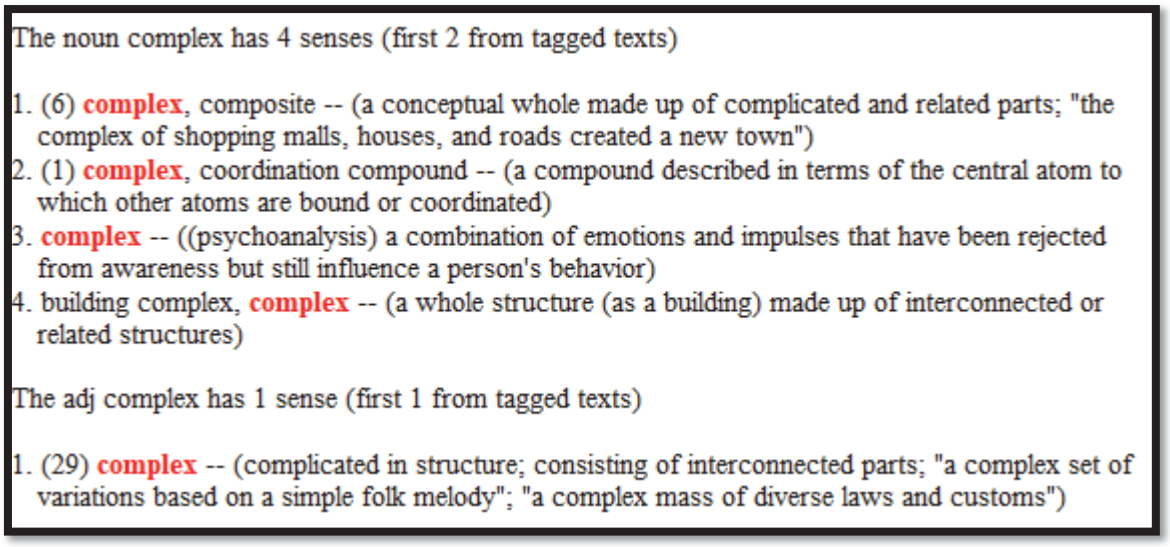


Figure 12: WordNet output for "complex" [76]

The tool introduces the idea of a Hyponym. If a word or phrase is a conceptual instance of another word it is deemed a Hyponym. Figure 13 shows some examples for the fourth noun sense of the word "complex"

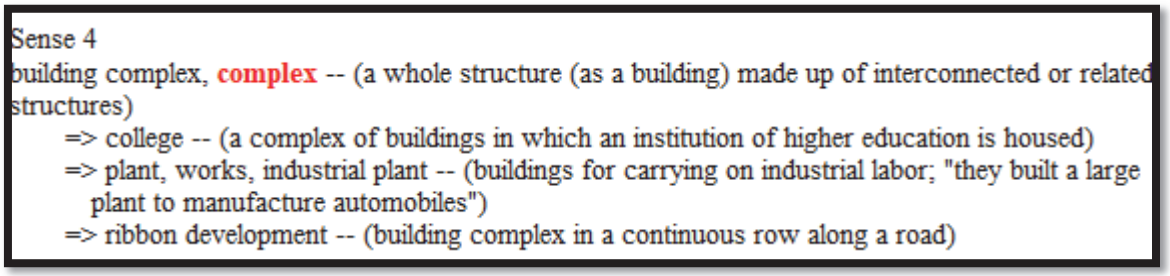


Figure 13: WordNet Hyponyms for "complex" [76]

Another relationship mapped in the tool is termed a Meronym. If a word or a phrase is a part of an object, represented by a word or phrase, it is its Meronym. The example in Figure 14 illustrates the concept.

Sense 4
building complex, **complex** -- (a whole structure (as a building) made up of interconnected or related structures)
=> structure, construction -- (a thing constructed; a complex entity constructed of many parts; "the structure consisted of a series of arches"; "she wore her hair in an amazing construction of whirls and ribbons")
HAS PART: foundation, base, fundament, foot, groundwork, substructure, understructure -- (lowest support of a structure; "it was built on a base of solid rock"; "he stood at the foot of the tower")
HAS PART: plate -- (structural member consisting of a horizontal beam that provides bearing and anchorage)
HAS PART: structural member -- (support that is a constituent part of any structure or building)

Figure 14: WordNet Meronyms for "complex" [76]

Hypernyms are closely related to Hyponyms. Hyponyms describe child instances of a concept. Hypernyms describe the parent concepts of which the concept in question is an instance. A Hypernym tree can be constructed that shows a hierarchal classification of a concept and the various levels of abstraction that can be applied to it.

Sense 4
building complex, **complex** -- (a whole structure (as a building) made up of interconnected or related structures)
=> structure, construction -- (a thing constructed; a complex entity constructed of many parts; "the structure consisted of a series of arches"; "she wore her hair in an amazing construction of whirls and ribbons")
=> artifact, artefact -- (a man-made object taken as a whole)
=> whole, unit -- (an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit")
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> physical entity -- (an entity that has physical existence)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Figure 15: WordNet Hypernyms for "complex" [76]

3 SYNTHESIS

3.1 INTRODUCTION

In this chapter the concepts that were explored in the Literature Study are applied and expanded. The purpose for this is to create a holistic understanding of the problem and the textual materials involved before venturing into design activities.

This chapter is primarily focused on analysing the utility of XML tags and their relation to patent components, but also includes a field test of natural language processing tools and general contextualisation of challenges and solutions.

3.2 A DEFINITION OF COMPLICATEDNESS

It was shown that the literature does not provide a cohesive, generally accepted definition for complexity. Most of the available definitions focus on complex systems. Some of the properties of these is emergent behaviour and anticipation. These cannot be ascribed to patented inventions, as they are made for a known purpose and emergent behaviour would only impair functionality. Most inventions are not capable of anticipation in any form.

It is also true that inventions differ in the amount of effort required to understand, utilise and materialise them. For this reason, a measure of complexity can be ascribed to them that does not fall under the general definition of systemic complexity.

To produce such a definition MECE logical expansion will be used [77]. This method consists of the hierarchical decomposition of a concept. Child objects should be collectively exhaustive in respect to their parent (∑children = parent) and mutually exclusive with respect to each other.

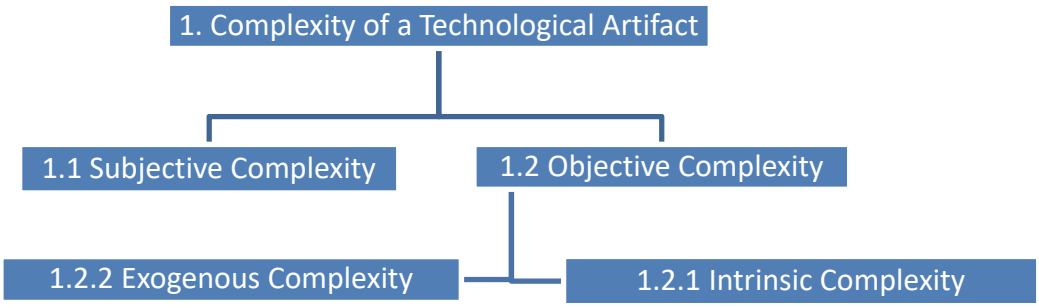


Figure 16: Complexity MECE Tree

The first distinction to be made when describing the complexity of a technological artefact is between objective and subjective complexity. Subjective complexity refers to the capacity of a technology creator, or consumer's ability to comprehend and utilise an artefact. By its nature it is difficult to measure, but its importance should not be underestimated. For example, if a new commercial software product is launched the user experience and learning curve will have a direct effect on its adoption at large. This in turn influences further development and affects the technological ecosystem at large.

In contrast, objective complexity is more directly measurable and refers to the nature of the object. The list below shows one possible expansion of objective complexity -

1. Exogenous complexity
    a. Market forces
    b. Political, social, legal, and moral forces
2. Intrinsic Complexity
    a. Structural
        i. Fixed state (e.g. A matchbox)
        ii. Multi-state
            1. Discrete (e.g. A light switch)
            2. Continuous (e.g. A spring)
    b. Internal Interaction
        i. Mechanical
            1. Statically fixed (e.g. Hammer head and shaft)
            2. Dynamic
                a. Permanent coupling (e.g. Cogs in watch)
                b. State based coupling (e.g. A gearbox)
        ii. Thermal
            1. Passive (e.g. A heat sink)
            2. Active (e.g. An air-conditioning unit)
        iii. Acoustic (e.g. Sonar)
        iv. Chemical (e.g. Electrochemical cell)
        v. Electrical (e.g. Microprocessor)
        vi. Gravitational (e.g. Hydro-electrical storage schemes)
        vii. Electromagnetic (e.g. Electric motors)

Interaction and structural complexity can be measured from patent information, when looking at part references and textual descriptions in patents. The literature has shown attempts to

extract more aggregated levels of interaction and interaction types from patent data, but it is difficult to use these to construct a cohesive complexity measure. For this reason, the number of parts and part interactions will be used as proxies for the complicatedness of patented inventions.

As a matter of convention, any reference to complexity from this point forward will reflect this definition.

3.3 MANUAL ANALYSIS OF PATENT PUBLICATIONS

Several manual and computer-assisted analyses were done on a sample set of 15 patent documents. These documents were chosen from a larger pool to represent a diversity of approaches in terms of component naming schemes, general document style and invention complexity.

Publication Number	Title
US8927141	Rechargeable battery
US8951028	Rotary machine of the deformable rhombus type comprising an improved transmission mechanism
US8943365	Computer program product for handling communication link problems between a first communication means and a second communication means
US8932884	Process environment variation evaluation
US8925121	Method and system for rapid and controlled elevation of a raisable floor for pools
US8925207	Axe
US8925122	Fully articulable shower curtain rod

Table 10: List of patents for manual analysis

3.3.1 COMPONENT NAME PLACEMENT

This analysis focused on the placement of XML tags within the sample patent documents and how these relate to component names and properties within the text. This analysis is done on a purely observational basis. Every sample patent contains a sizeable amount of parts and those parts are mentioned several times. It is therefore impractical to list every occurrence of a component name and how it correlates to its surrounding XML tags. Instead an approach was followed whereby the patent text was meticulously processed and every distinct way in which a tag could relate to a part name was documented.

It should be noted that all types of tags are used in this analysis. Some of the simplest formatting tags have proven to the most useful when extracting component names. As an example, consider the (embolden) tag. These are used to highlight components in the text that also features in the patent drawings. These tags are sparsely used in any other context and cases where numbers are highlighted that do not pertain to component numbers are rarer still. It thus becomes a powerful tool help extract the names and properties of components.

3.3.1.1 FORMATTING TAG PLACEMENT IN RELATION TO COMPONENT NAMES

Consider the following patent excerpts:

1. ...a stator 2...
2. ...uncoated regions 11a and 12a
3. Each end 27, 28
4. flotation bag (344)
5. The rolling bodies 16, 19, which can
   ... take the references 16, 21, being arranged ...
6. ... by two elongated ends, upper 27 and lower 28
7. ... rolling bodies 16 and 19, ...
8. ... electrical structures 104A and 104B
9. ... electrical structures 104A-C are used
10. ... into four variable volume cavities 10a, 10b, 10c and 10d.

In this set, the first entry is the most common and least complex. A noun referring to the component name is followed by an embolden tag that encapsulates the component number as it appears in the drawings. This noun-tag relationship is predominant to a degree that a fair extraction of information should be possible if only this relationship was used.

Multiple tags can, in some instances, refer back to the same noun. In the second entry "uncoated regions" is followed by two tags, each thus an uncoated region. The logic of this is simple enough, but it does beg the question on how pluralities are to be resolved.

It should be noted that reference to a part can extend beyond a given tag. Entries 2, 8 and 9 are good examples. In the first instance, adjacent and tags are used as a single reference. In most cases where this combined form of reference is used, reference is made to a feature or property of a part. For example, "uncoated regions" in entry 2 refers to regions on electrodes.

In entries 8 and 9, a character appended onto the closing tag provides a complete reference to a distinct part. This is expanded in 9, where a single tag is followed by "A-C".

Several conclusions can be drawn from the information -

A tag containing a part reference, or a list of such tags, will be directly preceded by a noun that strongly identifies the referenced part.

Component reference tags can take several forms. The most common of these are:

- {part reference number}
- {part reference number}{sub-part or part instance reference}
- {part reference number}{sub-part or part instance reference}

3.3.1.2 NOUN-TAG COUPLINGS

In this section the noun tag coupling is further explored by expanding the set.

Consider the following patent excerpts:

1. <p id="p-0037" num="0036">The case 34 is formed to be approximately cuboid...
2. The rechargeable battery 100 according to...
3. ...has an electrolyte injection opening 27 for...

As in the previous section, the first instance simply demonstrates the most common relationship between a tag and preceding noun.

The second and third excerpts follow the same basic noun-tag structure, but with some subtle differences. In the second case the noun “battery” is preceded by an adjective “rechargeable”. In the third case “electrolyte injection opening” are all classifiable as nouns. In Chapter 3 it was established that components should have names and a set of additional attributes. It is therefore appropriate to define a convention to stipulate when a descriptive word will form part of the name and when it will be added as an attribute.

Given a phrase extract from a sentence where all words are either a form of adjective or noun and the phrase is followed by a tag indicative of a part name, it would apply that

$$Name = \begin{cases} \{W_{n-k}, \dots, W_n\} & [W_{n-i}]_{POS} = NNx \; \forall \{i \in N \mid 0 \leq i \leq k\} \\ \{Wn\}, & [W_{n-i}]_{POS} \neq NNx \; \forall \{i \in N \mid i > 0\} \end{cases}$$

$$Properties = \begin{cases} \emptyset, & [W_{n-i}]_{POS} = NNx \; \forall \{i \in N \mid 0 \leq i \leq k\} \\ \{W_{n-k}, \dots, W_{n-1}\}, & [W_{n-i}]_{POS} = JJx \; \forall \{i \in N \mid 1 \leq i \leq k\} \end{cases}$$

for a given phrase in the form $[W_{n-k}] \dots [W_{n-1}][W_n] < t >$, where NN is a type of noun, JJ a type of adjective and W_x a word in a phrase.

3.3.2 THE EFFECT OF EMBODIMENTS

A prominent feature in many patents is the inclusion of multiple embodiments, or ways the technological artefact can be constructed. This can have a major impact on any automated attempt to measure the complexity of the artefact in question. Patent US8925122 will be used to demonstrate how the inclusion of multiple embodiments affects the mention and placement of parts.

Two of the patent figures are included below in Figure 17. Table 11 provides a list of names given to the parts in the patent figures. Two quandaries are immediately apparent. Firstly, let it be assumed that some metrics make use of the number of parts in a patent to gauge complexity. Consider a patent where multiple embodiments are included, which parts should then be fed as an input to such a metric? Several possibilities present themselves. All could be added, only the first could be added, or a combination of shared and unique parts per embodiment can be considered. To provide a meaningful answer something else must first be considered - which of these embodiments is more complex and how is it measured.

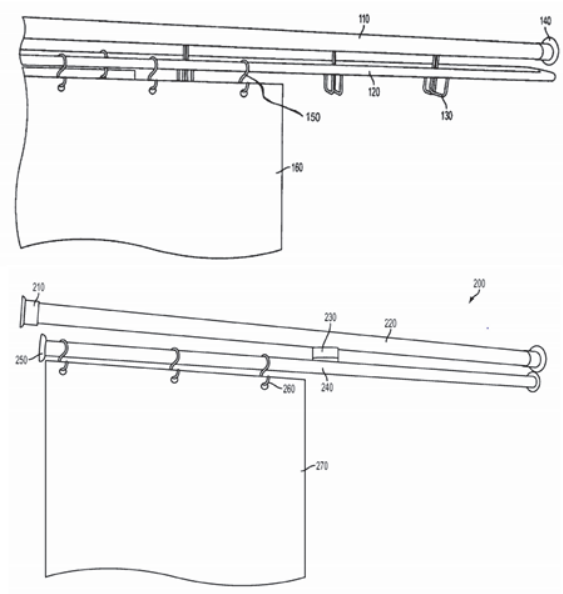


Figure 17: Two embodiments from patent US8925122 [78]

First Embodiment		Second Embodiment	
Part no.	**Part Name**	**Part no.**	**Part Name**
100	shower curtain rod	200	shower curtain rod
110	horizontal bar	220	horizontal rod
120	"loop-shaped" bar	240	lower curtain-hanging rod
130	"J" hooks	230	pivot mechanism
140	mounting cradles	210	mounting cradles
150	"S" hooks	260	"S" hooks
160	shower curtain	270	shower curtain
		250	Stoppers

Table 11: Part names and numbers for two embodiments of patent US8925122

3.3.3 THE COMPLEXITY OF SHOWER RODS

In the previous section two embodiments of a shower rod were provided. Before it can be considered how a jet engine is more complex than a teacup, it would be helpful to first analyse a simpler case, such as shower rods.

In the first embodiment, the shower curtain can be drawn all the way around the loop shaped bar. In the second, the shower curtain is attached to a lower rod that is pivoted on a second pole. This would allow the curtain to be swung outward at an angle from the anchoring rod.

A few observations need to be made before any analysis is done on the complexity of the shower curtains. Firstly, these are closely related artefacts, and so it would be expected that any measurement of complexity would not differ vastly between the embodiments. Secondly, it should be noted that no subjective metrics of complexity can be included in this analysis. To the engineer the first design of Sputnik might seem trivial while at the same time it may seem incomprehensible to a layperson.

A simple metric of complexity would be to simply count the number of parts present in an invention. Even this simple approach runs into trouble quite quickly – should multiple instances of the same part be considered? This would once more be contingent on the definition of complexity. If the *types* of parts are considered, the second embodiment is more complex as it has one additional type of part. If every part *instance* is counted, the first embodiment is more complex as it has multiple instances of "J" hooks. Considering that almost all definitions of complexity consider the interaction of all the different parts in a system, it would seem logical to include multiple instances of a part if one is to measure *structural* complexity. If the complexity of design or manufacture is under scrutiny, it may be less important – to produce one or many of the same artefact does not require more knowledge intrinsic to that part.

To account for these difficulties, multiple approaches for extracting parts are considered. The first is a brute count of all part instances in a patent. The second approach would focus on finding unique parts in a patent. These proposed metrics for complexity would be highly compatible with a process that uses patent data as a feedstock.

4 METHODOLOGY

4.1 INTRODUCTION

This section addresses the steps and methods required to satisfactorily answer the research questions identified earlier. Figure 18 provides an overview of the process.

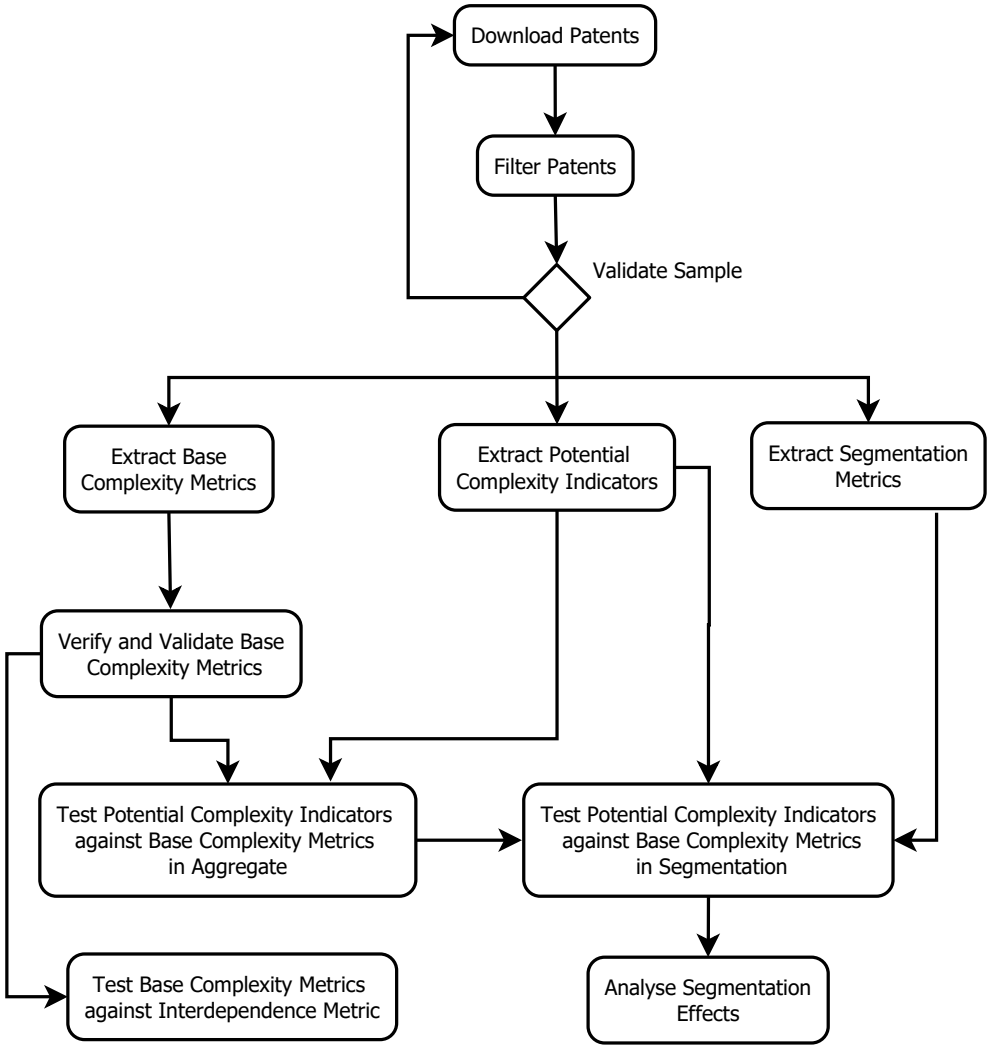


Figure 18: Methodology Diagram

The first part of this methodology is to identify and collect a set of patents. This sample set is then assessed for suitability. Several controls are put into place to ensure that the sample size and nature is fitting.

A set of base complexity metrics are then extracted from the sample. These aim to capture the complexity inherent in patented inventions. A set of tests are conducted on these to assess if they do indeed measure complexity and are accurately extracted.

The base complexity measures are used to assess how other data fields in the patent document relate to the complexity of the invention. These potential complexity indicators are plotted against the base metrics to identify potential correlations. Pearson coefficients are used to assess if a linear relationship exists between the indicator and the base complexity measure.

All comparisons of metrics are done both in aggregate and segmented form. Some patent data fields can be used to segment the sample. This adds additional perspective to the results.

All of the above-named comparisons aim to address the first research question, namely -*What is the relationship between patent metadata and invention complexity in patented inventions?*

A second comparison is made between the complexity measures used in the Fleming and Sorenson study [11]. This is done to answer the second research question, namely - Given *a set of parameterised relationships between complexity metrics and patent data (Research question 1), would the conclusions of the Fleming and Sorenson study remain valid?*

Several assumptions are prevalent in the base complexity metrics. These include -

1. Modularity and interaction are adequate indicators of the complexity of a technological artefact.
2. The number of components named in a patent is a valid proxy for the patented invention modularity.
3. All the parts critical to an invention's functionality are included in its patent and patent diagrams.
4. Differences in writing style and reference to images do not differ to a degree that disqualifies the proposed method.

To minimise the risk of these, several controls are put into place to filter and validate the sample documents used in this study.

4.2 METRICS AND DATA FIELDS

In the literature section, the different data fields in patent documentation were explored. A distinction is made between the different types of metrics and indicators that will be used.

4.2.1 BASE COMPLEXITY METRICS

The two metrics that will be used to describe the complexity of the object in a patent are modularity and interaction. All other metrics will be measured against these. It is therefore important to establish sufficient controls in their theoretical derivation and test their practical implementation thoroughly.

4.2.1.1 PART COUNT

Part count refers to the number of constituent objects within the invention's construction. One of the legal principles of patenting is that any person with an average understanding of the field should be able to replicate the invention from the patent. It follows that all parts critical to the functionality of the patent will be mentioned, explained and referenced in the patent drawings. This implies that the data source is sufficiently complete so that this metric can be derived therefrom within an acceptable error margin.

It also follows that an increase in the number of parts within a patent will have a direct impact on the intricacies of designing, maintaining, understanding and manufacturing the patented object. If these elements increase, the knowledge requirement and production steps needed to materialise the object will increase as well. This justifies the use of this metric as a proxy for complexity as an increase in the aforementioned elements would constitute an increase in complexity, under most of the definitions provided in the literature section.

The synthesis section demonstrated that XML tags are present where numbered parts in the patent figures are referenced in the text. In most cases, the tags are preceded by the part's name.

The part names will be extracted in the following way:

1. Find the start of a relevant XML tag.
2. Iterate backwards over the preceding words.
3. Stop when a stop word, XML tag or the beginning of the sentence is reached.

The extracted terms are then normalised. This is done to ensure that parts are not double counted for patents with multiple embodiments. The normalisation process compares the extracted part names against the figure reference number and concatenates duplicates.

4.2.1.2 PART INTERACTION

Part interaction constitutes any physical or conceptual coupling between the parts in a patent. The premise that critical parts are named, referenced and explained also holds for this metric. A secondary premise is that an intricate relationship between parts will require a more detailed explanation than a simple one. If the effort of explaining the relationship can be measured, then it can act as a proxy for the interaction between parts.

Two methods are proposed to extract interaction. The first is to make use of a syntactic parser to extract the semantic relationships between words in every sentence in the patent description that contains at least two parts. From these relationships the nature of the interaction can be deduced.

A second option to measure this metric is to assume that if two parts co-occur in a sentence, it denotes some type of relationship between them. From manual inspection of patent texts, this seems to be the case.

Part interaction will be measured in two ways. Firstly, the total amount of interactions between all parts will be counted. A second measure will be taken where at least one connection is present between parts.

4.2.1.3 BASE COMPLEXITY METRIC VALIDATION

The literature has shown that complexity in the innovation process increases over time. It follows that the base complexity measures should reflect this increase.

To test if the base complexity metrics capture this increase their average values are plotted over time. A steady increase would denote that they capture aspects of complexity in inventions.

One implication of this test is that sample patents should be taken evenly over a time span of at least ten years.

4.2.1.4 BASE COMPLEXITY METRIC VERIFICATION

The accuracy of the extracted metrics needs to be tested and confirmed. To this end several patents will be manually processed to recreate the base complexity metrics. This will be compared to the automated process to ensure correct extraction.

4.2.1.5 BASE COMPLEXITY METRIC CONTROLS

Controls are put into place to ensure the integrity of the extracted base complexity metrics. Manual inspection of patent documents has shown that not all patent documents are fitting for this method.

The first instance where control is required is in relation to chemical process patents. These rarely reference parts and would typically contain descriptions of chemical formulas instead. Chemical formulae descriptions are tagged through the XML schema, but not in the same way as parts. To control for this, all patents related to chemistry are removed from the sample set. Filtering these out is a trivial task as all patents related to chemical processes contain at least one classification from the “chemistry” section under the CPC system or a <chemistry> XML tag to indicate a formula.

Inspection also revealed that a portion of patents contain figures, but reference them poorly in the description text. To control for this phenomenon all patents with part counts less than twice the amount of drawings in the patent are removed from the sample.

4.2.2 POTENTIAL COMPLEXITY INDICATORS

It has been shown in the literature that patents contain many measurable data fields beyond the direct description of the invention. These will be measured against the base complexity metrics. The goal is to assert whether they codify or relate to the measured base complexity measurements.

4.2.2.1 CLASSIFICATION CODES

The Fleming Sorenson study [11] that sparked this inquiry relies heavily on classification codes. The classifications are used as proxies for technology fields and their co-occurrence on a patent is used to calculate the level of interaction of the different technologies in the patent.

Classification codes are dissected into several metrics. The first is to simply count the unique classification codes at different levels of the classification hierarchy and to test these against the base complexity metrics. Unique classifications are counted at subclass, main group and sub group level.

Another metric that will be derived from the classification codes is interdependence, as used by Fleming and Sorenson:

$$Ease\ of\ recombination\ of\ sub-class\ i\ \equiv E_i = \frac{\sum sub-classes\ co-occuring\ with\ i}{\sum previous\ patents\ in\ sub-class\ i}$$

$$Interdependence\ of\ patent\ l\ \equiv\ K_l = \frac{\sum sub-classes\ on\ patent\ l}{\sum_{l \in i} E_i}$$

This will be measured against base complexity metrics in several ways. Firstly, the ease of recombination of for every sub-class i will be compared against the average brute and unique interactions measured in the base complexity metrics for that sub-class. The ease of recombination will be compared to brute and unique interactions normalised to the amount of parts counted.

Comparisons will also be done on a per patent basis. The interdependence metric will be measured against the interaction metrics in the base complexity metrics set, both as is and normalised.

The broad scope of comparison of classification based metrics is necessitated by uncertainty in how complicatedness is captured in the base metrics.

4.2.2.2 CITATION COUNTS

Citation counts are used throughout the literature as proxies of various elements of innovation. In this instance the supposition for its use is that more complex technologies will consist of a larger knowledge base, and therefore would need to cite a larger array of sources.

Citations are classified under two categories - citations to previous patents and citations to external sources, such as academic publications and product specifications.

Citation counts will be compared against base complexity metrics.

4.2.2.3 CLAIM COUNT

Patent claims contain concise descriptions of an invention's form and function. It can be argued that a greater number of features and functions would increase the complexity of an invention.

The claim count for each patent is tested against the base complexity metrics for correlation.

4.2.2.4 PUBLICATION LAG

Patent documents contain both the dates on which the patent was first applied for and the date on which it was finally published. The time span between these dates differs from patent to patent and can potentially a function of the difficulty to validate the invention. This speaks directly to the complexness of the patented invention.

The publication lag can also be influenced by legal aspect surrounding the patent. To test if the time lag correlates with complexity metrics the two sets of metrics will be plotted and analysed. A Pearson coefficient will be calculated to test linear correlation.

4.2.3 SEGMENTATION METRICS

It was noted in the literature that an array of factors influences the complexity of inventions and the innovation process. It is therefore prudent to disaggregate the sample set according to these variables to reduce noise in comparisons and results. These metrics are binary in nature. By filtering on these the result set can be viewed at different levels of aggregation.

4.2.3.1 SYSTEMS AND METHOD PATENTS

The literature has shown that patents can be disaggregated according to "systems" inventions and standalone units of technology. Patents with the word "system[s]" in the title are considered to be a combination of smaller inventive units. This has the potential to influence the complexness of the invention, and is therefore adopted.

The metric is, admittedly, a rough proxy. Wording and style of patent authors can differ and there is no guarantee that disaggregation according the presence of specific words will be effective. To control for this, results from sample groups should be compared. If there is a significant difference in complexity measures, it would support the assertions that "systems" patents differ from standalone patents in complexity and that it is standard practice to indicate "systems" inventions in the patent title.

The concept is extended to filter patents not only on the occurrence of “systems” in the title, but also on the occurrence of “method”. Manual inspection has shown that a sizeable portion of patents contain the word in the title.

5 DESIGN AND IMPLEMENTATION

5.1 INTRODUCTION

The design and implementation of the methodology was a highly iterative process. These are presented together for congruency and readability. Validation and verification of metrics are also presented here.

5.2 SAMPLE SELECTION

5.2.1 DATA SOURCES

5.2.1.1 SOURCE SELECTION

Two sources of patent information were considered in the literature review, namely the USPTO and the EPO. Both of these contain a dataset that is representative of the technological landscape, but the data format differs substantially.

The first design choice regarding the patent data source(s) to be used is thus:

a. Implement an architecture to source data from the USPTO.
b. Implement an architecture to source data from the EPO.
c. Implement an architecture to facilitate both EPO and USPTO interfaces.

The two patent offices are compared below in terms of benefits and drawbacks:

1. USPTO benefits
    a. The data provided by the USPTO is more complete than that provided by the EPO. In many instances the EPO provides only the bibliographical information of patents without the patent description or claims.
    b. The XML annotation of text is more in depth than that of the EPO.
2. USPTO drawbacks
    a. The data is available only in compressed XML files intended for bulk download. As such all patents published over a period of interest need to be downloaded and processed before a search process can be initiated.
3. EPO benefits
    a. The EPO provides an API that allows the user to download a subset of patents relevant to a specific query.
    b. Access to a global range of patents is provided.

4. EPO drawbacks
    a. Data sets provided by the EPO are in many cases incomplete.
    b. Data usage through the EPO API is throttled. (2 GB / week)
    c. No XML annotation for claims or descriptions.

Taking into consideration the arguments provided it follows logically to implement a patent database using USPTO patent data. There are two reasons for this. Firstly, the nature of the data, both in terms of completeness and annotation depth, is critical to the success of the proposed methodology. Secondly, it is simpler to implement. USPTO data requires a process to convert the raw XML files into a usable database structure while EPO data will require an additional interface to the EPO API while still having to process the XML- formatted data.

5.2.1.2 SAMPLE SELECTION

The USPTO publishes patents on a weekly basis. The data is available in a single zipped XML file usually containing between 1000 and 2000 patents.

Data published between January 2007 and September 2017 was downloaded. The selection included at least 3 weeks of published data from 2007 to 2010 and at least a monthly publication thereafter.

5.2.1.3 SAMPLE PRE-PROCESSING

The zipped XML archives were firstly split into smaller sections. Each patent was saved in its own XML file. A total of 185 068 patents were extracted.

Patents were then filtered according to the methodology controls. All patents covering chemical processes and patents that did not have accompanying figures were removed from the sample set. The filtering process also served to remove plant and design patents, leaving only utility patents in the set. A total of 105 465 patents remained after filtering.

5.3 IMPLEMENTATION ENVIRONMENT

5.3.1 SOFTWARE DEVELOPMENT

A C# console application was created to process the XML documents. The language was chosen partially out of preference and familiarity. However, working in the .NET framework does have some advantages. The framework provides excellent data structures when working with large sets of nuanced data. One of these is the LINQ class which allows the user to execute SQL-like queries on a collection or XML based dataset. Consider the code snippet

in Figure 19. The element variable contains a parsed paragraph from the description section of a XML patent document. Figure 20 provides a typical example. The **Where** method filters out XML tags that are not in the form .. or ... The **ForEach** method is used to iterate over the remaining set of nodes and the predicate inside the **ForEach** is a call directly to a member of the class instances being iterated over.

```
element.Descendants()
    .Where(childNode => childNode.Name != "b" && childNode.Name != "i")
    .ToList()
    .ForEach(e => e.Remove());
```

Figure 19: LINQ Code Snippet

The snippet above will filter out unwanted tags, such as <figref> from the paragraph shown below. This approach is very useful, as complicated actions can be presented very concisely in the code.

```
<p id="p-0023" num="0022">As shown in <figref idref="DRAWINGS">FIGS. 1
and 2</figref>, the filtering apparatus <b>1</b> according to the
present invention installed between the water-receiving tank <b>22</b>
and the water-pumping tank <b>24</b> of the circulating flush toilet <b>
20</b> has a casing <b>2</b>, a toothed gate <b>3</b> formed at the
upper part of one sidewall of the casing <b>2</b> close to the
water-receiving tank <b>22</b>, and a filtering unit <b>5</b> disposed
in the upper part of the casing <b>2</b> in multiple stages.</p>
```

Figure 20: Typical Patent Paragraph

A second reason for using the .NET framework is the availability of NuGet packages. These contain coding resources for a vast array of tasks. Figure 21 shows a list of packages used during implementation. The presented implementation is in compliance with all package licencing.

The IKVM package [79] contains a Java virtual machine and other tools required to enable Java and .Net interoperability. This package was needed to support the Stanford Core NLP package [80], which is implemented in Java. The Stanford Core NLP package is used for natural language processing tasks.

The MySQL.Data package [81] is an ADO.Net driver for MySQL. It was used facilitate queries to the results database.

The NUnit package [82] is a unit-testing framework that was used to verify the correctness of the implementation. StyleCop [83] contains a list of consistency rules that help ensure good quality code.

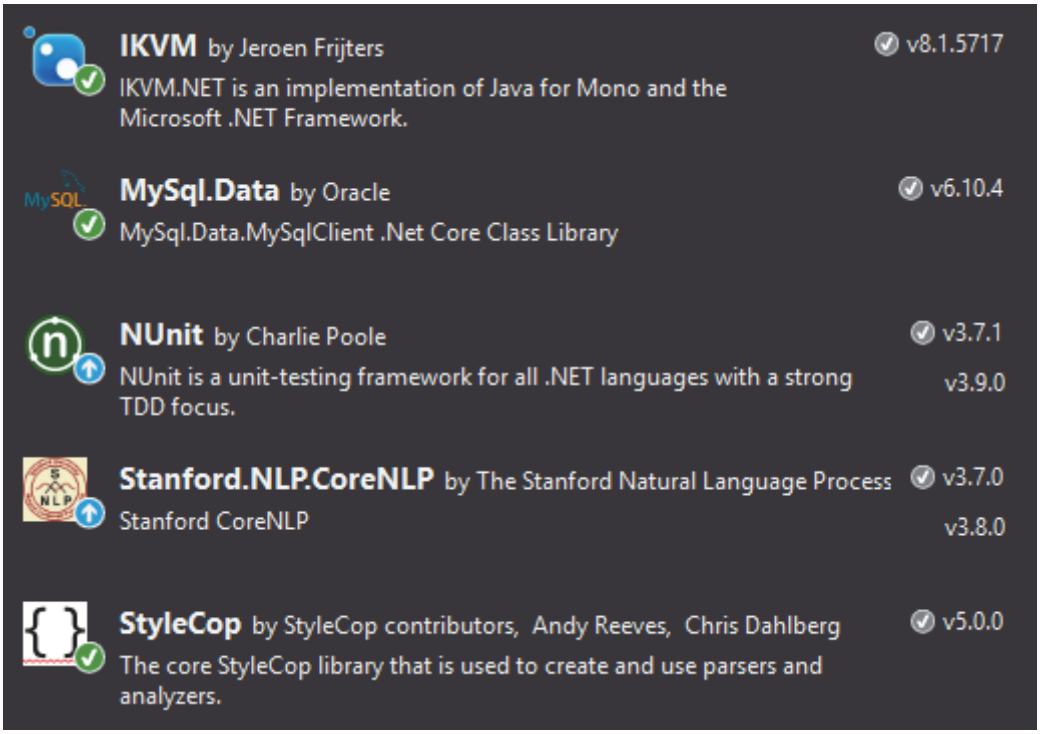


Figure 21: NuGet Packages used during Implementation

5.3.2 RESULTS AND METRIC STORAGE

A Microsoft Access database was used to store extracted metrics during the first implementation iteration. Access was chosen for the following reasons:

1. It integrates well with MS Excel, which would make the processing of results easier.
2. Data backup and data mobility is easy.

Despite its advantages Access was replaced by a MySQL database midway through implementation. The two main reasons for this were:

1. Queries between the C# implementation and MS Access were dismally slow, even after optimisation.
2. It was found that the variant of SQL used by MS access did not support all of the required functionality.

5.4 BASE COMPLEXITY METRICS

5.4.1 OVERVIEW

The methodology requires that both the number of parts and their interaction be measured from patent data. To measure the interaction between parts requires that parts should first be identified in the text before their interconnectedness can be determined.

5.4.2 TOKENISATION

A structure needs to be imposed on the XML data before any information can be extracted from the patent documentation. For this purpose, the patent document is split up into parts according to the XML hierarchy. The sections consisting of free text descriptions, such as the paragraph shown in Figure 20, are then further decomposed in tokens. Tokens represent the smallest unit of knowledge and a blueprint for the class is shown in Figure 22.

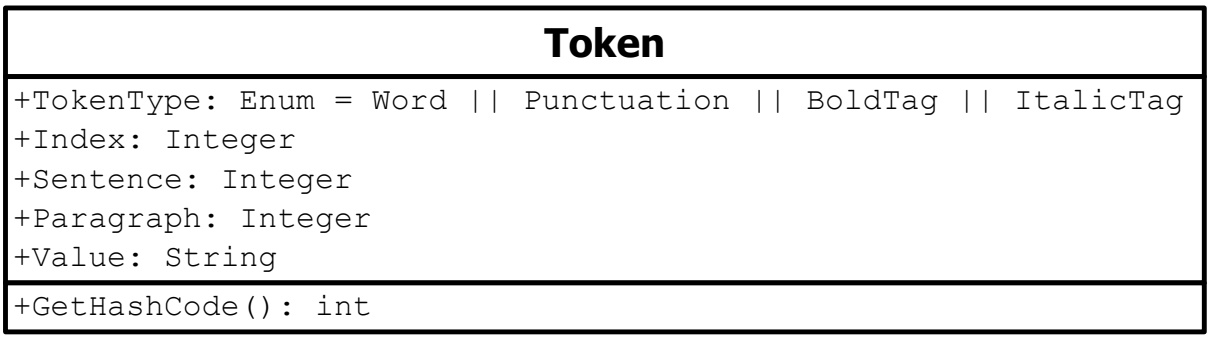


Figure 22: The Token Class

A token can represent a word, punctuation mark, bold tag or italic tag. Each token is indexed on three levels. It is stamped with the paragraph number, the number of the sentence in the paragraph and the token's position in that sentence.

Every token also has a value field. In the case of words and punctuation the field contains the word or punctuation mark. In the case of tags the field contains the number in the part number enclosed by the tag. A search of over 12 000 patents revealed that the bald and italic tags are almost exclusively used to reference part numbers from the patent drawings. Fringe cases are filtered out.

The class is immutable, in other words once an instance of the class is created its members cannot be amended from outside the class. This prevents any erroneous changes to the class members after its creation.

5.4.2.1 TOKEN EXTRACTION

A patent is first loaded as an XDocument object. The XDocument class is part of the LINQ namespace and allows for XML documents to be loaded as objects. The different levels of the XML hierarchy can then be traversed as needed.

The XDocument instance is then passed to a token parser. The parser iterates over every element of the document until it finds a paragraph of text, denoted <p>. In the LINQ architecture an element is synonymous with a tag and all the contents between the opening and closing tag.

Once a paragraph is isolated the parser sanitises the text. The main purpose of this is to remove common abbreviations. Abbreviations with full stops make sentence border detection difficult. The presence of any remaining full stops followed by capital letter or tag is considered as the break between sentences.

The parser then iterates over the sentences. Regular expressions are used to identify tags and extract their content. For tags the regular expression reads

$$< (?< Tag > [a-z]+).*? > (?< Content >.+?) <\/\1 >$$

Any text that conforms to the pattern is matched and the data therein is catalogued. The process is repeated until all free text paragraphs in the patent has been converted to tokens.

The accuracy of the process was confirmed by manually comparing patent documents to generated tokens.

5.4.3 COUNTING PARTS

5.4.3.1 DATA STRUCTURE

One of the base complexity metrics is the amount of parts present in a patent. A data structure was put into place to represent all the parts in a patent. The outline of the Part class structure is shown in Figure 23.

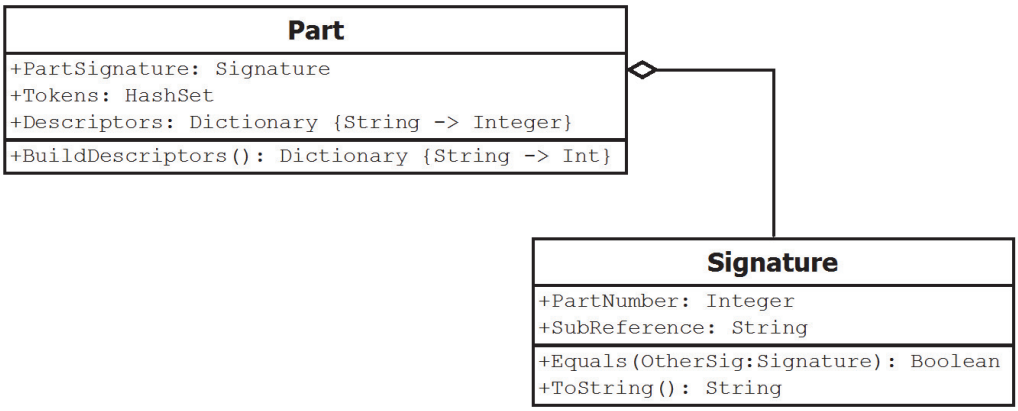


Figure 23: Part and Signature Classes

Every part is provided with a unique signature. Tags are generally used in one of two ways. In most instances a bold tag encapsulates the part reference number on the drawings. In a smaller amount of cases it is followed by an italic tag with a sub reference. For example, 321a is meant to be displayed as **321***a*. The signature class uses the tag contents as a unique identifier.

Multiple instances of the same reference are usually found in the text. All the tokens that preceded the reference tag and are found to contain words that describe the part are added to the part class instance. A HashSet was chosen as a collection container for performance reasons. The process of identifying tokens that describe a specific part will be presented in the following sections.

The Part class also contains a Descriptors member. This member maps the lemmatised token values of tokens that describe the part to their frequency of occurrence in the patent.

5.4.3.2 TOKEN COLLECTION

Up to this point, the patent has been stripped down to tokens and tags were converted to part signatures. A brute count could be done by simply looking at the amount of unique part signatures in the document. The problem with this approach is that it does no control for duplicate part entries. Many patents describe multiple embodiments of the patented invention. In many of these descriptions different tag references are used to describe the same part. Textual descriptions are therefore the only identifiers of the uniqueness of a part.

To overcome the problem of embodiments, the tokens that describe each part needs to be extracted and allocated to that part. For example, consider the extract from a patent paragraph:

```
Preferably, the helmet mount connector <b>160</b> is configured to connect to
```

The words “helmet”, “mount” and “connector” are descriptive of part number 160. These are extracted by iterating backwards from the token containing the tag and measuring a set of criteria against the value of each token. If all the criteria are matched for a token, it gets added to the part description and the iterator moves on to the preceding token. The criterion includes:

1. The token value is not a stopword. Stopwords are a list of commonly occurring words that are highly unlikely to form part of a part description. The stopword list contains all prepositions, conjunctions pronouns and articles.
2. The token value should have at least one noun or adjective lexical classification. A WordNet interface [76] is used to assess all the possible lexical classifications for each word.
3. The iterator must not go past the first word in a sentence.

In the example above the words “helmet”, “mount” and “connector” all have at least on noun interpretation and are added to the list of tokens that describe the part. The iterator stops when it reaches the word “the”. It is a definite article and included in the stopword list.

5.4.3.3 NORMALISATION

At this stage in the process, the part signature has been defined and it is linked to a list of tokens that describe it. The next step is to normalise the count by removing double counted parts from different embodiments.

Firstly, each token value describing the part is lemmatised and its frequency of occurrence is counted. Lemmatisation is also done from the WordNet interface. The lemmatisation is contextualised according to lexical classification of the token value. For example, if the word “mounting” is lemmatised under a noun context it would be returned as is. If it is done in verb context the suffix will be removed and only “mount” would be returned.

The similarity of parts is then estimated. The different descriptors are proportionally weighted according to their frequency of occurrence, or calculated according to:

$$Inner\ Weight = IW_i = \frac{Descriptor\ i\ count}{\sum Descriptor\ Count}$$

The similarity between two parts are then estimated according to:

$$Similarity(P_1, P_2) = \frac{IW_{P_1 \cap P_2}}{2}$$

The results of applying this process for patent US8925122 is shown in Figure 24. The top right half of the matrix represents the similarity of parts. The bottom left counts the co-occurrence of parts mentioned within the same sentence. The patent contained three different embodiments for the invention. These are indicated by the purple rectangles.

	100	110	120	130	140	150	160	200	210	220	230	240	250	260	270	300	310	320	330	340	350	360	360a	360b	380
100	0	0%	0%	0%	0%	0%	76%	100%	0%	33%	0%	57%	0%	27%	71%	100%	0%	96%	62%	83%	0%	39%	0%	0%	0%
110	2	0	26%	0%	0%	0%	0%	0%	0%	67%	0%	18%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
120	1	3	0	0%	0%	0%	0%	0%	0%	0%	0%	45%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
130	1	2	2	0	0%	50%	0%	0%	0%	0%	0%	0%	42%	55%	0%	0%	0%	0%	46%	0%	0%	0%	0%	0%	0%
140	1	1	0	0	0	0%	0%	0%	100%	0%	0%	0%	0%	0%	28%	0%	42%	0%	0%	0%	0%	68%	0%	45%	0%
150	0	0	5	0	0	0	0%	0%	0%	0%	0%	0%	42%	90%	0%	0%	0%	0%	46%	0%	0%	0%	0%	0%	0%
160	0	0	3	0	0	2	0	73%	0%	0%	0%	36%	0%	35%	97%	76%	0%	68%	71%	93%	0%	58%	0%	0%	0%
200	0	0	0	0	0	0	0	0	0%	35%	0%	59%	0%	25%	68%	100%	0%	96%	59%	81%	0%	37%	0%	0%	0%
210	0	0	0	0	0	0	0	1	0	0%	0%	0%	0%	0%	28%	0%	42%	0%	0%	0%	0%	68%	0%	45%	0%
220	0	0	0	0	0	0	0	3	1	0	0%	29%	0%	0%	0%	33%	0%	36%	0%	0%	0%	0%	0%	0%	0%
230	0	0	0	0	0	0	0	1	0	2	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
240	0	0	0	0	0	0	0	2	0	3	4	0	0%	21%	33%	57%	0%	57%	25%	36%	0%	14%	0%	0%	0%
250	0	0	0	0	0	0	0	0	0	0	0	2	0	47%	0%	0%	0%	0%	38%	0%	83%	0%	0%	0%	0%
260	0	0	0	0	0	0	0	0	0	0	0	3	2	0	32%	27%	0%	23%	76%	35%	0%	13%	0%	0%	0%
270	0	0	0	0	0	0	0	0	0	0	1	5	2	3	0	71%	0%	63%	66%	88%	0%	59%	0%	0%	0%
300	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%	96%	62%	83%	0%	39%	0%	0%	0%
310	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%	0%	0%	0%	85%	0%	37%	0%
320	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	54%	76%	0%	32%	0%	0%	0%
330	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	79%	0%	35%	0%	0%	0%
340	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	0	0%	56%	0%	0%	0%
350	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0%	0%	0%	0%
360	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	1	2	1	0	0%	58%	0%
360a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0%	0%
360b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	4	2	0	0%
380	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0

Figure 24: Part Similarity and Count Matrix

An example calculation of the similarity of two parts is provided for clarity on the process. Table 12 shows the text that preceded parts tags 160 and 200 in the patent text. Part 160 was referenced 12 times throughout the patent. Part 200 was referenced 5 times. The text is broken up into descriptors, as shown in Table 13. Every descriptor is assigned a proportional inner weighting. These are summed and divided by two for all descriptors that are found on both parts.

Part No	Text	Part No	Text
160	shower curtain	200	shower curtain rod
160	liner	200	shower curtain rod
160	curtains	200	shower curtain rod
160	shower curtain liner	200	shower curtain rod
160	shower curtain	200	rod
160	shower curtain		
160	shower curtain		
160	liner		
160	shower curtain		
160	curtain		
160	shower curtain		
160	curtain		

Table 12: Text extracts describing parts 160 and 200

Descriptor		Descriptor Count		Inner Weight		Overlap
P160	P200	P160	P200	P160	P200	
shower	shower	7	4	0.350	0.308	0.658
curtain	curtain	10	4	0.500	0.308	0.808
liner		3		0.150		
	rod		5		0.385	
						0.73

Table 13: Example calculation of part similarity

5.4.3.4 VERIFICATION AND CALIBRATION

In the previous section it was shown that two parts, a shower curtain liner and a shower curtain rod, was calculated to be 73% similar. This brings up the question of how similar two parts should be before they are considered equal. To answer this and to verify the correctness of the implementation several patents were analysed by hand. The manual results were matched against the implementation results. There was an 8.26% difference in the amount of words identified that describe a part. This is a cumulative representation of false positives and negatives. It may also include a small amount of errors from the manual analysis. The amount of unique tag signatures counted differed by 17, out of a total of 1099 profiled parts.

The manual analysis also included a list of parts in each patent that was thought to be equivalent. A simulation was set up that concatenated parts with a similarity score higher than a cut-off threshold. The threshold was varied and the difference between the manual assessment and implemented part counts were assessed. The results are shown in Figure 25.

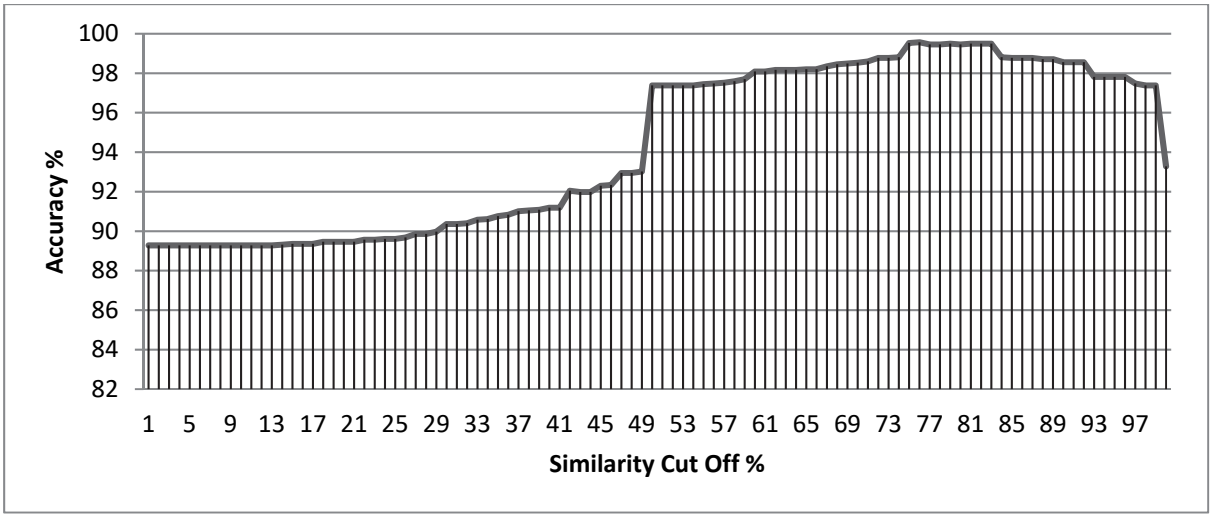


Figure 25: Calibration for part similarity

It was found that the highest level of agreement between the manual assessment and the implemented model was 99.5%. This maximum was achieved when a similarity score of 82.5% or higher was deemed to indicate that two parts were the same.

5.4.4 MEASURING INTERACTION

5.4.4.1 PARSER APPROACH

The methodology proposes two methods for extracting the interaction between parts. The first is to use a syntactic parser to extract the relationship between parts. The parser is fed the raw text from the patent and produces the most likely relationships between words.

The Stanford Core NLP parser was used to analyse a set of sample patents. Sample output from this process is shown in Figure 26. This approach was later abandoned for the following reasons:

1. The parser was not trained to handle the language use in patents. It was mostly trained on trade publications and newspaper articles. As a result, the accuracy was low.
2. The second problem was that the parser was implemented in Java and produced results at a sub-glacial pace. The parser took 30 to 40 seconds per patent. On a sample set of more than 100 000 patents this was impractical.
3. The approach is complex and convoluted, and there is a simpler method proposed in the methodology.

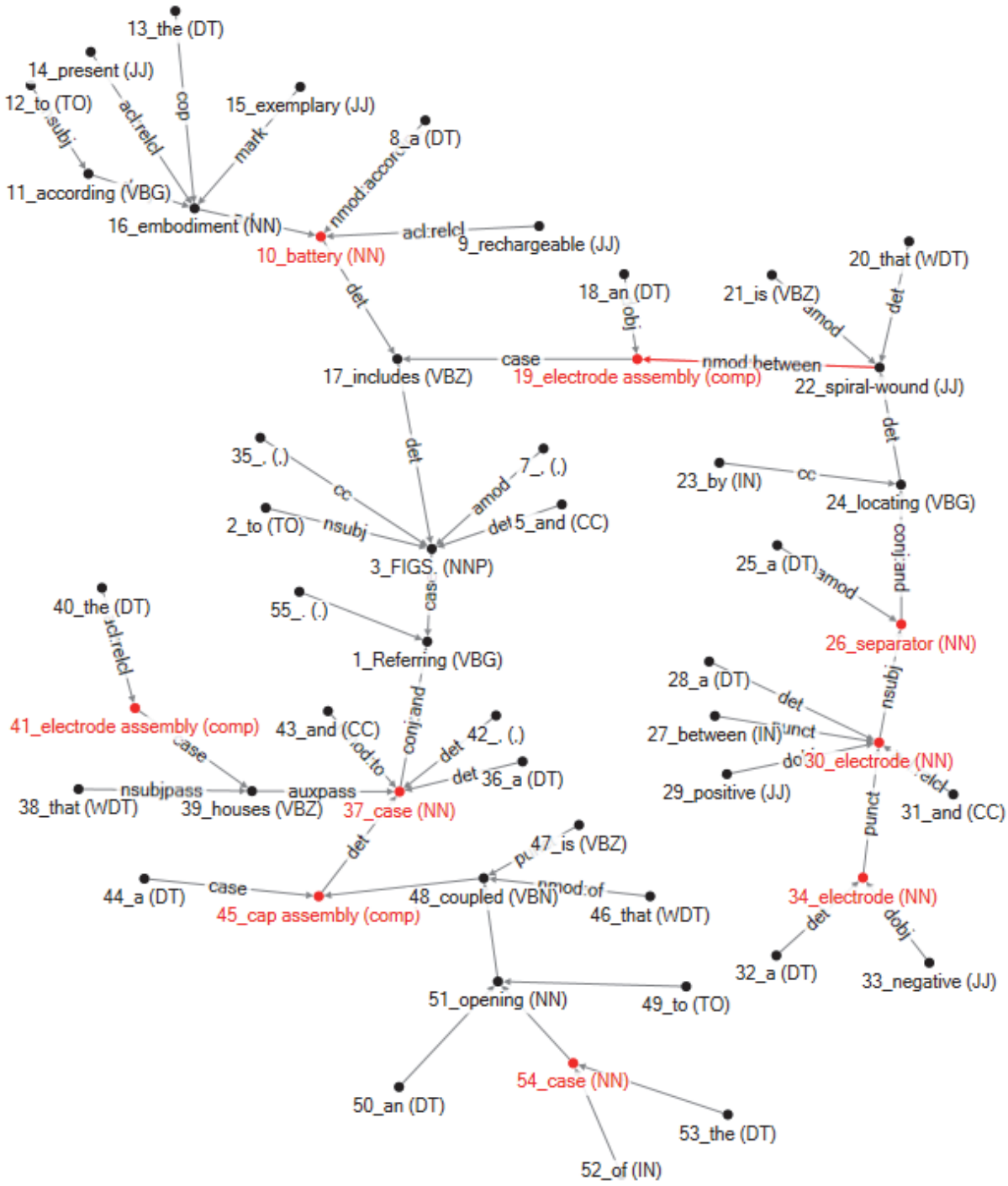


Figure 26: Parsed Sentence

5.4.4.2 PART PROXIMITY APPROACH

A second way to measure interaction is to assume that if two parts co-occur within the same sentence there is some connection between them. The implementation of this approach was simple as the pre-processing infrastructure was already in place from the part counting implementation.

A jagged two-dimensional array was created to store part proximity information. Indexes were created from the part signatures in the patent text. The programme then iterated over all tokens representing tags. These were grouped by paragraph and sentence index. If two were found in the same sentence, the array was incremented at that index.

The total amount of interactions is measured by summing all the entries in the array. Unique interaction is measured by counting all indexes with a non-zero value.

A unit test with dummy data was used to confirm the correctness of the implementation.

5.4.5 VALIDATION

The literature has shown that the innovation process is increasing in complexity. It therefore follows that any measure of complexity derived here should also show this gradually increasing trend, if the patent sample is of sufficient size.

The first action in the patenting process happens when the application for the patent is handed in. The time span between application and publication is not set and differs from patent to patent. Figure 27 shows the distribution of application dates per quarter for the sample used. Quarters where less than 1000 patent applications exist are not used in the validation process. This is done to reduce noise in the metric average.

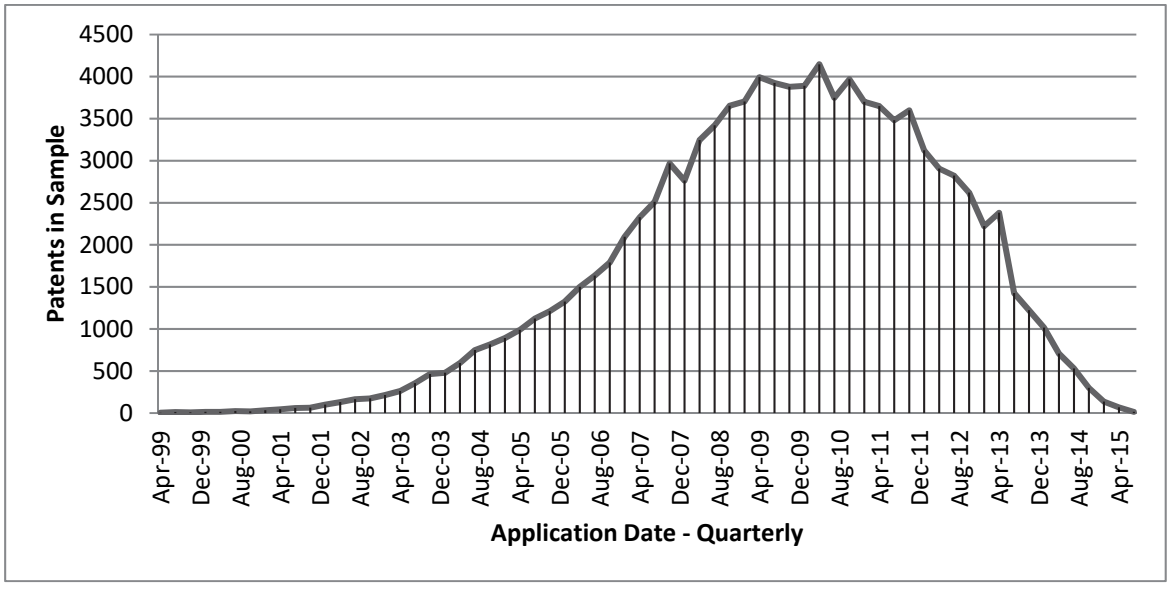


Figure 27: Patent sample application date distribution

Figure 28 shows the average part count in the sample over time. A gradual increase in part count is seen over the almost 9-year span between April 2005 and January 2014. The Pearson coefficient for this set was calculated as 0.853. This suggests a strong linear relation between average part count and patent application dates.

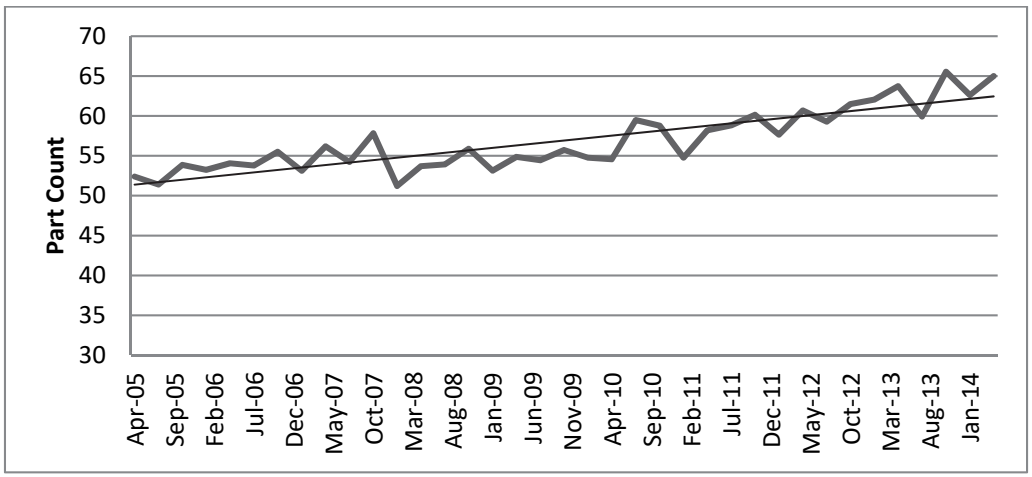


Figure 28: Average Sample Part Count per Quarter

Figure 29 shows the average part interaction (dark grey) and normalised average part interaction (light grey) as measured from the sample. Both instances show a growth in the amount of parts co-occurring, but the increase is more pronounced in the direct measurement. The Pearson coefficients for the direct and normalised interaction counts are 0.921 and 0.913, respectively.

These measurements show that, at least in aggregate, the proposed base complexity metrics are a valid indication complexity.

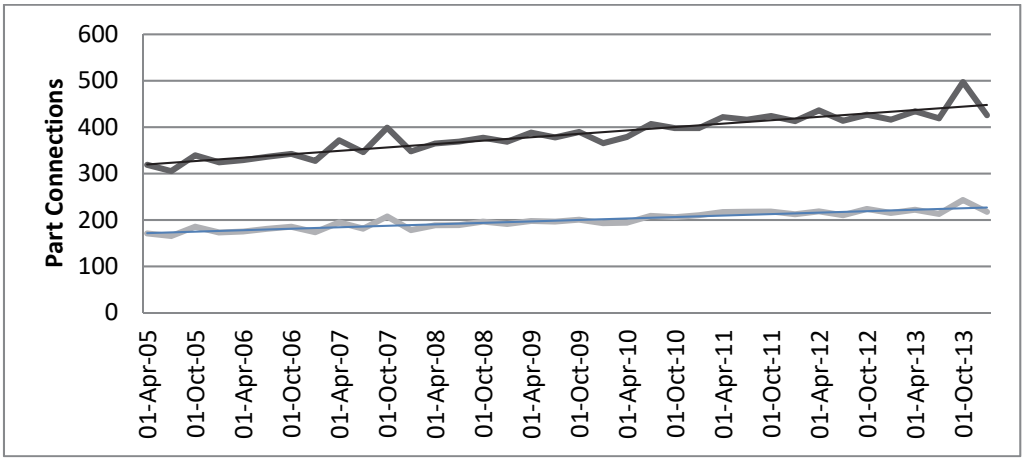


Figure 29: Average Part Interaction and Normalised Interaction per Quarter

5.5 POTENTIAL COMPLEXITY INDICATORS

The methodology requires that a set of potential complexity indicators be measured from the patent data. Most of these are explicitly stated in the patent XML structure and did not require advanced methods to extract.

5.5.1.1 CLASSIFICATION CODES

On inspection, it was confirmed that a significant proportion of patents in the sample was classified only according to the old IPC standard. Most were classified under both IPC and CPC. To overcome this shortcoming in the data the parser first searched for CPC classifications, if none were available the parser fell back to IPC classification codes.

```
<classifications-ipcr>
  <classification-ipcr>
    <ipc-version-indicator>
      <date>20060101</date>
    </ipc-version-indicator>
    <classification-level>A</classification-level>
    <section>A</section>
    <class>42</class>
    <subclass>B</subclass>
    <main-group>1</main-group>
    <subgroup>24</subgroup>
    <symbol-position>F</symbol-position>
    <classification-value>I</classification-value>
    <action-date>
      <date>20070102</date>
    </action-date>
    <generating-office>
      <country>US</country>
    </generating-office>
    <classification-status>B</classification-status>
    <classification-data-source>H</classification-data-source>
  </classification-ipcr>
```

```
<classification-cpc>
  <cpc-version-indicator>
    <date>20130101</date>
  </cpc-version-indicator>
  <section>B</section>
  <class>64</class>
  <subclass>D</subclass>
  <main-group>10</main-group>
  <subgroup>00</subgroup>
  <symbol-position>F</symbol-position>
  <classification-value>I</classification-value>
  <action-date>
    <date>20150106</date>
  </action-date>
  <generating-office>
    <country>US</country>
  </generating-office>
  <classification-status>B</classification-status>
  <classification-data-source>H</classification-data-source>
  <scheme-origination-code>C</scheme-origination-code>
</classification-cpc>
```

Figure 30: CPC and ICP Examples in USPTO XML

This approach is justified in that the CPC and IPC follow the same hierarchal structure and the classification codes are closely related. To prevent double counting in cases where CPC and IPC codes were both present on patents only CPC codes were stored.

5.5.1.2 CITATION COUNTS

Figure 31 shows the XML structure used for citations in USPTO documentation. Every citation is either marked as a patent citation or is tagged as a miscellaneous citation. The patent parser simply counted the <patcit> and <othercit> tags present in every document to measure the amount of citations.

```
<patcit num="00001">
  <document-id>
    <country>US</country>
    <doc-number>1105569</doc-number>
    <kind>A</kind>
    <name>Lacrotte</name>
    <date>19140700</date>
  </document-id>
</patcit>
<nplcit num="00020">
  <othercit>European Search Report dated Jan. 20, 2011 as
  received in European Patent Application No. GB1016384.8.
  </othercit>
</nplcit>
```

Figure 31: Patent Citation examples in USPTO XML

5.5.1.3 CLAIM COUNT

Every patent has a <number-of-claims> XML tag. This tag encloses the number of claims on all patents. The patent parser reads in the contents from this tag and parses it to an integer to get the claim count.

5.5.1.4 PUBLICATION LAG

Publication lag is the defined as the difference between the application date and the publication date of a patent. These date values are always encoded in a patent document and can be found under the tags <publication-reference> and <application-reference>. The parser reads in and stores these data fields.

6 RESULTS AND DISCUSSION

6.1 SAMPLE ANALYSIS

The patent sample consisted of 105 456 patents published between January 2007 and September 2017. Table 14 shows the sub-classes assigned most frequently.

Sub-Class	Occurrence	Sub-Class Description
A61B	17982	DIAGNOSIS; SURGERY; IDENTIFICATION
H01L	13599	SEMICONDUCTOR DEVICES; ELECTRIC SOLID STATE DEVICES
A61F	10195	FILTERS IMPLANTABLE INTO BLOOD VESSELS; PROSTHESES; DEVICES PROVIDING PATENCY TO, OR PREVENTING COLLAPSING OF, TUBULAR STRUCTURES OF THE BODY, E.G. STENTS; ORTHOPAEDIC, NURSING OR CONTRACEPTIVE DEVICES; FOMENTATION; TREATMENT OR PROTECTION OF EYES OR EARS; BANDAGES, DRESSINGS OR ABSORBENT PADS; FIRST-AID KITS
A61M	8413	DEVICES FOR INTRODUCING MEDIA INTO, OR ONTO, THE BODY DEVICES FOR TRANSDUCING BODY MEDIA OR FOR TAKING MEDIA FROM THE BODY DEVICES FOR PRODUCING OR ENDING SLEEP OR STUPOR
B41J	5078	TYPEWRITERS; SELECTIVE PRINTING MECHANISMS, {e.g. INK-JET PRINTERS, THERMAL PRINTERS}, i.e. MECHANISMS PRINTING OTHERWISE THAN FROM A FORME; CORRECTION OF TYPOGRAPHICAL ERRORS
B65D	4814	CONTAINERS FOR STORAGE OR TRANSPORT OF ARTICLES OR MATERIALS, e.g. BAGS, BARRELS, BOTTLES, BOXES, CANS, CARTONS, CRATES, DRUMS, JARS, TANKS, HOPPERS, FORWARDING CONTAINERS; ACCESSORIES, CLOSURES, OR FITTINGS THEREFOR; PACKAGING ELEMENTS; PACKAGES
B29C	4764	SHAPING OR JOINING OF PLASTICS; SHAPING OF MATERIAL IN A PLASTIC STATE, NOT OTHERWISE PROVIDED FOR; AFTER-TREATMENT OF THE SHAPED PRODUCTS
B01D	4321	SEPARATION
H01M	4111	BATTERIES, FOR THE DIRECT CONVERSION OF CHEMICAL INTO ELECTRICAL ENERGY
A63B	3898	APPARATUS FOR PHYSICAL TRAINING, GYMNASTICS, SWIMMING, CLIMBING, OR FENCING; BALL GAMES; TRAINING EQUIPMENT
B32B	3627	LAYERED PRODUCTS
G01N	3484	INVESTIGATING OR ANALYSING MATERIALS BY DETERMINING THEIR CHEMICAL OR PHYSICAL PROPERTIES
H01R	3290	LINE CONNECTORS; CURRENT COLLECTORS
B65H	3220	HANDLING THIN OR FILAMENTARY MATERIAL, e.g. SHEETS, WEBS, CABLES

Table 14: Top 14 Sub-Classes in Patent Sample

It is clear that sub-classes differ in size and utility. The most prevalent sub-class, A61B, is assigned 17 982 times to patents. This represents 7.3% of the entire sample. The 14th most prevalent class is assigned to only 1.3% of patents in the sample.

6.2 BASE COMPLEXITY METRICS

6.2.1 PART COUNT

The base complexity metrics were calculated according to the proposed methodology. Figure 32 shows the distribution of directly counted and normalised part counts in the patent sample. There is only a marginal difference between the distributions. This would suggest that the effect of multiple embodiments within the same patent is less than what was anticipated. It could also imply that parts described in separate embodiments differ to such an extent that they are measured as distinct units during normalisation.

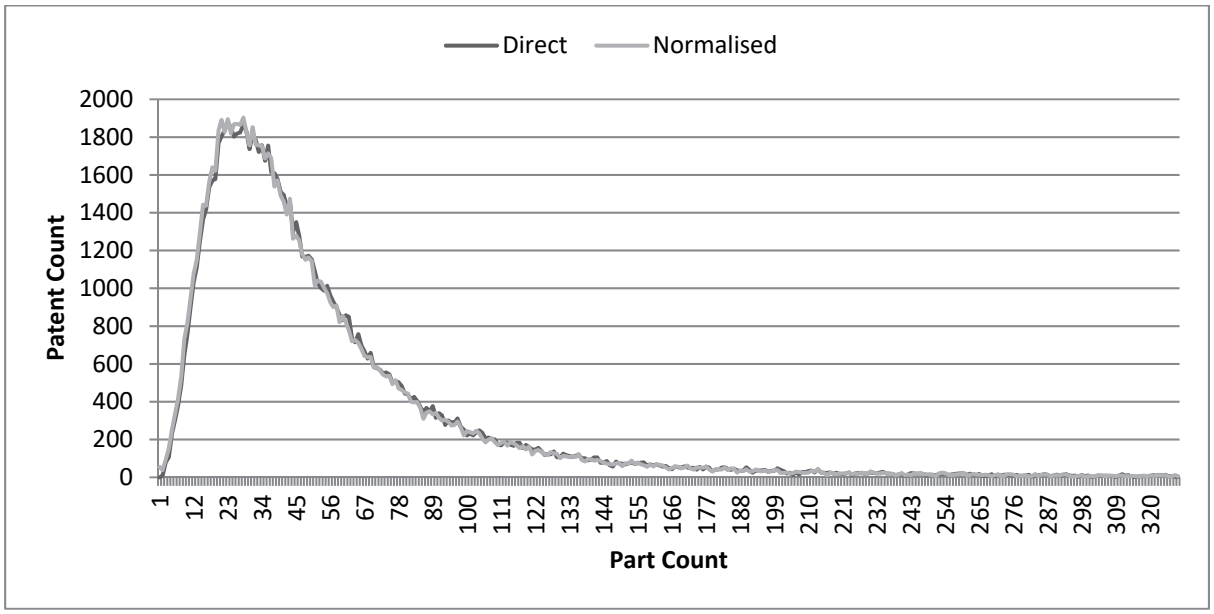


Figure 32: Part Count Distribution

The number of parts in a patent forms a Poisson-like distribution with a mean at 56.8. The normalised distribution has a slightly lower mean at 55.9. The distribution implies that most inventions can be described by between 15 and 35 function-critical units.

Figure 37 shows the distribution of directly counted and normalised interactions between patent parts. Directly measured interactions include all instances where parts were tagged within the same sentence. The normalised measure shows count interactions where parts co-occurred at least once in a sentence.

The direct and normalised interactions follow a Poisson like distribution with means at 385.6 and 200.4, respectively. The difference between the distributions gives an indication to the amount of effort required to describe a relationship between two parts.

Figure 33 shows the normalised cumulative distribution of parts as found in several sub-classes. Normalisation is done on the number of patents in every sub-class, so that the resulting distributions can be compared. The sub-classes in the figure are those most prevalent in the sample. This measurement only included patents where the subclass is indicated as the main classification. One main classification is added to each patent and describes the invention as a unit. Other classifications refer to either the invention as a whole or to one specific feature.

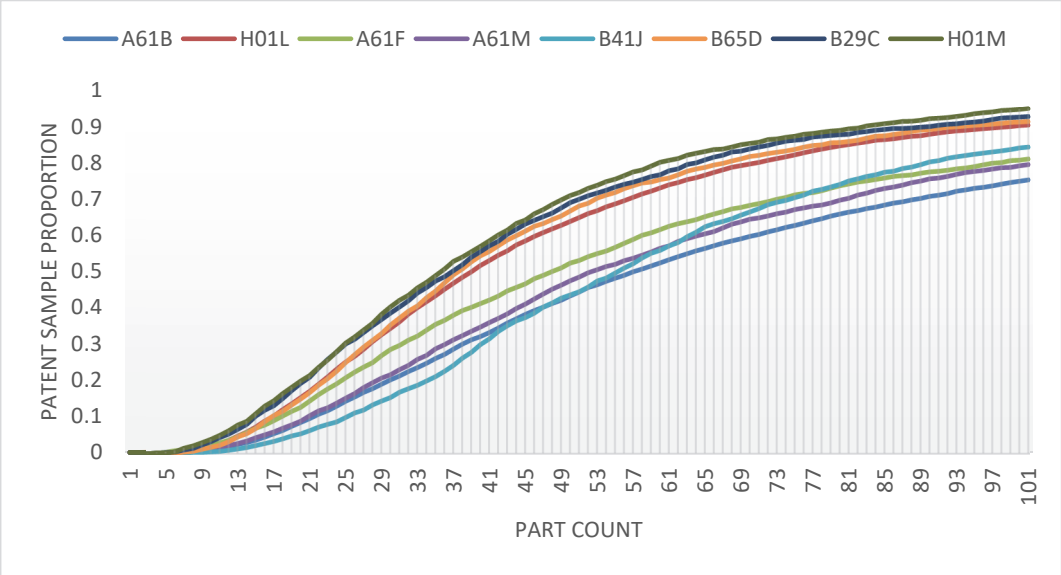


Figure 33: Normalised Distribution of Parts in Sub-Classes

It is evident that the distribution of parts in different sub-classes vary significantly. 95.0% of patents, with a main classification under sub-class H01M, have less than 100 parts. At the other end of the spectrum only 75.3% of patents classified primarily under sub-class A61B have less than 100 parts. Towards the centre of the spectrum, 90.4% patents classified under sub-class H01L has less than 100 parts.

Sub-class A61B is very broad and contains classifications for all medical scanning equipment such as MRI machines, sonar and X-ray diagnosis. It also includes surgical equipment ranging from surgical gloves to robotic devices optimised for surgical procedures.

Sub-class H01L is also broad and contains classifications for all semiconductor and solid state electronic devices. The sub-class is extended to include processes and apparatus used to produce semiconductor and solid state electronic devices.

Sub-class H01M is smaller and contains classifications for the direct conversion of chemical into electrical energy. It includes classifications for batteries, fuel cells and the manufacture thereof.

The three sub-classes present one very broad sub-class with multiple technology streams (A61B), a broad sub-class loosely based on a single technology stream (H01M) and a sub-class focused on a single technology stream (H01L). Figure 34 compares the distribution of part counts within these sub-classes in a box-and-whisker plot. The bottom whisker shows the minimum of the first quintile and the top whisker shows the maximum entry of the third quintile. The box section presents the first to third quintiles. The line in the box presents the median and the cross shows the mean.

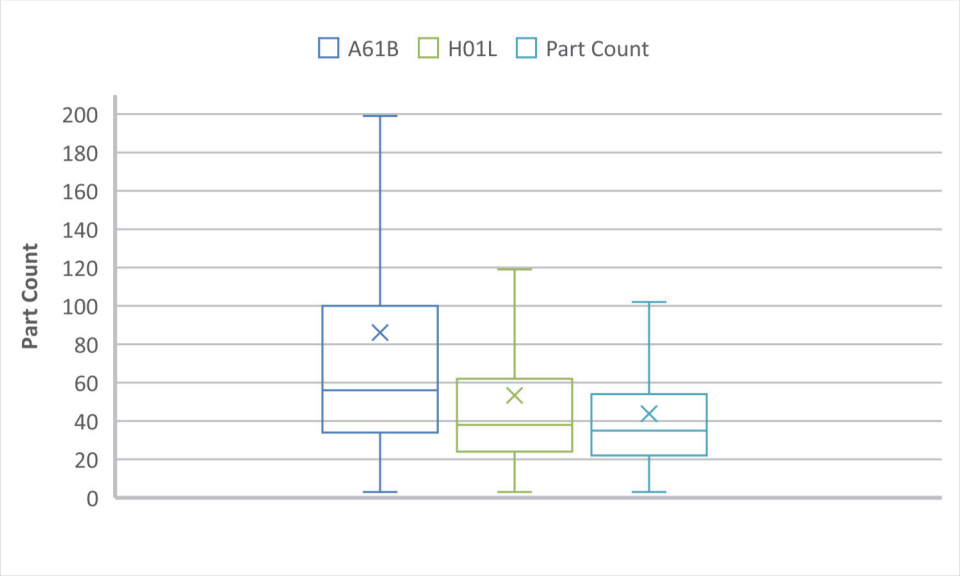


Figure 34: Box Representation of Part Count in Sub-Classes

The difference in the breadth of part distribution between sub-classes reflects the coverage of technology streams within them. Note that the first quintiles lie closer together than the third quintiles. This suggests that all of these sub-classes cover a portion of patents with low complexity. The difference in third quintiles is far more pronounced and indicates that the complexity of technology streams within a sub-class is better expressed in the upper sections of the distribution.

Figure 34 also shows that the amount of parts in a sub-class that focusses on a specific technology stream is more closely distributed around the mean than a sub-class with a broader scope.

The average number of parts in patents with a main classification under sub-classes A61B and H01M is 86.0 and 43.8 respectively. It confirms the intuitive deduction that medical scanning equipment consists of more contingent parts than batteries.

Figure 35 shows the normalised distribution of parts in sub-classes A61B and H01L along with two lower level classifications under each. Classification A61B17/70 covers prosthetic implants for stabilising the spinal cord and classification A61B18/18 covers surgical instruments that apply electromagnetic radiation. Under sub-class H01L classification H01L21/00 covers processes for the production of semiconductors and classification H01L21/4763 covers the deposition of non-insolating layers onto insolating layers within a semiconductor. The most prevalent classifications under each sub-class was chosen to bolster statistical significance.

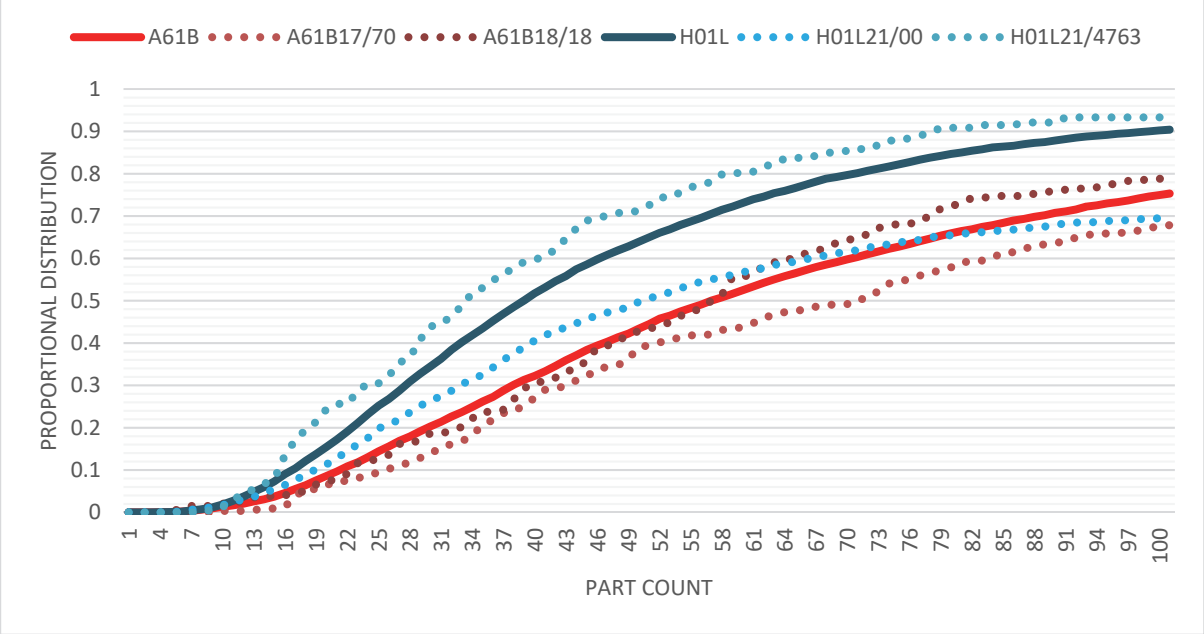


Figure 35: Normalised Proportional Distribution of Parts at Sub-Class and Sub-Group Levels

Note that the distribution of lower level classifications varies significantly from the aggregated distribution of parts in the sub-class.

Figure 36 compares the distributions of parts in patents with the words “system” and “method” in the title. The distributions of parts in patents with other title words, including “process” and

"assembly", were also measured. These did not yield a large enough result sets to provide significant comparable distributions.

The sample contained 16 507 patents with "system" in the title and 30 173 patents with "method" in the title. This presents 15.6% and 28.6% of the sample, respectively. The distribution of parts in patents with "method" in the title differs slightly from the full sample distribution. The mean amount of parts in method patents is 59.5. This is 4.2 parts per patent higher than the sample mean. The difference is more pronounced when the third quintiles are compared. For "method" patents the third quantile is 12 parts per patent greater than that of the sample.

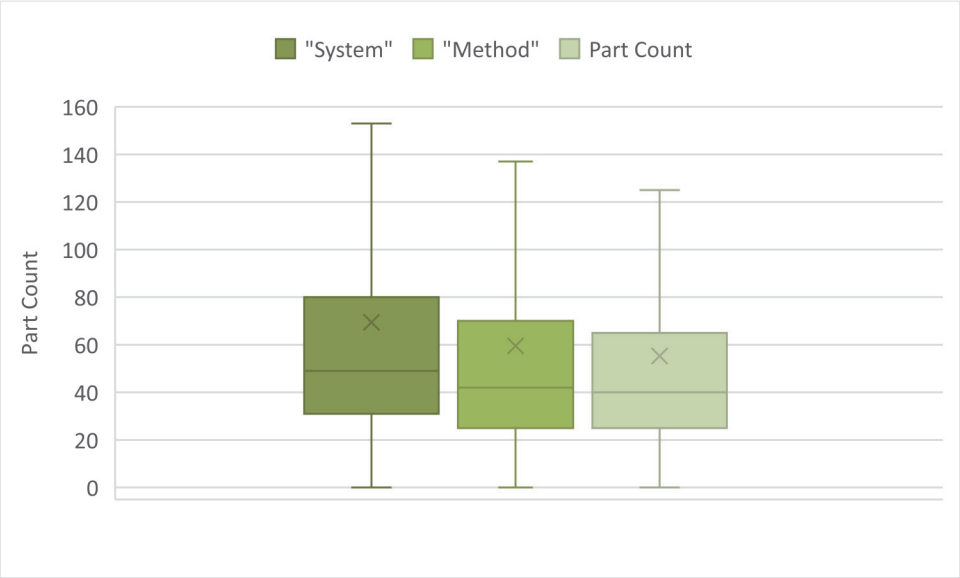


Figure 36: Distribution of "System" and "Method" Patents

The distribution of parts in patents with "system" in the title differs more distinctly from the full sample distribution. In this instance the average amount of parts per patent is 69.4. This is 14.2 parts per patent more than the full sample mean. Both the first and third quintiles are significantly larger than that of the full sample distribution.

One explanation for the difference between the sample and "systems" patent parts distributions is that "systems" patents describe the combined workings of several technological units. The other patents in the sample may only describe single technological units.

6.2.2 PART INTERACTION

Figure 37 shows the distribution of interactions between parts, as measured from the patent sample. The direct measurement is a count of every co-occurrence of two parts in a sentence. The normalised count represents a binary measure where it is counted if two parts co-occur within at least one sentence.

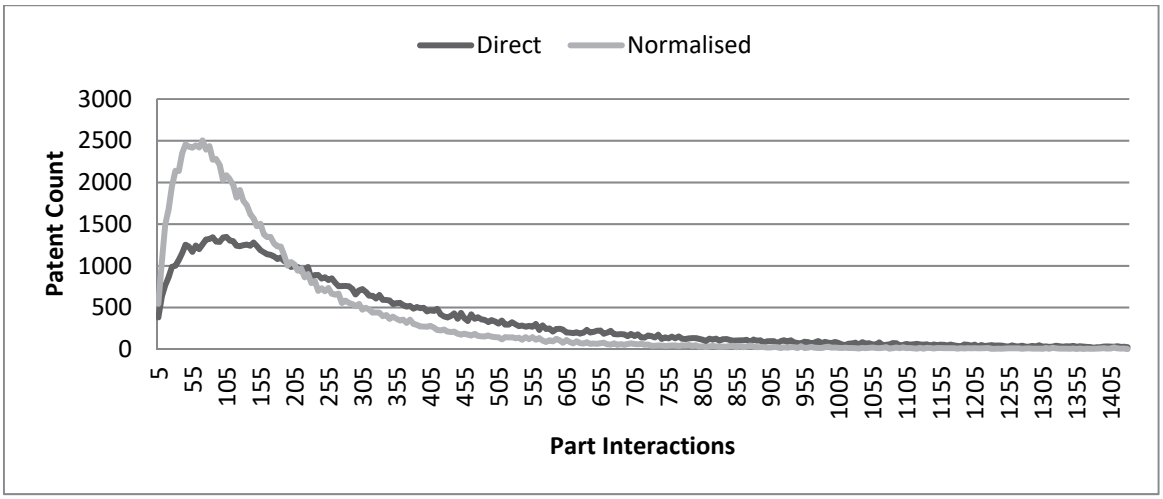


Figure 37: Part Interaction Distribution

Figure 38 shows the relationship between parts counted and the average amount of interactions in patents with a particular amount of parts. A strong linear relationship is present with Pearson coefficients of 98.8 and 99.5 for directly measured and normalised interactions, respectively.

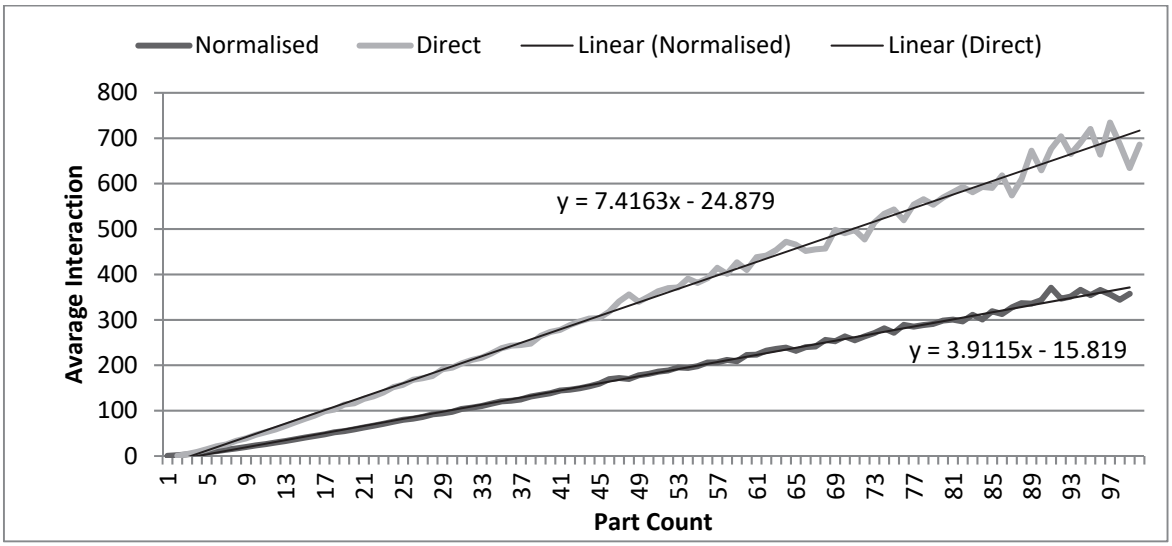


Figure 38: Average Interaction per Part Count

The relationship indicates that the amount of effort to describe the nature of an invention is directly related to the number of components. It should be noted that the metric of interaction only shows that there is some connection between two parts, not how they are connected. It therefore does not take into account the nature of the connection. For example, positioning gears in a clock is not as complex as doping a field effect transistor, but the connection count may not reflect this.

6.3 POTENTIAL COMPLEXITY INDICATORS

6.3.1 CITATION COUNT

Figure 39 shows the relationship between the amount of citations in a patent and the average amount of parts for patents with a specific amount of citations. The distribution of citations is also shown.

There is a positive correlation between the amount of parts and citations in a patent. With only one citation a patent has on average 46.1 parts. This amount remains fairly stable up to patents with 15 citations, resulting in an average amount of 47.0 parts. From this point it the average amount of parts rise more steeply. Patents with 25, 35 and 45 citations respectively have 53.8, 61.6 and 67.1 parts on average.

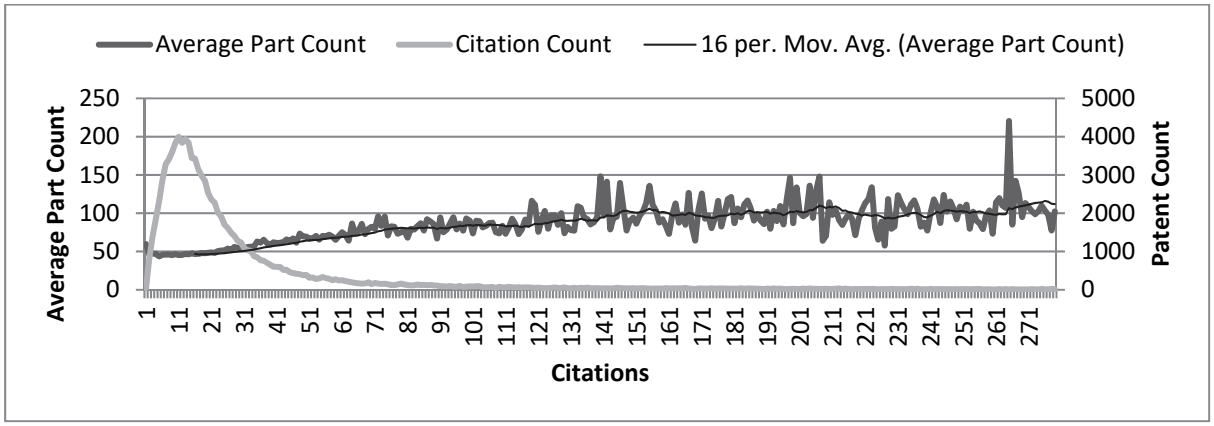


Figure 39: Relationship between Part Count and Citation Count

Measurement beyond 50 citations becomes difficult. The distribution of citations shows that very few patents have more than 50 citations. The sample loses statistical significance in the upper region as the average amount of parts is calculated on fewer patents. The resulting volatility is visible in the average part count.

Parts, as generally indicated on patents, may represent a basic object such as a pipe or button. It can also refer to a more complex artefact such as a battery pack or LED display unit. It follows that many of the more complex artefacts may be themselves patented and therefore cited.

Figure 40 shows a close up of the lower range of average of average part count per citation. The amount of citations appears to have little effect on the average amount of parts for patents with less than 15 citations. A possible explanation for this is that patents comprised of simpler parts are found in this region.

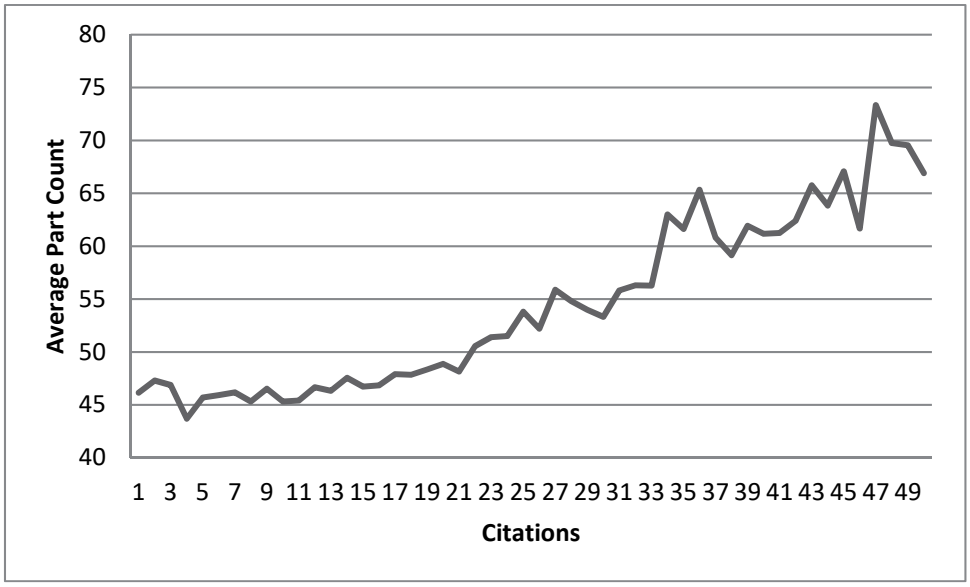


Figure 40: Average Part Count per Citation

Figure 41 shows the average interaction count per citation. As with patent part counts there is a clear positive correlation, but it is more pronounced. There is a 30% increase in average interaction between patents with 1 and 25 citations. As with part counts the statistical significance in averages degrades as citations increase.

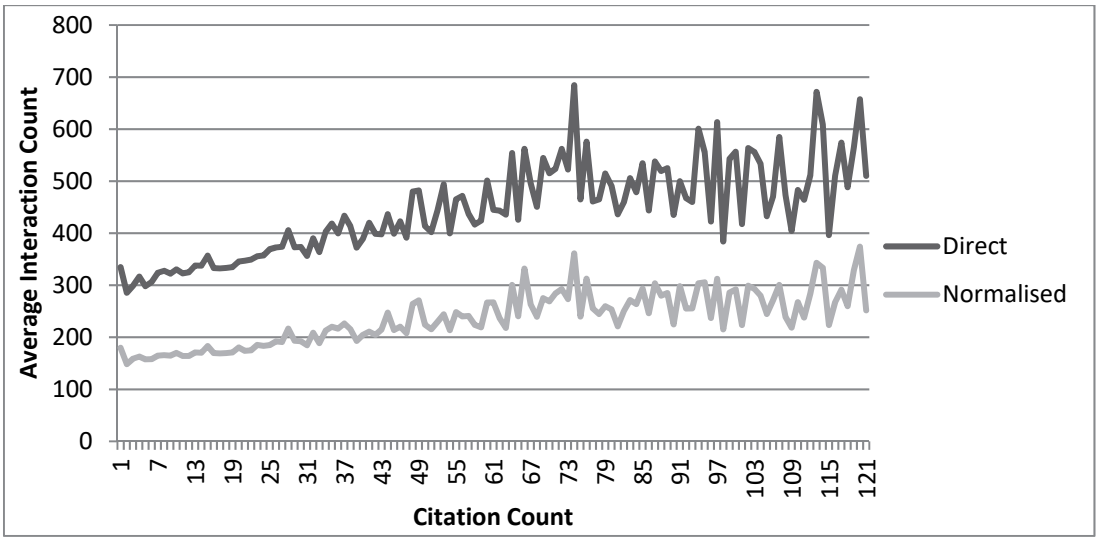


Figure 41: Average Interaction Count per Citation Count

Figure 42 shows the distribution of patent parts in patents with only patent citations and patents containing at least one other citation. Other citations generally refer to trade publications, academic publications or product manuals. It was noted that patents with other citations tend to have more parts than patents only citing other patents.

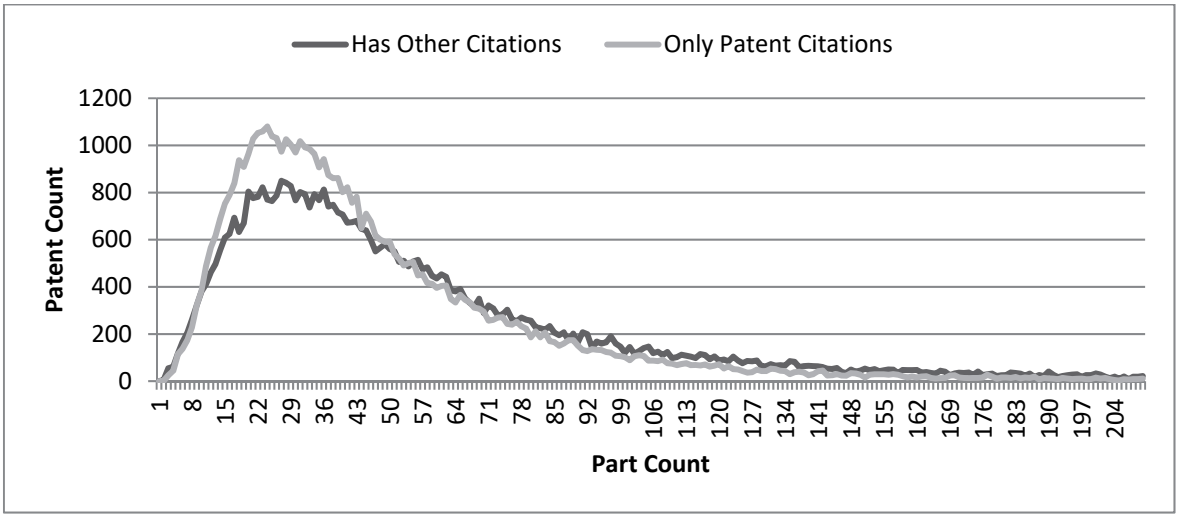


Figure 42: Patent Citation Distribution per Citation Type

The mean amount of parts for patents only citing patents is 49.93. It is 63.83 for patents with other citations. This represents a difference of about 28%. Two possible explanations are available for the difference in distributions. The first is that more complex and cutting-edge technologies need to draw and cite from the academic literature. A second possibility is that more complex patents might incorporate knowledge from multiple disciplines and therefore make it more likely to cite a broader range of publication types and knowledge sources.

Figure 43 shows the citation count vs. part count for patents in sub-class H01N. The data was filtered to sub-class level to eliminate as many unknown variables as possible. There is no distinguishable direct relationship between part count and citation count on a per patent level.

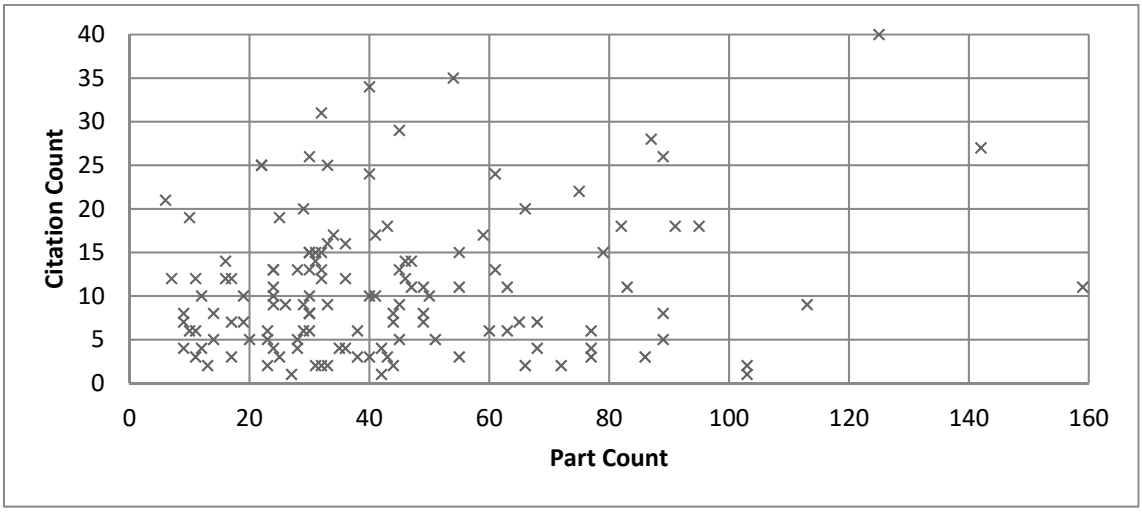


Figure 43: Citation Count vs. Part Count for Sub-Class H01L

It is therefore apparent that citation counts do reflect aspects of technological complexity in aggregate, but it is not a very meaningful indicator of complexity on a per patent basis.

6.3.2 CLAIM COUNT

Patent claims define the scope of protection of a patent in a list of succinct statements. Figure 44 shows the distribution of claims in the patent sample. It shows that almost no patents have only one claim. The portion of patents with more claims raises evenly to about 4% at 8 claims per patent. Patents with between 9 and 19 claims all contribute about 4% to the sample. The plateaued distribution ends as 10.1% of the sample has exactly 20 claims. After this point the number of claims per patent decays rapidly.

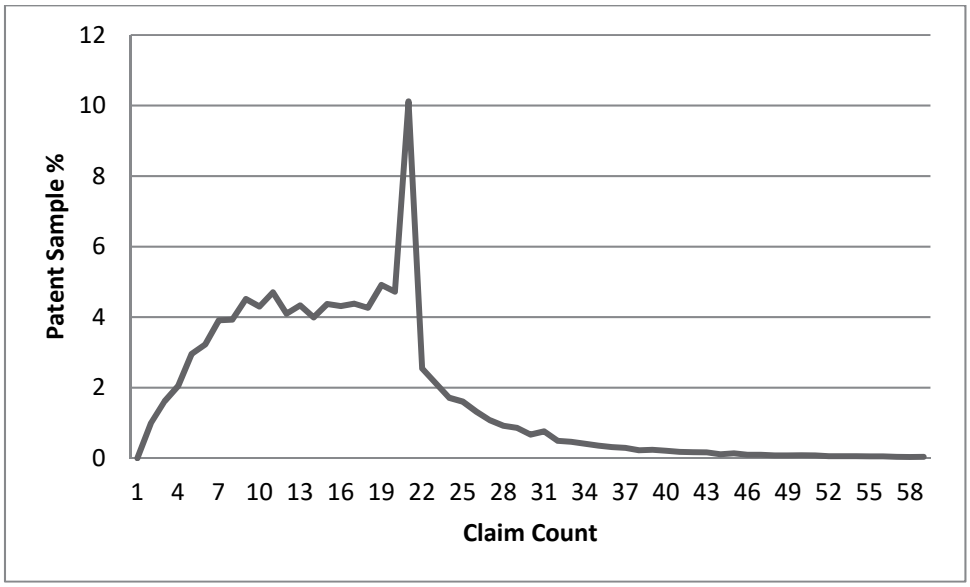


Figure 44: Distribution of Claims in Patent Sample

The anomaly in the distribution at exactly 20 claims triggered a review of the implementation. All steps were rechecked to ensure that the distribution was in fact correct. Upon deeper investigation it was found that the USPTO charges an additional fee of $80 for every claim in excess of 20 [84]. Patentees want the broadest possible application for a patent while at the same time minimising patenting cost. It would seem that for many patentees having 20 claims in a patent is the optimal method of achieving this balance.

This effect has profound implications for the usage of claim data in innovation studies. It implies that the procedures and conventions of the USPTO can cause distortions in the data. The question then becomes to what degree these distortions drown out other signals in the data.

Figure 45 shows the average amount of parts in patents grouped per claim count. Information is provided on the total sample as well as three sub-classes, namely A61B, H01M and B65D.

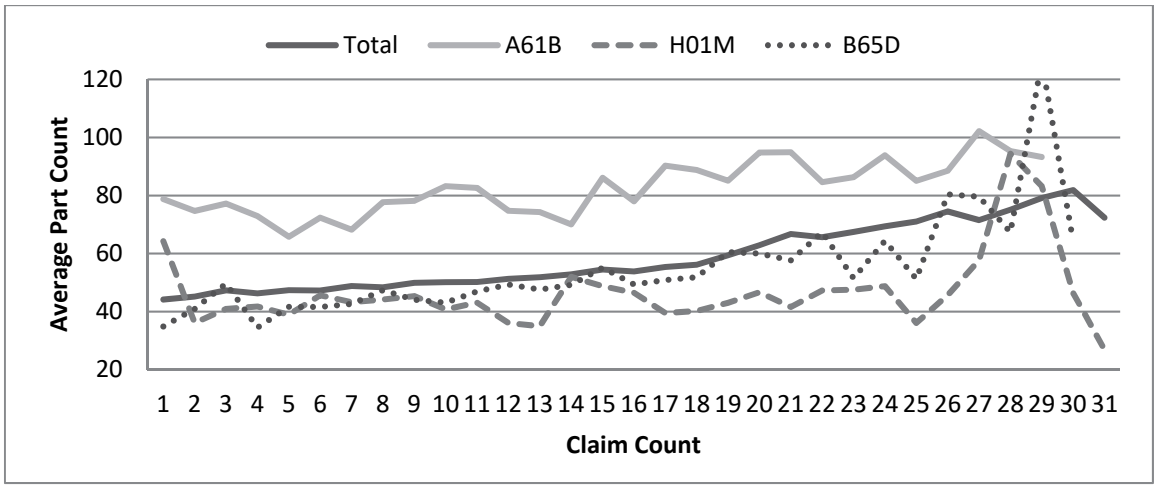


Figure 45: Average Part Count per Claim

A strong positive correlation between claim count and part count is indicated. Patents with a single claim on average has 44.2 parts. Patents with 20 claims on average has 62.9 parts and those 30 claims have an average of 81.8 parts.

Patents in sub-class A61B covers a broad range of medical technologies. This sub-class has about double the average part count per claim and also shows an increasing trend as claim count rises. Sub-class A61B covers almost 18 000 patents in the sample, that amounts to 17.05% of the total sample. Note the difference in volatility between the full sample average and that of Sub-class A61B. This shows a surprising amount of variance in part count per claim. This confirms that, as with citation counts, the number of claims on patents reflect innovation trends in aggregate, but not on individual patent level.

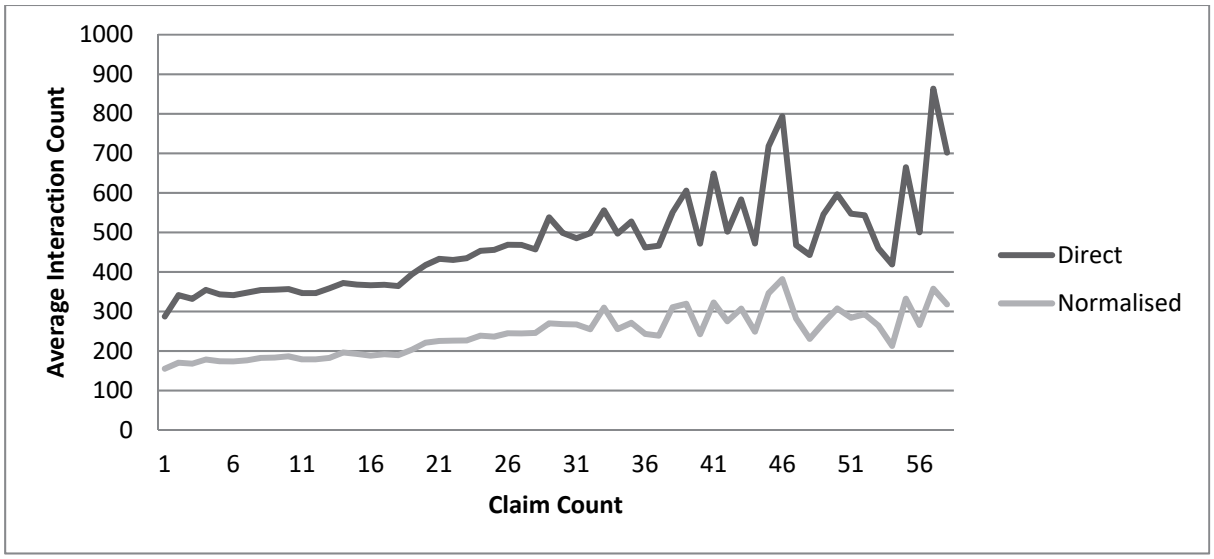


Figure 46: Average Interaction Count per Claim

Figure 46 shows the direct and normalised average interaction counts per patent claim. The positive relationship between complexity and aggregate claim count is confirmed. Note that the statistical significance of the sample degrades as the claim count increases in the region above 20 claims per patent.

6.3.3 PUBLICATION LAG

Publication lag was calculated by measuring the time that elapsed between the application and publication of a patent. Figure 47 shows the distribution of publication lag in the sample population. The same fat-tailed distribution is seen as in the case of part counts. This implies that patents are most likely to get published around three years after an application was made. In extreme cases a patentee would have to wait a decade before a patent is granted.

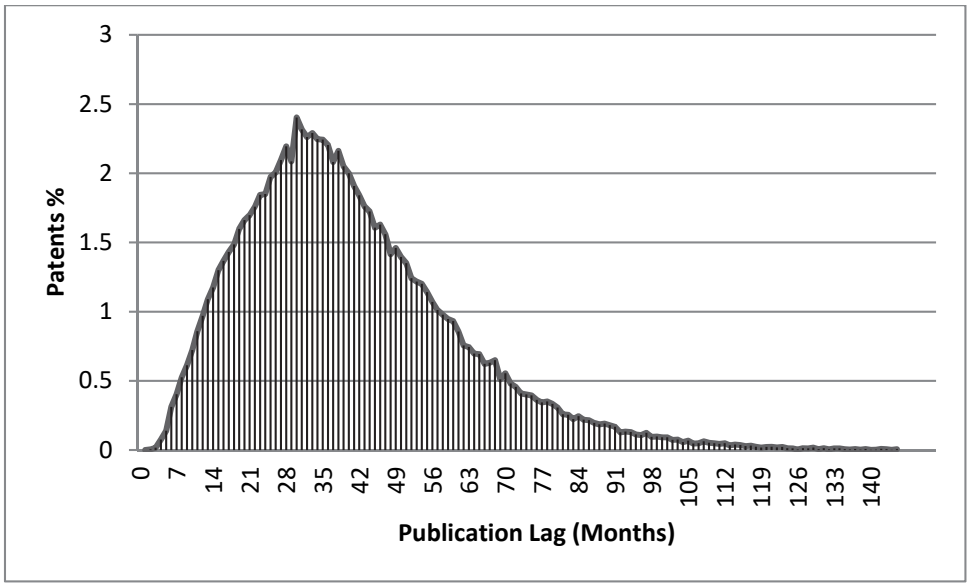


Figure 47: Publication Lag Distribution

Figure 48 shows the average interaction count against lag time for the statistically significant portion of the publication lag distribution. A negative relationship is visible between publication lag and the average amount of interaction between parts in a patent. Note that the lag time is graphed for 6 months to over 8 years. This is a long enough timespan that the aggregate increase in the complexity of technology would become manifest in the measurement. A higher lag time makes it probable that the application was done earlier than other patents in the sample. This accounts for the negative relationship. Other factors such as legal disputes and the time to assess more complex patents may also contribute to the result, but to split out these effects would require a much larger sample and complexity normalisation over application time.

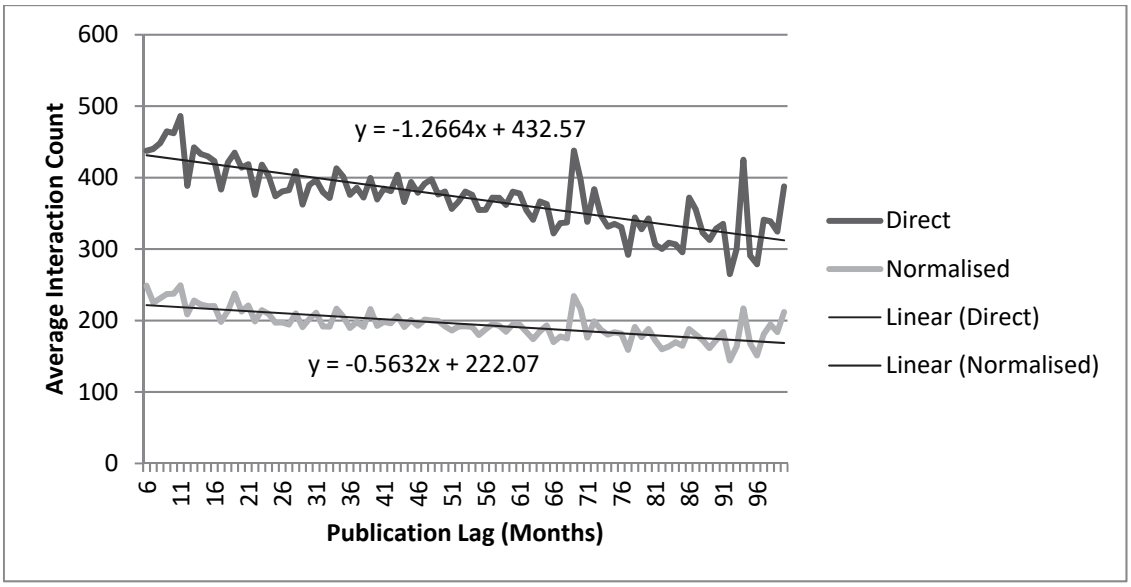


Figure 48: Average Interaction Count per Publication Lag Month

The publication lag once again validates the base complexity metrics. It shows that a measure of complexity is reflected in the amount of parts and interactions in inventions.

6.4 INTERACTION, RECOMBINATION AND INTERDEPENDENCE

In this section Fleming and Sorenson's Interdependence metric is measured against the base complexity metrics. Interdependence is calculated in two steps. The first is to determine the Ease of Recombination according to -

$$Ease\ of\ recombination\ of\ sub-class\ i\ \equiv E_i = \frac{\sum sub-classes\ co-occuring\ with\ i}{\sum previous\ patents\ in\ sub-class\ i}$$

Figure 49 shows the average interaction count per sub-class against the Ease of Recombination for every sub-class in the patent sample. Note the concentration of E_i at integer values. These are caused by sub-classes occurring only once or twice in the sample. As a result, the E_i value is normalised by dividing by 1, leaving the sum of co-occurring sub-classes on the single patent unchanged.

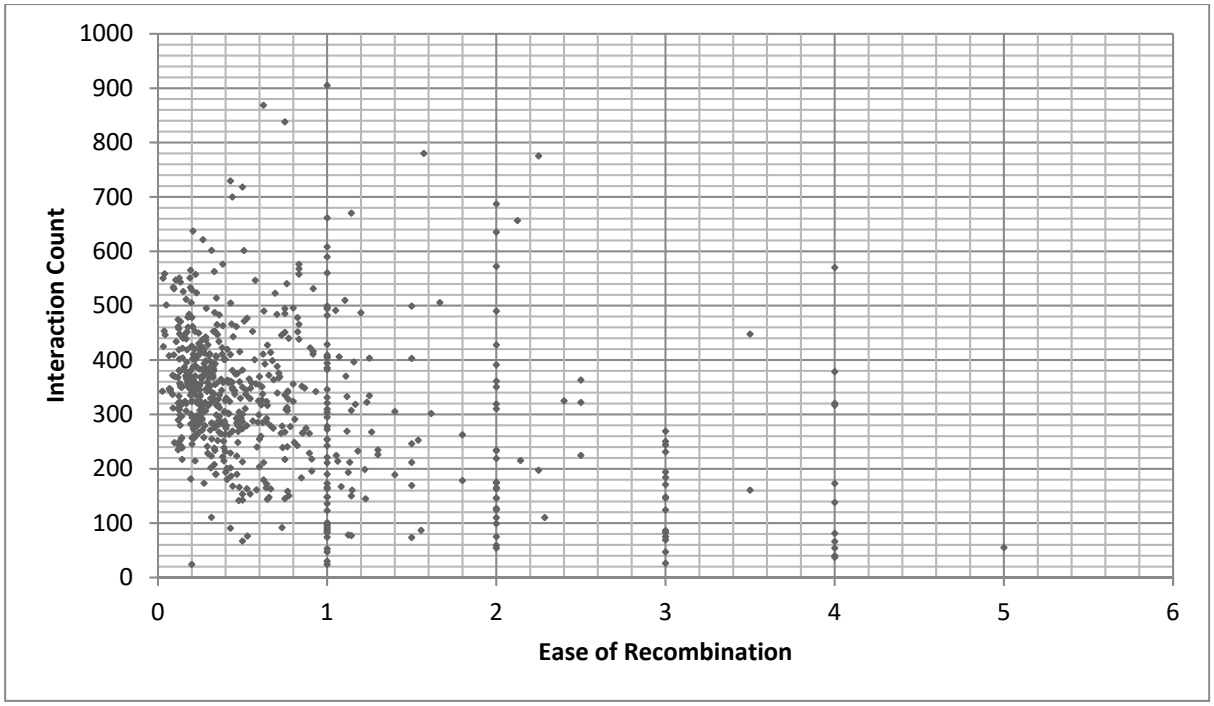


Figure 49: Average Interaction vs. Ease of Recombination

No relationship can be drawn between the interaction count and the Ease of Recombination. This leaves only the possibility that one or both measures do not capture the level of complex interactions in patents.

The second step of Fleming and Sorenson's metric is to calculate the interdependence of an invention at patent level. This is done according to -

$$Interdependence\ of\ patent\ l\ \equiv\ K_l = \frac{\sum sub - classes\ on\ patent\ l}{\sum_{l \in i} E_i}$$

Figure 50 shows the interdependence over the normalised interactions per part count. The horizontal clusters each consists of patents in a particular sub-class with the same configuration of sub-classes co-occurring on each patent.

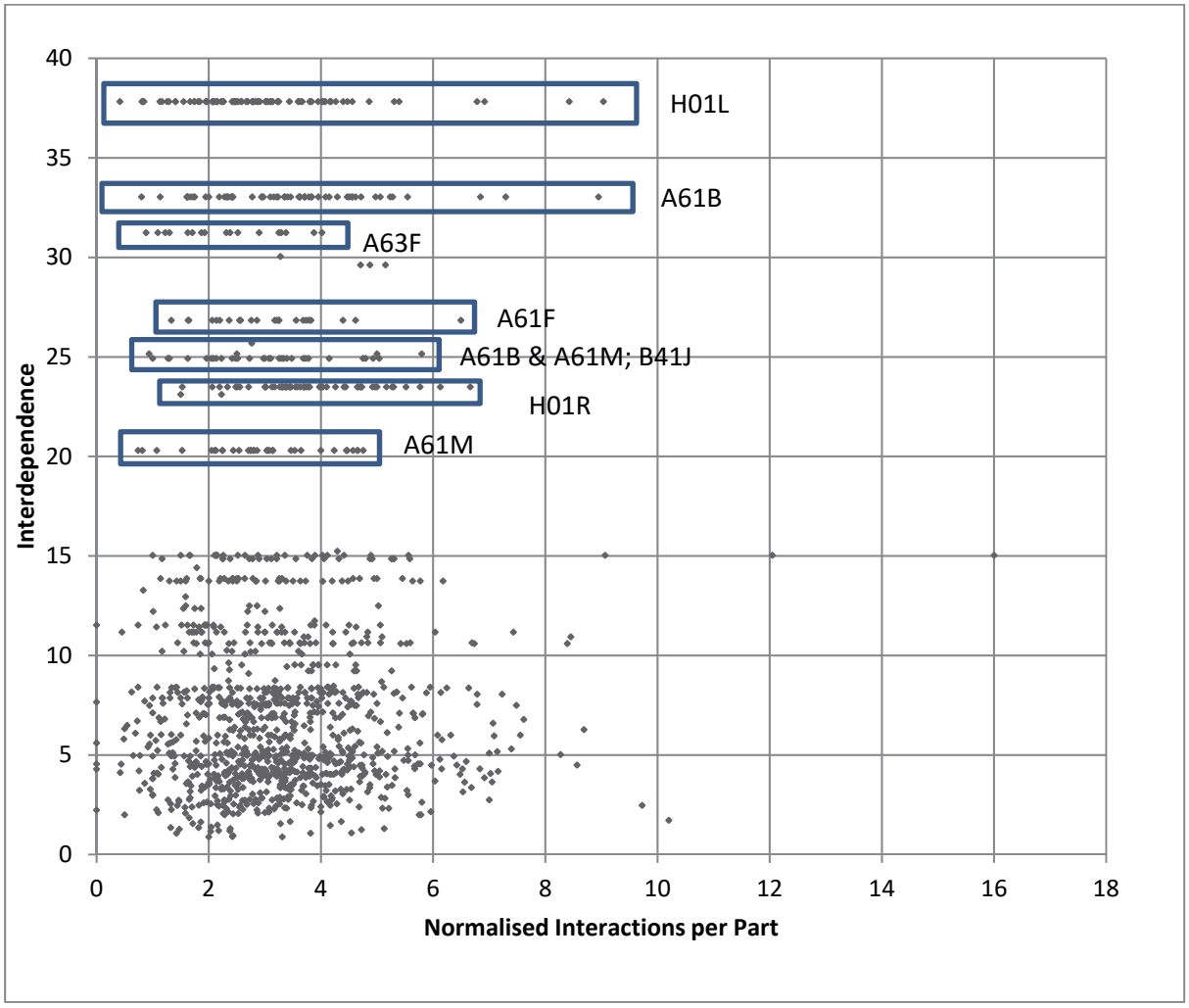


Figure 50: Interdependence vs. Normalised Interactions per part

Sub-class H01L covers semiconductor technology and the manufacture thereof. The sub-class is assigned a total of 13 599 times on 6 885 unique patents. Of the patents 5 980 has assignments exclusively under sub-class H01L. This means that only 13.1% of patents under sub-class H01L has other sub-classes assigned to it. The high volume of patents under the sub-class in combination with the low co-occurrence with other sub-classes would result in a very low ease of recombination, and therefore a high interdependence.

There is some irony in the fact that a high level of interdependence is indeed present *within* the H01L sub-class. The degree of accuracy required to manufacture quality FETs and similar devices is very high. The same level of sophisticated interaction is not present when integrating transistor technology into other systems. This can be demonstrated by looking at the acceptable tolerances of the components. It has been noted that if the concentration of the dopant in a silicone based semiconductor changes by one part in 10^8, its resistance can fluctuate by a factor of 24100 [12]. Most electronic components have tolerances of more than 0.1% and for most applications mush larger tolerances are acceptable. The interdependence within transistors are much greater than that of a transistor operating as a part in a larger system.

It has been noted earlier that the amount of parts and part interactions in a sub-class do not necessarily reflect the average amount of parts and interactions in more granular classifications under that subclass. There is a lot of variance in complexity within sub-classes. The horizontal clustering of patents related to a specific sub-class confirms that classification sub-classes are too aggregated to be used as meaningful complexity measures.

There is no clear relationship between the Interdependence metric and any of the base complexity measures.

6.5 DISCUSSION

The analysis has shown that the effect of embodiments and different naming of similar parts is much smaller than what was anticipated. This implies that most patents describe critical parts only once and that it is indicated under only one tag in the patent figures.

The distribution of parts in patents shows that few patents have a very small amount of parts, half of patents have between 26 and 67 parts and a quarter of patents have more than 67 parts.

The analysis shows that sub-classes are not equally utilised. The most prevalent sub-class was assigned to 17.05% of the sample and the ten most prominent was assigned to only 3.7% of the sample. This creates the problem that only a few sub-classes can be analysed in a statistical significant way. Prevalent classes were analysed and was found to present a broad enough view of the technological landscape.

It was shown that classes differed in coverage scope. Some pertain to a narrower set of innovations while others encompass a vast array of invention types. The variance of scope

was reflected in the variance of parts within the sub-classes. This confirms that classes are not always representations of specific technology streams.

It has furthermore been demonstrated that the complexity within sub-classes is not uniform. Classification codes under the same sub-classes vary quite far from the sub-class mean.

Citation and claim counts both correlate positively with increases in the base complexity measures.

7 CONCLUSION

7.1 REFLECTION

This study set out to determine the relationship between patent metadata and invention complexity in patented inventions. This broad scope was focused by investigating the use and validity of complexity metrics, as proposed by Fleming and Sorenson. To illuminate the relationship between innovation and complexity several smaller questions and tasks were set out, in combination they formed the platform on which the central question could be staged.

The first of these tasks was to investigate the nature and availability of patent data and metadata. The USPTO, EPO and CIPC databases were analysed, both in terms of patent content and format. The South African patent database (CIPC) unfortunately did not meet the requirements of this study, as patent data is provided in image format and no other textual data source could be found. After careful consideration the USPTO was chosen as a data source. The main reason for this was the richness of the XML structure used in USPTO patent documents.

The data within patent documentation was analysed and the data fields that could be of use was catalogued. In particular, patent citations, classifications, prominent title keywords, claims, assignment and the general process from application to grant was considered.

The literature has demonstrated that most metadata fields in patents are used as data sources in a plethora of study fields. No study could be found with a method that leveraged the XML structure in patent documents to assist in the extraction of information.

The next subsection of work considered the usefulness of patent information in the study of technology, innovation and complexity. To achieve this, definitions and basic principles of technology and innovation were explored. The nature of complexity and complex systems were also investigated. It was demonstrated that there is very little consensus on what constitutes a complex system. One of the outflows from this enquiry was the determination that a distinction needs to be made between complexity in a systemic form and the intrinsic complexity within an invention. The main justification for this split in thinking is that technological artefacts have different levels of complexity, but they do not fulfil the definition for a complex adaptive system.

Several modes of innovation were explored and design methods scrutinised. One such method was the TRIZ design principles. It provided insight into problem abstraction and made

several observations on the lifecycles of technology. It was later found that a substantial volume of research focusses on using patent information in combination with TRIZ principles to extract information. Applications ranged from advanced patent search methodologies to solution discovery in design problems and patent value estimation.

A detailed description was given on how Fleming & Sorenson derived and implemented their measures for complexity from patent documents. This formed part of a review of previous work on how to extract complexity metrics from patents. The literature on the subject was sparse, but it did show that the increase in complexity in technology is manifest in the measurements that can be made from patent document metadata.

The last portion of inquiry covered the tools and methods available to extract data from patents. Several excellent software packages were found that could assist in the extraction of natural language structures.

All of these smaller questions were synthesised and the following premise was derived from the literature:

1. Patent data is freely available, and it has been shown that the metadata fields capture aspects of invention complexity.
2. The XML structure of USPTO patents mark part names explicitly, but it would seem that this has never been utilised to facilitate the extraction of part information.
3. A recurring theme in the study of complexity is the number of components in a system and the modes in which they interact.

From these a method was conceived to extract complexity measures from patents. The method proposes a primary set of metrics, or base complexity metrics, and a set of periphery measurements. The base metrics include the interaction between parts and the number of parts in an invention. The base metrics are then validated and verified to ensure that they do capture aspects of innovation. These are then compared to the periphery measurements and tested for correlation. A more focused test is carried out on Fleming & Sorenson's complexity metric. It is compared to the base complexity metrics on a sub-class and per patent level.

The method was then implemented. The implementation underwent several iterations. It was found that the initial language parsing method was impractical and inaccurate. The method was adjusted accordingly –

1. The XML structure contains tags that highlight references to numbered parts in the patent figures. These tags are exclusively used for this purpose.

2. The name of the part can be extracted with reasonable accuracy from these tags.
3. Where two parts co-occur within the same sentence it is assumed that there is some connection between the two, physical or conceptually.

7.2 FINDINGS

The base complexity measures, both part count and interaction, have Poisson-like distributions. It has been noted that the removal of duplicate part names only made nominal differences to the measurements. Two possible explanations are available for the low effect of normalisation. Firstly, patentees describe similar parts with non-overlapping names. This could be done to better the description and also as part of a strategy to widen the patent cover. Secondly, if multiple embodiments are described within a patent it is possible that they differ enough to be deemed separate technological artefacts.

Over a 9-year sample period the average number of parts in patents increased. This increase was gradual and very linear. This validates the notion that the number of parts in an invention is a valid proxy for the complexity of the invention. The linear nature of the complexity increase over time was surprising, and in a world where exponential trends is the advertised norm, a little anticlimactic. From Moor's law and other trends, it is clear that the expanse of knowledge and innovation is not linear, and newer technology branches are evolving even faster. Consider the cost of sequencing a human genome, illustrated in Figure 51. The first time a human genome was sequenced it took 15 years and cost $2.7 billion. Now it can be done for under $1000 and in less than a week. Part of this growth needs to be attributed to the establishment of an economy of scale within the genomics industry, but the greater part comes from innovation activities.

The question remains – why is the increase in base complexity measures linear? Two facts need to be considered before a conclusion can be drawn –

1. One or two examples of spectacular growth cannot be extrapolated to all technology streams.
2. Patent documents have natural limitations. The number of parts per document cannot grow *ad infinitum* at a greater than linear pace.

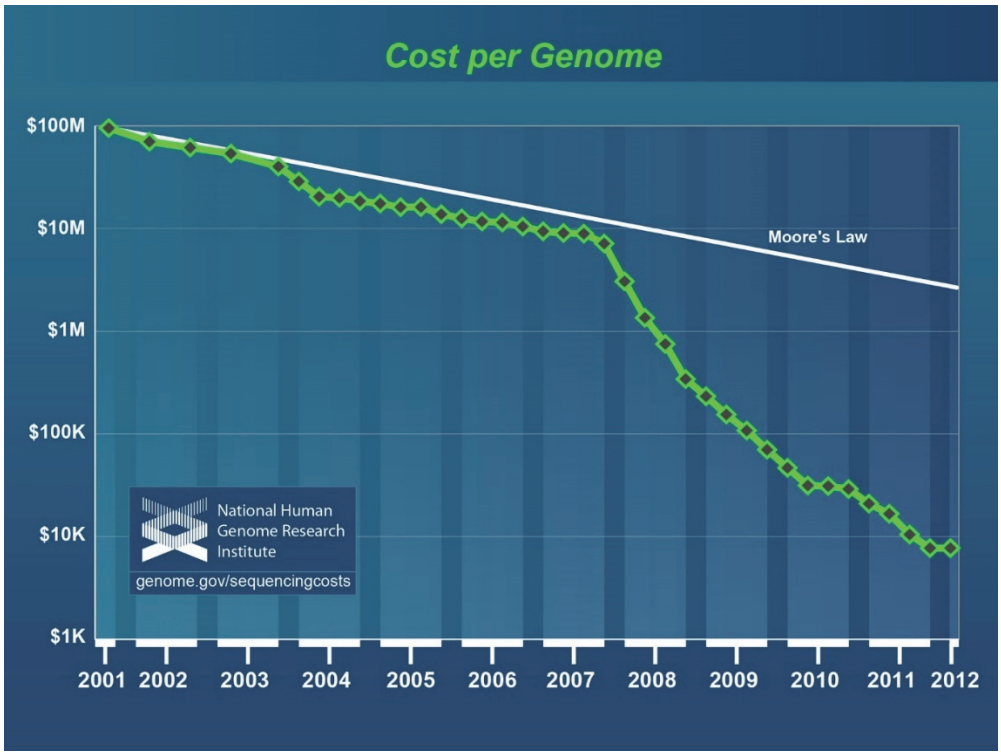


Figure 51: Moor's Law & Genome Sequencing Cost [85]

The answer lies in *modularity* – A continuous adaptation is found in patent documentation where the borders of what constitutes a part is shifted and expanded. A description of a part in a patent could move from *battery* to *power system*. A battery will most likely be less complex than a power system, containing a battery and charging and control circuitry. This continuous adaptation in the borders of modularity implies that counting constituent parts in an invention is not enough to show true growth. A second measurement is needed to measure the rate at which modules within inventions transform and grow. Together these will produce a better quantitative result.

It has been shown that the distributions of parts in patent differ when aggregated by sub-classes in the patent classification scheme. Some of the contributing factors to these differences included the nature of the technology covered by the sub-class and the scope of the sub-class. Some sub-classes focus on a specific technology stream while others describe properties of inventions that can be classified very broadly. For example, sub-class E01D covers the construction and assembly of bridges. Sub-class E04C covers all structural elements and building materials for fixed constructs. The one clearly has a broader scope than the other, but they are at least still focused within the broad category of construction. Then there are sub-classes like B32B covering all "layered products". Cases such as these is not even loosely coupled to any specific technology type. Therefore, it can be concluded

that sub-classes are not a good representation of different technology types. A much more granular approach is required.

The shortcomings of using sub-class information as complexity proxies was highlighted when Fleming and Sorenson's metrics were compared to the base complexity measures. No correlation could be found on either sub-class or per patent level. During the comparison process several weaknesses were revealed in their metric –

1. Normalisation is done by dividing the sum of sub-class co-occurrences by the number of patents in the sample, tagged with a specific sub-class. Even with a massive sample there will still be instances where a sub-class occurs on only one or a hand full of patents. This effect skews the measurement.
2. The metric is based on sub-classes which are not adequate descriptions of technology streams.
3. Internal interdependence is ignored. Many classes, such as H01N, are assigned multiple times to the same patent. For example, patent US7157331 describes an ultra-violet protection layer for semiconductors. It has the following classifications – H01L21/336, H01L21/8234, H01L21/44, H01L31/0288, H01L29/80 and H01L29/76. All of these fall under H01L and the nuance and interplay between these will be lost in this metric.

Several patent metadata fields were investigated. These included citations, claims, publication lag and the presence of certain terms in the patent title. It was shown that citation count and claims count are positively correlated with the amount of parts in the patent. This gives credence to the first research question in that it confirms a relationship between the metrics. It is difficult to elucidate this relationship beyond a qualitative observation. There are several reasons for this –

1. The base complexity metrics are themselves proxies for complexity. This limits accuracy.
2. Patents remain *legal* documents. As such there is no guidance in the documentation process to capture the nuances of a technology in a standardised way. The patent description is free form text and measurements from it will be influenced by the writer's style and motive.
3. It has been shown that a disproportionate number of patents have exactly 20 claims (see Figure 44). This is attributable to a combination of the USPTO's costing structure and patentees trying to make a claim as wide as possible. It is therefore clear that

other systemic forces influence the data in patent documents in non-trivial ways. This instance was highly visible and discernible, but more subtle influences from policy or other sources can have a profound cumulative influence.

In conclusion, aspects of complexity are encoded in patent metadata. Due to the noise in data they only reveal trends in aggregate, and is not very useful to gauge the complexity of an invention on a per patent bases.

Regarding the second research question, Fleming and Sorenson noted that their model was highly sensitive to the number of parts and their interactions. A variation in these can make the difference between an invention being average versus being in the top 6% of successful patents, according to their method. This study has illuminated several shortcomings in their complexity metrics. It is therefore reasonable to conclude that their results are questionable until such time that the experiment is repeated with better data and complexity measures.

7.3 FUTURE RESEARCH

The nature of this study necessitated a broad analysis of several data forms, methods and ideologies. In many cases measurements and snippets from the literature prompted questions and illuminated opportunities for further study. There is also a lot of room to improve the proposed methodology and measures set forth in the preceding chapters.

Natural language processing is not limited to the statistical methods presented in this study. Additional methods can possibly be applied in future refinements of this work. Other methods, such as deep leering can also be applied for more nuanced measurements of the evolution in the world of innovation and technology.

An exact method, and set of well tested complexity metrics, that is applicable to patent data remains elusive. Although this study shows that various measures from patents encapsulate aspects of complexity, it does not provide a quantification of the underlying dynamics of these interactions. There is therefore ample opportunities to search for these elusive relationships.

The broader definition of complexity is another matter that can be further developed. Most definitions focus on the systemic aspects of complexity while the complexity of an invention or object is rarely described. Such a definition, along with the theoretical expansion thereof, could potentially be applied in manufacturing and design activities.

The demonstration that complexity is encoded in patent data creates opportunity for study beyond the field of engineering. One possible application is in quantitative finance, and more specifically in share portfolio management. This could be very valuable if a relationship between a company's stock value and the complexity of its patented inventions exits. In a broader sense it can also enrich the fields of innovation studies and macroeconomics, if the complexity of produced artefacts and the effects of an increase in complexity is studied in aggregate.

7.4 GENERAL OBSERVATIONS

During the course of this investigation several unintended discoveries were made. They are a haphazard collection of observations that do not tie back to the research questions directly, but still hold value.

The first unexpected surprise was the lack of technical infrastructure used to support South African intellectual property rights. Patents are not digitised properly and scanned in as images. Many of them are incomplete and only show some biographical information, but not the actual patent. This makes searching within the database highly impractical. To add to this difficulty, the CIPC patent search engine is badly implemented. The website crashes sporadically and displays the full server-side debug log. One possible remedy to this state is to form a research group tasked with upgrading the CIPCs systems. Research opportunities abound; from OCR projects to properly digitise older patents to the design of a complete new system.

It was noted that the classification hierarchy was much more granular than originally anticipated. Consider the excerpt of classifications shown below. All of the codes presented here are at the lowest level of the hierarchy, but some still fall under others. In other words, all the classifications here imply H01L29/66, even though they are technically on the same level. This should be taken into account in future studies.

H01L 29/66	. Types of semiconductor device {; Multistep manufacturing processes therefor}
H01L 29/66007	. . {Multistep manufacturing processes}
H01L 29/66015	. . . {of devices having a semiconductor body comprising semiconducting carbon, e.g. diamond, diamond-like carbon, graphene}
H01L 29/66022	. . . . {the devices being controllable only by variation of the electric current supplied or the electric potential applied, to one or more of the electrodes carrying the current to be rectified, amplified, oscillated or switched, e.g. two-terminal devices}
H01L 29/6603	. . . . . {Diodes}
H01L 29/66037	. . . . {the devices being controllable only by the electric current supplied or the electric potential applied, to an electrode which does not carry the current to be rectified, amplified or switched, e.g. three-terminal devices}
H01L 29/66045	. . . . . {Field-effect transistors}
H01L 29/66053	. . . {of devices having a semiconductor body comprising crystalline silicon carbide}
H01L 29/6606	. . . . {the devices being controllable only by variation of the electric current supplied or the electric potential applied, to one or more of the electrodes carrying the current to be rectified, amplified, oscillated or switched, e.g. two-terminal devices}
H01L 29/66068	. . . . {the devices being controllable only by the electric current supplied or the electric potential applied, to an electrode which does not carry the current to be rectified, amplified or switched, e.g. three-terminal devices}
H01L 29/66075	. . . {of devices having semiconductor bodies comprising group 14 or group 13/15 materials (comprising semiconducting carbon H01L 29/66015; comprising crystalline silicon carbide H01L 29/66053)}
H01L 29/66083	. . . . {the devices being controllable only by variation of the electric current supplied or the electric potential applied, to one or more of the electrodes carrying the current to be rectified, amplified, oscillated or switched, e.g. two-terminal devices}

This study has demonstrated that the complexity of innovation is increasing. The linear nature of the increase in part count within patents indicate that the parts in patents are themselves getting more complex. This effect will also then increase the complexity in the innovation process and require greater areas of specialisation in engineering and other technical vocations.

A last observation is that patents are not equal units of innovation. Some will contribute more to both technological and economic growth than others. Consider the patent drawings shown in Figure 52. The first is described as a “Banana protective device” and the second as a “Method of exercising a cat”. The first can be made into a product and sold, to what would probably be a small niche market. The rights on the second patent would be near impossible to enforce – it would be very hard to collect royalties from all individuals who point laser pointers at cats. The value of these patents pale in comparison to others describing, for example, the design of an internal combustion engine or smartphone camera. The inequality of patents, both in monetary and inventive terms, adds a great deal of uncertainty to technology studies.

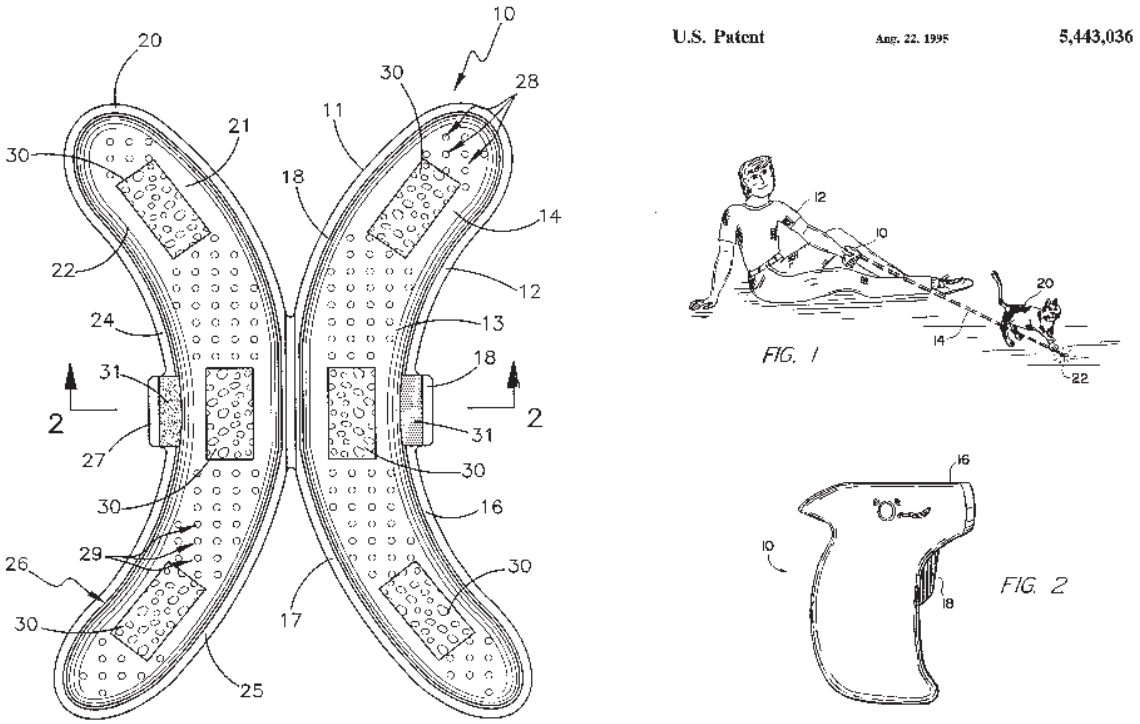


Figure 52: Drawings from Patents US5443036 and US6612440

8 WORKS CITED

[1] M. Frumkin, “The Origin of Patents,” *Journal of the Patent Office Society,* vol. 27, no. 3, pp. 143-149, March 1945.

[2] Government of Canada, “Justice Laws Website,” 21 09 2017. [Online]. Available: http://laws-lois.justice.gc.ca/eng/acts/P-4/FullText.html. [Accessed 28 10 2017].

[3] USPTO, “First U.S. Patent Issued Today in 1790,” 2001.

[4] World Intelectual Property Organisation, February 2017. [Online]. Available: https://www3.wipo.int/ipstats/index.htm.

[5] World Bank, “Data: The World Bank,” 2015. [Online]. Available: http://data.worldbank.org/indicator/IP.PAT.RESD/countries/1W?display=graph. [Accessed 19 11 2015].

[6] European Patent Office; European Commission, “Why researchers should care about patents,” 2007. [Online]. Available: http://ec.europa.eu/invest-in-research/pdf/download_en/patents_for_researchers.pdf.

[7] Z. Griliches, “Market value, R&D, and patents,” *Economics Letters,* vol. 7, no. 2, pp. 183-187, 1981.

[8] A. Jaffe, M. Trajtenberg and R. Henderson, “Geographic localization of knowledge spillovers as evidenced by patent citations.,” *Quarterly Journal of Economics,* vol. 108, no. 3, pp. 577-598, 1993.

[9] M. Sakakibara and L. Branstetter, “Do stronger patents induce more innovation? Evidence from the 1988 Japanese patent law reforms,” *RAND Journal of Economics,* vol. 32, no. 1, pp. 77-100, 2001.

[10] G. Moor, “Cramming more components onto integrated circuits,” *Electronics,* vol. 38, no. 8, April 1965.

[11] L. Fleming and O. Sorenson, “Technology as a complex adaptive system: evidence from patent data,” *Research Policy,* no. 30, pp. 1019-1039, 2001.

[12] J. Millman, Microelectronics: digital and analog circuits and systems, New york: McGrawHill, 1979, p. 13.

[13] J. Kim and S. Lee, “Patent databases for innovation studies: A comperative analysis of USPTO, EPO, JPO and KIPO,” *Technological Forecasting & Social Change,* no. 92, pp. 332-345, February 2015.

[14] A. Pouris and A. Pouris, “Patents and economic development in South Africs: MAnaging intellectual property rights,” *South African Journal of Science,* vol. 11, no. 107, pp. 24-33, November 2011.

[15] M. White, “Anatomy of a Patent,” May 2008. [Online]. Available: http://library.queensu.ca/webeng/patents/Anatomy_of_a_US_patent.pdf.

[16] WIPO, “What is a Patent?”.

[17] S. Gilfillan, Inventing the Ship, Chicago: Follett Publishing Co., 1935.

[18] J. Schumpeter, Business Cycles, New York Toronto London: McGraw-Hill , 1939.

[19] H. Chesbrough, Open Innovation: The New Imperative for Creating and Profiting from Technology, Harvard Business Press, 2003.

[20] K. Barry, E. Domb and M. Slocum, 2010. [Online]. Available: http://www.triz-journal.com/triz-what-is-triz/.

[21] C. Y. Baldwin and K. B. Clark, Design Rules: The power of modularity. Vol 1, MIT Press, 2000.

[22] J. H. Hollard, “Complex Adaptive Systems,” *Daedalus,* vol. 121, no. 1, pp. 17-30, December 1992.

[23] J. Ladyman, J. Lambert and K. Wiesner, “What is a complex system?,” *European Journal for Philosophy of Science,* vol. 3, no. 1, pp. 33-67, January 2013.

[24] “Various,” *Science,* vol. 284, no. 5411, April 1999.

[25] T. Montecchi, D. Russo and Y. Liu, “Searching the Cooperative Patent Classification: Comparison between keyword and concept-based search,” *Advanced Engineering Informatics,* vol. 1, no. 27, pp. 335-345, April 2013.

[26] M. Montecchi and D. Russo, “FBOS: Function/Behaviour–Oriented Search,” *Procedia Engineeing,* vol. 1, no. 131, pp. 140-149, 2015.

[27] B. Hall, A. Jaffe and M. Trajtenberg, “Market value and patent citations,” *Rand Journal of Economics,* 2005.

[28] H. Park, J. Ree and K. Kim, “Identification of promising patents for technology transfers using TRIZ evolution trends,” *Expert Systems with Applications,* vol. 1, no. 40, pp. 736-743, 2013.

[29] J. Luo and K. Wood, “The growing complexity in invention process,” *Research in Engineering Design,* vol. 28, no. 4, pp. 421-435, October 2017.

[30] C. D. Manning and H. Schutze, Foundations of Statistical Natural Language Processing, 2nd Edition ed., H. Schutze, Ed., Cambridge: MIT Press, 1999.

[31] A. Uysal and S. Gunal, “The impact of preprocessing on text classification,” *Information Processing & Management,* vol. 50, no. 1, pp. 104-112, January 2014.

[32] T. Brants, “Part-of-Speech Tagging,” in *Encyclopedia of Language & Linguistics*, K. Brown, Ed., Elsevier, 2006, pp. 221-230.

[33] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard and D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” in *Trancactions of the ACL*, Baltimore, 2014.

[34] G. Miller, “WordNet: A Lexical Database for English,” *Communications of the ACM,* vol. 37, no. 11, pp. 39-41, 1995.

[35] European Patent Office, 1 2016. [Online]. Available: http://www.epo.org/searching/data/data/patent-additions.html.

[36] A. Khval, D. Katsubo, P. Passarelli, S. Szymura and D. Afriat, “Open Patent Services RESTful Web Services,” 2015.

[37] REED Tech, January 2016. [Online]. Available: http://patents.reedtech.com/pgrbft.php.

[38] SA Department of Trade and Industry, 2014. [Online]. Available: http://patentsearch.cipc.co.za/.

[39] A. Abbas, L. Zhang and S. Khan, “A literature review on the state-of-the-art in patent analysis,” *World Patent Information,* vol. 37, no. 1, pp. 3-13, 2014.

[40] WIPO, “International Patent Classification Guide,” 2015. [Online]. Available: http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf.

[41] C. Harris, R. Arens and P. Srinivasan, “Comparison of IPC and USPC classification systems in prior art searches,” in *Procedings of the 3rd international workshop on Patent Information Retreval*, Toronto, 2010.

[42] J. List, “Editorial: On Patent Classification,” *World Patent Information,* vol. 1, no. 41, pp. 1-3, 2015.

[43] USPTO; EPO, “CPC Annual Report,” 2014. [Online]. Available: http://www.cooperativepatentclassification.org/publications/2014CPCAnnualReport.pdf. [Accessed 19 January 2016].

[44] J. Baruch, L. Parker and W. Lawson, “Technology assessment & forecast,” Washington, 1977.

[45] H. Mucke, “Relating patenting and peer-review publications: an extended perspective on the vascular health and risk management literature,” *Vasc Health Risk Manag.,* vol. 1, no. 7, pp. 265-276, April 2011.

[46] Oxford University Press, 2017. [Online]. Available: https://en.oxforddictionaries.com/definition/technology.

[47] Merriam-Webster, Incorporated, 2017. [Online]. Available: https://www.merriam-webster.com/dictionary/technology.

[48] A. Ganguly, R. Nilchiani and J. V. Farr, “Defining a Set of Metrics to Evaluate the Potential Disruptiveness of a Technology,” *Engineering Management Journal,* vol. 22, no. 1, pp. 34-44, March 2010.

[49] K. e. a. Catchpole, “Patient handover from surgery to intensive care: using Formula 1 pit-stop and aviation models to improve safety and quality,” *Pediatric Anesthesia,* no. 17, p. 470–478, 2007.

[50] H. Chesbrough, W. Vanhaverbeke and J. West, Eds., Open innovation: Researching a new paradigm, Oxford: Oxford university press, 2006.

[51] Google, [Online]. Available: https://scholar.google.co.za/scholar?q=ISBN+1422102831.

[52] A. Kovacs, B. Van Looy and B. Cassiman, “Exploring the scope of open innovation: a bibliomentric review of a decade of research,” *Scientometrics,* vol. 104, pp. 951-983, 2015.

[53] V. Petrov, “The laws of system evolution,” *The TRIZ Journal,* no. 3, pp. 9-17, 2002.

[54] Alphabet, July 2017. [Online]. Available: https://www.google.co.za/search?q=define+complexity.

[55] N. Goldenfeld and L. Kadanoff, “Simple Lessons from Complexity,” *Science,* vol. 284, no. 5411, pp. 87-89, April 1999.

[56] M. W. Whitesides and R. F. Ismagilov, “Complexity in Chemistry,” *Science,* vol. 284, no. 5411, pp. 89-92, April 1999.

[57] G. Weng, U. Bhalla and R. Iyengar, “Complexity in Biological Signalling Systems,” *Science,* vol. 284, no. 5411, pp. 92-96, April 1999.

[58] J. Parrish and L. Edelsein-Keshet, “Complexity, Pattern and Evolutionary Trade-Offs in Animal Aggregation,” *Science,* vol. 284, no. 5411, pp. 99-101, April 1999.

[59] D. Rind, “Complexity and Climate,” *Science,* vol. 284, no. 5411, pp. 105-107, April 1999.

[60] W. Arthur, “Complexity and the Economy,” *Science,* vol. 284, no. 5411, pp. 107-109, April 1999.

[61] H. Simon, “The Architecture of Complexity,” *Proceedings of the American Philosophical Society,* vol. 106, no. 6, pp. 467-482, Dec 1962.

[62] Editorial, “No Man is an Island,” *Nature Physics,* vol. 5, no. 1, January 2009.

[63] The Lean Methods Group, 1996-2007. [Online]. Available: https://triz-journal.com.

[64] D. Russo and T. Montecchi, “Creativity techniques for a computer aided inventing system,” in *International Conference on Engineering Design (ICED11)*, Copenhagen, 2011.

[65] T. Dao and T. Simpson, “Measuring Similarity between sentences,” *WordNet Tech. Rep.,* 2005.

[66] S. Aharonson and M. Schilling, “Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution,” *Research Policy,* no. 45, pp. 81-96, 2016.

[67] M. Roach and W. Cohen, “Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research,” *Management Science,* vol. 2, no. 59, pp. 504-525, February 2013.

[68] C. Manning, P. Raghavan and H. Schutze, An Introduction to Information Retreval, Online Edition - http://www-nlp.stanford.edu/IR-book/ ed., Cambridge: Cambridge University Press, 2009.

[69] Y. Chen and Y. Chang, “A three-phase method for patent classification,” *Information Processing & Management,* vol. 48, no. 6, pp. 1017-1030, November 2012.

[70] D. Eisinger, J. Monnich and M. Schroeder, “Developing Semantic Search for the Patent Domain,” in *Proceedings of the First International Workshop on Patent Mining and its Applications (IPAMIN)* , Hildesheim, 2014.

[71] J. Gomez and M. Moens, Professional Search in the Modern World, Springer International Publishing, 2014, pp. 215-249.

[72] A. Ratnaparkhi, “A Simple Introduction to Maximum Entropy Models for Natural Language Processing,” Philadelphia, 1997.

[73] T. Brants, “TnT - a statistical part-of-speech tagger,” in *Sixth Conference on Applied Natural Language Processing ANLP-2000*, Seattle, 2000.

[74] K. Toutanova, D. Klein, C. Manning and Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-03)*, Edmonton, 2003.

[75] A. Taylor, M. Marcus and B. Santorini, “The Penn treebank: An overview,” in *Treebanks: Building and Using Parsed Corpora* , A. Abeille, Ed., York, Springer, 2003, pp. 5-22.

[76] D. Slowmin and R. Tengi, “WordNet Browser - A graphical interface to the WordNet online database,” 2005.

[77] C. Brown, “Decompisition and prioritization in engineering design,” in *Proceedings of the Sixth International Conference on Axiomatic Design*, Daejeon, 2011.

[78] D. B. Kitfield, “Fully articulable shower curtain rod”. USA Patent US8925122, 6 January 2015.

[79] J. Frijters, June 2014. [Online]. Available: http://www.ikvm.net.

[80] The Stanford Natural Language Processing Group, November 2017. [Online]. Available: http://sergey-tihon.github.io/Stanford.NLP.NET/.

[81] Oracle, 2017. [Online]. Available: https://dev.mysql.com/downloads/.

[82] C. Poole and R. Prouse, 2017. [Online]. Available: http://nunit.org/.

[83] A. Reeves and C. Dahlberg, 2017. [Online]. Available: https://github.com/StyleCop.

[84] U1SPTO, January 2014. [Online]. Available: https://www.uspto.gov/sites/default/files/documents/USPTO%20fee%20schedule_current_6.pdf.

[85] National Human Genome Research Institute, “DNA Sequencing Costs: Data,” NIH, 31 October 2017. [Online]. Available: https://www.genome.gov/sequencingcostsdata/. [Accessed 03 December 2017].

[86] H. Park, J. Yoon and K. Kim, “Using function-based patent analysis to identify potential application areas of technology for technology transfer,” *Expert systems with applications,* no. 40, pp. 5260-5265, 2013.

[87] J. Scannell, A. Blanckley, H. Boldon and B. Warrington, “Diagnosing the decline in pharmaceutical R&D efficiency,” *Nature Reviews Drug Discovery,* vol. 11, pp. 191-200, March 2012.

[88] Symantec Corporation, “Symantec Intelligence Report - June 2015,” Mountain View, 2015.

[89] C. Laorden, X. Ugarte-Pedrero, I. Santos, B. Sanz and J. B. P. Nieves, “Study on the effectiveness of anomaly detection for spam filtering,” *Information Sciences,* no. 277, pp. 421-444, 2014.

[90] WIPO, "WIPO Patent Drafting Manual," [Online]. Available: http://www.wipo.int/edocs/pubdocs/en/patents/867/wipo_pub_867.pdf. [Accessed 18 January 2016].

[91] A. Gregory, "The phones of Dr. Moreau," New Yorker, New York, 2014.

[92] The Research Institute for Innovation and Sustainability (RIIS), December 2014. [Online]. Available: http://www.openix.theinnovationhub.com/index.php?page=view_challenge&challenge_id=36.

[93] Prinston University, "WordNet 3.0 Reference Manual," 2010. [Online]. Available: https://wordnet.princeton.edu/wordnet/documentation/.

[94] D. Ariely, January 2013. [Online]. Available: https://twitter.com/danariely/status/287952257926971392.

[95] B. Hall, "Market value and patent citations," *Rand Journal of Economics,* no. 36, 2005.