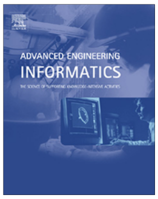
Advanced Engineering Informatics 27 (2013) 317–334

Contents lists available at SciVerse ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



Automatic extraction of function–behaviour–state information from patents

G. Fantoni [a,*], R. Apreda [b,d], F. Dell'Orletta [c], M. Monge [d]

[a] Department of Mechanical, Nuclear and Production Engineering, University of Pisa, Largo Lucio Lazzarino, 2, 56126 Pisa, Italy
[b] Department of Energy and Systems Engineering, University of Pisa, Largo Lucio Lazzarino, 2, 56126 Pisa, Italy
[c] Istituto di Linguistica Computazionale "Antonio Zampolli", ILC–CNR, via G. Moruzzi, 1 Località S. Cataldo, 56124 Pisa, Italy
[d] Erre Quadro s.r.l., via S. Andrea, 59, I-56122 Pisa, Italy

ARTICLE INFO

Article history:
Available online 18 June 2013

Keywords:
Patent informatics
Semantic elaboration
Function–Behaviour–Structure
Product development

ABSTRACT

Patents contain a large quantity of technical information not available elsewhere and therefore very interesting for both academia and industry. The purpose of the research is to try to detect and extract information about the functions, the physical behaviours and the states of the system directly from the text of a patent in an automatic way. The above three categories constitute a well-known set of relevant entities in the theory of engineering design, and their study allows powerful analysis of individual artefacts as well as that of groups of products or technologies. The focus is in providing a handy tool that could speed up and facilitate human analysis and allow tackling also large corpora of documents. A second goal is to develop a protocol based on free software and database resources, so that it could be replicable with limited effort by everyone without having to rely on commercial databases.

Extracting technical and design information from a document whose aim is more legal than technical, and that is written using a specific jargon, is not a trivial task. The approach chosen to overcome the various issues is to support state-of-the-art Computational Linguistic tools with a large Knowledge Base. The latter has been constructed both manually and automatically and comprises not only keywords but also concepts, relationships and regular expressions. A case study about a very recent patent describing a mechanical device has been included to show the functioning and output of the entire system.

1. Introduction

Patent literature is growing at an impressive speed, causing an overflow of information and an increased difficulty of performing technological surveys [1,2].

Moreover, intentional IP policies such as patent hiding, patent proliferation, bombing, etc. contribute to the generation of confusion and to loss of time both in research and in analysis.

A side effect of such explosion affects also the performance of patent examiners: actually their available time per patent to be analysed is decreasing and as a consequence the quality of both search and analysis diminishes as well [3,4].

Patent writing is understandably a challenging job, but patent reading needs for a quite long experience as well. That is due not only to the legal jargon used there, but also to the will of disclosing as little as possible of the patenting device. Information are intentionally left to the bare minimum needed to make the device capable of being understood and the claimed invention of being replicated without undue experimentation by a person with an ordinary skill in the art (as requisite for patentability), but no additional information is provided in order to limit the possibility of design around.

Therefore, given the above-mentioned problems, and since the existence of a plethora of existing and expired patents that can be related to those under study, it is clear that human analysis of patents needs for a huge amount of human labour [5].

On the other hand information in patents are of fundamental importance, and not only for legal or intellectual property rights purposes.

Quite often, information in patents cannot be found elsewhere since the interest of some inventors is more on hiding than in disclosing [6], thus no papers or technical documents can be found except for the patent itself.

Moreover, patents are a source of technical data or descriptions that are essential, precise and specific to the domain under study. Finally the overall patent archive accessible from the various international patent offices websites concentrates in only one database a huge (almost complete) amount of technical solutions and inventive ideas coming from all over the world, temporally ordered and

* Corresponding author. Tel.: +39 3286171576; fax: +39 0502218069.
E-mail addresses: g.fantoni@ing.unipi.it (G. Fantoni), apreda.riccardo@gmail.com (R. Apreda), felice.dellorletta@ilc.cnr.it (F. Dell'Orletta), maurizio.monge@sns.it (M. Monge).

1474-0346/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aei.2013.04.004

thoroughly organised and classified. A similar unified and structured database does not exist for example for scientific papers.

Therefore patent literature represents a necessary complement to the traditional technical and scientific literature, while patent repositories constitute an invaluable tool in engineering design, supporting tasks such as the representation and modelization of technologies, the study and the foresight of technological evolution, or even the generation of new ideas and problem solutions [7,8].

Needless to say, the proper solution to make the huge amount of data in patent literature manageable and available for the above mentioned engineering purposes, is the semiautomatic processing of the raw data by software tools: they can extract and aggregate the useful information that human experts will then use for the subsequent analysis. Indeed, totally automated solutions are not yet sufficiently accurate to allow getting totally rid of human intervention, while on the other hand the prohibitive amount of work required to process documents makes a completely human hard to apply, even on a relatively small subset of the published patents.

The system we implemented has the aim of rapidly supplying conceptual information to the technicians who want to investigate a patent or a patented technology (i.e. a group of closely related patents).

The engineering design model of reference we are interested in is the so-called Function–Behaviour–State (FBS) model, a well-known and very powerful methodology to represent and modelize products and processes. The proper study of functions, linked to structural and causal information, allows an abstract and yet rigorous description of artefacts, that can be used for many useful design activities such as product comparison, innovation generation through the study of functional variants or technological transfer, technology foresight and so on.

The FBS approach has a large consensus in the engineering design community from the theoretical point of view, but its actual application during everyday practice is still limited, due to some hindering factors.

First of all the manual analysis of a technical description to create a full functional, behavioural and structural representation is a very time consuming activity, of order of several hours for artefacts even of medium complexity. If cumbersome for one patent, manual analysis becomes impossible to apply on an entire set of products. Secondly, the possibility of errors, ambiguities, and even different interpretations is relatively high and even expert analysts can come up with slightly different representations of the FBS information. Finally, there are several variants of the FBS paradigm, and depending on the particular focus of the analysis, one approach can be more useful than others, but there is no easy, standard procedure to convert a representation already obtained with one model to another representation, equivalent but based on a different model.

The idea of the present research is to develop a methodology (rather than a single software tool) that would allow the designer to automatically extract and visualise information about the functions, behaviours and states of a device or artefact in the form of graphs that can be read alongside the drawings of the invention. This way the information about the device layout and architecture is shown by the drawings while the properties of each component, the functional interactions and the physical, chemical, logical relations (behaviours) are displayed through Functional-Behaviour-Structure graphs.

An important feature of the tool is that it provides within a unique global picture or in single separate pictures where the FBS relationships are shown separately.

The other main characteristic is to be flexible, allowing including different variants of the FBS approach, such as states, properties, features, etc. in an easy way.

Clearly such tool would address all three of the previously described issues: it would speed up the construction of the FBS representation and therefore the analysis; it will reduce the possibility of errors and ambiguities and help different interpretations to converge; it will also allow, given the proper conversion procedure, to shift automatically from one representation to the other.

Of course the extraction of meaningful knowledge is not an easy task, even given the advancements in the field of Natural Language Processing, but the particular nature of the patent documentation, rigidly structured and based on technical concepts with determined characteristics, allows finding relevant entities with analysis of lower complexity with respect of ordinary texts. We will review some of the issues to be solved and how we have tackled them in the dedicated section.

An additional, but not minor, goal of our research was to elaborate a procedure that could be replicated rather easily, or at least with contained effort, by any interested designer, being based only on free software and free databases, in contrast with analogous existing approaches. Indeed there are already various attempts to extract some variant of FBS information from patents in the literature, usually with good performances, but they rely partly on commercial, proprietary knowledge databases. The developed approach may still present a limited degree of imprecision that has to be removed through manual refinement of the outputs (but we note here that commercial databases guarantee good performances precisely because they have been largely revised manually at the source), but is fast, reliable and above all simple. Indeed the rigid nature of patent language and structure allowed using the simplest rules and patterns to the maximum effect.

We stress that the research work puts in synergy engineering and computational linguistic approaches, but although we are using state-of-the-art NPL software and even if some software tools have been specifically adapted for the particular nature of the present investigation, the advancement mainly concerns and benefit the engineering design community. Still, the application of rather new NPL techniques to such a particular context as that of technical patents is a good test for their flexibility and analytical power.

The paper is then organised as follows. Section 2 gives an overview of the FBS model and discusses how it can be related to patents using software tools. In particular Sections 2.1 and 2.2 describe the FBS approach and its main features, and explain why such model and patent analysis can strengthen each other; Section 2.3 briefly discusses the issues and the state of the art of the automatic processing of documents written in natural language; Section 2.4 reviews the existing approaches to the automatic extraction of FBS information from patents. Section 3 presents the proposed methodology; particular care has been given to the knowledge database construction. In Section 4 the method and its outcome are discussed, and applied to a case study to show its potentialities. Finally Section 5 concludes with an overview of future developments.

2. FBS: a design model well suited for patent analysis

2.1. Design theory and patents

Studying patents is not useful only for intellectual property rights purposes, i.e. to protect and valorise a specific product and with just industrial applications in mind.

It can be relevant for engineering design as well, both from the theoretical and the practical point of view.

Indeed one of the goals of the research in the field of engineering design is to construct a comprehensive model of human artefacts that would allow representing them in abstract and yet precise and rigorous terms.

The definition of such theoretical framework to describe products and processes would significantly improve many important design activities, such as product comparison, retrieval of information about previous solutions, knowledge sharing and standardisation, and would even support the generation of new ideas, thanks to the systematic exploration of design variants and alternatives.

On the other side, patents are among the best sources of information for innovative solutions, technological trends, and domain specific information; sometimes, as already mentioned, they might even be the only information source available on a particular technology tout-court.

It is then clear that the benefits of processing technical documents, and patents in particular, in the context of engineering design are numerous and bi-directional.

First, applying a representation coming from the theory of design will help better organise both the single patent content or the aggregate information coming from a large number of patents.

Visualisation of patents, patent clustering, patent comparison, automatic summarisation, identification of recurring solutions and many other tasks can be performed more efficiently.

Second, meta-analyses such as technology foresight and the individuation of technological trends and potentialities can be performed more rigorously.

Third, it would be possible to construct repositories of products and individual technical solutions that, in turn, would help comparative studies, product improvement and idea generation.

Finally, theories of design are usually top-down constructions and, although there are many valid approaches, there is no such thing as a universally accepted canon, or a conclusive experimental proof of sort. Conversely, the huge patent corpus, aggregating the results of a century of human inventive activity in all engineering fields, constitutes the perfect "test bench" for any design theory.

Thus another advantage of automatic patent analysis is the possibility to validate the ontologies of the various design models, or to help improving them, this time following a bottom-up procedure; on a larger scale it can even provide a way to compare and benchmark different design approaches.

2.2. The FBS framework

Many theoretical models or frameworks have been proposed in engineering design over the years. Some focus more on the technical aspects, others on the interaction with the user; some are very prescriptive and analytic while others try to stimulate intuitive reasoning, and so on.

One of the most efficient and complete approaches is called Functional Analysis, since the key elements are assumed to be the activities (functions) that the artefact must perform to achieve its desired outputs.

Studying functions as independent entities allows abstracting from the particular technical solutions and constitutes therefore a very powerful design strategy. Moreover functions capture both the physical action that produces the output and the desired goal as seen from the user point of view.

Functional Analysis is more a paradigm, comprising several methodologies sharing a common philosophy and some common aspects, rather than a specific methodology in itself.

This is not the place to enter a full review of the various models (see Erden et al. [9] for more on the subject); in the present paper we focus on the so-called FBS model and its variants, since it is one of the most articulated and complete.

During the 1990s the Function Behaviour Structure approach has been proposed as a theoretical framework to analyse products by Umeda et al. [10], and shortly after reframed as Function Behaviour State in Tomiyama et al. [11] and Umeda et al. [12,13], in order to shift from a "device centric" point of view to an "event based" one. The approach has been adopted and modified by several authors (starting from Gero et al. [14,15]) and it is considered of great interest, among other reasons because it allows modelling cognitive design aspects. Here we refer mainly to Umeda et al. approach (with reference to the formulation of [13]), more formal than Gero's one, and more suitable for automatic patent analysis.

The FBS model basically assumes that the three entities that constitute the acronym, and their mutual relationships, encode all the relevant information about a product or a process.

During the paper we adopt the following definitions of the key components of the FBS ontology. Some of these definitions can be found already in the original works by Umeda et al. and in those by Gero; others have been derived from works on qualitative physics [16,17] and from two recent works [18,19,29] that complete the FBS framework. According to such extended ontology a system can be abstracted and decomposed into the following entities.

2.2.1. Needs

The exigencies from where the very existence of the artefact is originated. While in engineering they are treated as the voice of customers, or as external data from marketing, and then converted into engineering requirements, in patents sometimes they are explicitly mentioned in order to explain the novelty of the patented artefact or method [17].

2.2.2. Goals

Every product is designed and manufactured with the precise purpose of satisfying certain needs (of any kind: material, spiritual, social) of the user. The product's aim at addressing a specific need is conventionally referred to as a goal. And the way of addressing an existing problem and satisfying a specific goal is the aim of a patented device.

2.2.3. Functions

They are the interpretation of physical behaviours according to the user's goals. While Umeda et al. [13] define the functions as "descriptions of behaviour recognised by a human through abstraction in order to utilise it", Gero [15] describes them as the motivation for the product existence or, more generally, ascribes them to teleology (what the object is for).

2.2.4. Behaviours

Behaviours are the "physical phenomena" that cause the change of the "states" of the system. In our view behaviours are the descriptions in natural language of the equations (belonging to physics, chemistry, mechanics, etc.) that describe the evolution of a system.

2.2.5. Scenes

Homogeneous groups of phases belonging to the same life cycle stage, where **Phases** are homogeneous set of functions belonging to/performed by the same components and characterised by the same physics/chemistry/logics [20]. The concept of scene is very close to that of History by Hayes [16] and it is fundamental to trace the logical and temporal evolution of the functioning into an artefact.

2.2.6. States

A state "is a property at an instant of time of a system (and environment), that is involved in an interaction between a system and its environment. As a consequence of an interaction [behaviour], the property of a system (and environment) changes and this is called a state change" [13].

A state corresponds to a particular set of entities, attributes of entities, and relations between entities. The concept of states is closely related to that of **Structures**. In fact, in Umeda et al. [13] the authors argue there is no meaningful distinction between "state" and "structure". They claim that there is no difference except for the duration: structures that change in a short time are usually called as states. Somasekhara and Chakrabarti describe each product state through the evolutions of the parameters characterizing each state.

Both Chakrabarti [21] and Russo [22] introduced also the **Physical Effects** in the view. In our opinion a Physical Effect is (theoretically) a property while its manifestations are instead behaviours and so its definition is redundant with respect to the FBS approach. However, more pragmatically, Physical Effects (e.g. Joule's effect, Coulomb's law of friction, Hertz's contact etc.) provide remarkable keywords that imply both structure and behaviour information simultaneously.

Continuing the discussion about state [13] and structure [15] we cite also Wie [23], where "Structure is the most tangible concept with various approaches to partitioning structure into meaningful constituents such as features [24] wirk elements [25] and interfaces [26] in addition to the widely used assemblies and components."

2.2.7. Features

In agreement with Umeda et al. [13] we think that the term structure, even if correct, is misleading since it pushes the designer to focus on physical entities (subassemblies and components) instead of on their parameters. Therefore we prefer the term feature defined as "the specific characteristics of a single part of the product, in terms of the geometrical entities that define it and in terms of the chemical, physical, mechanical, biological, etc. properties of the material it is composed of (e.g. the Young's modulus, resistance to acids or to flames, transparency to certain light frequencies, thermal or electrical conductivity, porosity, etc.). Such concept of feature [23] is very close to Suh's Design Parameter [27] but is even closer to designer's lexicon and common understanding [20,28].

We summarise in Fig. 1 how the various authors in the literature have intended the FBS model and how the various entities in the various approaches are correlated.

Fig. 1 shows the different views of the various authors that have contributed to the FBS approach, and how and where the same definitions of the various entities are adopted. More precisely, the same colour represents the same (or very similar) definition. A few arrows highlight some of the numerous existing dependencies (for a detailed analysis please refer to Erden et al. [9]).

2.3. Computer aided systems for patent analysis

2.3.1. Problems related to natural language processing

Patents, as many other technical documents, are written in plain English (or other national languages); the automatic processing of Natural Language however is not a straightforward activity and presents several difficulties. Consider that computers see texts in input just as a continuous string of characters, while the user wants to extract structured information as output. Thus it is necessary to start from the very elementary tasks, such as teaching the machine how to divide the string into separate words and into sentences, up until the more sophisticated ones such as to recognise the role of a word in the sentence and its relationship with other words.

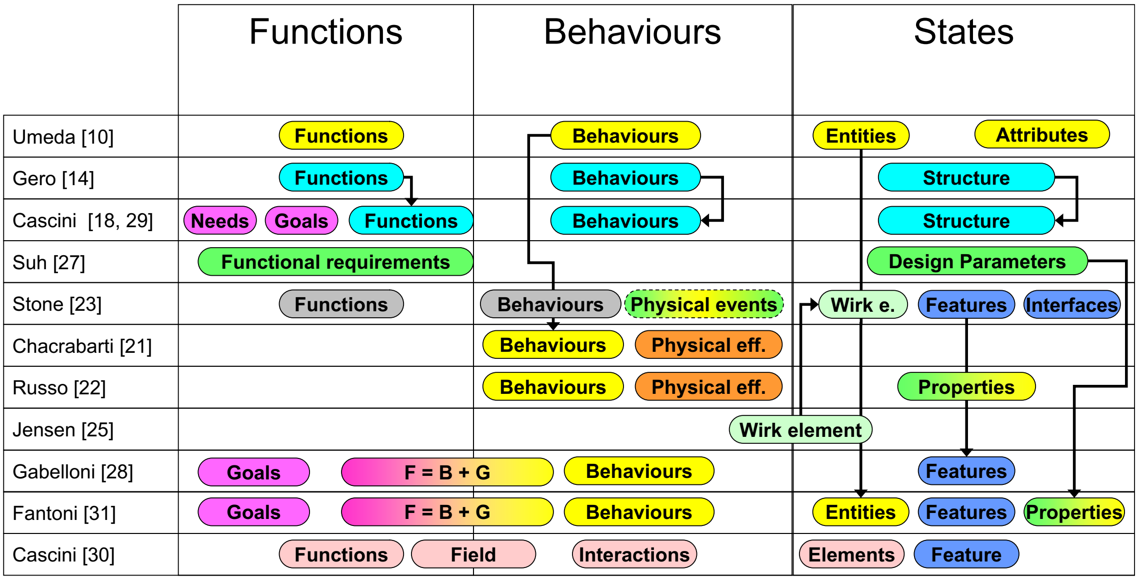
Until a certain point, Computational Linguistic has now reached a very high level of sophistication and reliability and we don't need to discuss it further, since it is now the standard knowledge of the field. As for the degree of precision reached, the recognition of the various parts of speech has for example achieved a very high level of reliability, up to 97% for English [32].

There are however other critical points, related to the semantic aspects rather than to the purely lexical ones, that are still to be fully addressed and therefore need to be properly taken into account also in the analysis of patents.

While we postpone to Section 3 the accurate description of the problems and of the solutions adopted for the specific field of patents, we list here the general issues that apply to the processing of any technical document.

The first problem is about determining exactly the meaning of a certain word or expression (in the technical context this should always be possible).

Sometimes this is just related to the general vagueness of the natural language, which is not always precise as much as desired for technical descriptions. Even if some ambiguities can persist, this cause is usually dealt with and solved when a technical specific jargon is used, as in patents.



	Functions	Behaviours	States
Umeda [10]	Functions	Behaviours	Entities Attributes
Gero [14]	Functions	Behaviours	Structure
Cascini [18, 29]	Needs Goals Functions	Behaviours	Structure
Suh [27]	Functional requirements		Design Parameters
Stone [23]	Functions	Behaviours Physical events	Wirk e. Features Interfaces
Chacrabarti [21]		Behaviours Physical eff.	
Russo [22]		Behaviours Physical eff.	Properties
Jensen [25]			Wirk element
Gabelloni [28]	Goals F = B + G	Behaviours	Features
Fantoni [31]	Goals F = B + G	Behaviours	Entities Features Properties
Cascini [30]	Functions Field	Interactions	Elements Feature

Fig. 1. Synthetic map of FBS approaches in literature.

There is however a more fundamental and unavoidable source of ambiguity. Indeed some words or expressions are homograph, i.e. they are written with the same spelling but have different meanings. Think for example at the couple lead (metal) and lead (guide). The computer is not always able to discriminate them. A particular case of homography is polysemy, when the same word has more than one semantic meaning. A classical example is the English word "bill" that as over 15 meanings as noun and 4 as verb. Technical language is usually more precise than common day language, yet a certain degree of polysemy, especially for certain categories of objects, still exists. In the context of technical documents particular crucial are polysemic verbs describing actions, such as for example *to get*, *to cut*, etc. It is therefore necessary to disambiguate all the dubious words, otherwise the subsequent analysis can be severely compromised.

A symmetrical problem is that of total synonymy, i.e. when the same concept or meaning can be expressed with different, equivalent words or periphrases (i.e. radiofrequency vs. electromagnetic waves between 3KHz and 300 GHz). A particular case of this concept is that of alternative spellings (i.e. radio frequency, radiofrequency, radio-frequency, RF).

Of course in the analysis of technical documents it is quite important to recognise that the different forms in fact refer to the same object, while the computer usually cannot do that if not properly trained. For this purpose proper knowledge databases are a very useful complement to statistical software.

Another source of problems comes from the fuzzy nature of relationships between words in human languages. Hierarchical relationships and affinity relationships are fundamental in any technical analysis, to determine structures, to make comparison and so on. However the same relationships that need to be rigorous and unique between engineering objects, are not so when considering words.

Partial synonymy, expressing similarity between concepts, is often vague, definitively not quantitative, and more crucially does not form a close ring (i.e. if A is synonym of B and B of C, it can happen that A is not synonym of C).

Moreover synonymy also changes with the context (certain words are synonym in some context but not in others, such as *to die* and *to expire*), and mix with the problem of polysemy (since a certain synonymy relationship applies to one of the meaning but not to all).

Hierarchical relationships (hyperonymy and hyponimy) as well can be vaguely-defined in natural language; a particular very interesting case is that of multiple inheritance, where a word can belong to more than one semantic categories or class of generality (for example *to channel* implies, from the functional point of view, both a movement from A to B and a constraint along a pipe). Again which class is more relevant often depends on the context.

Finally sometimes the use of particular jargons, even technical ones, can complicate the analysis if the software has not been trained to recognise the non-standard terms (as for example the hydroxymethyl acetate, which is definitely a not a common English word).

2.3.2. Approaches to the automatic analysis of patents

The approaches to the automatic-semiautomatic analysis of data and metadata contained in patents are numerous and various. Bonino et al. [33] revise the literature and propose an overview of the field focusing both on systems able to improve the precision and recall of a search [22] and on semantic-based solutions.

Here we just sketch the possible methodologies; the interested reader can refer to Moehrle and Gerken [1] and Bonino et al. [33] for comprehensive reviews.

The first distinction that can be made is about the role of patent metadata (inventor, applicant, cited patents and papers, etc.) versus the technical content proper. Indeed some search and analysis engines exploit those non-technical data to build maps of competitors [35] or evaluate the potential of certain technologies [40] on the base of the citation network. While such approach can be very interesting it has to be noted that these data are often very heterogeneous (some patents may just happen to have no metadata), and it can be very hard to obtain homogeneous results.

The second fact to consider is that almost all existing tools try to generalise from the single search originally given as input by the user (and usually consisting in a series of keywords or a short description in natural language). Thus the various software applications can be characterised according to the particular method used for the expansion:

- Synonyms can be expanded to their synset using a linguistic ontology (such as Wordnet [36]).
- Statistics can be used to infer semantic information about words as it is standard in computational linguistics. Sometimes, to reduce the complexity of the analysis Latent Semantic Analysis and Latent Semantic Indexing are used [37,38].
- Another approach is to transform the queries into SAO-structures (Subject–Action–Object-structures), which are close to the FBS model [39,34] and allow exploring the functional relationship of an invention.
- Hand-crafted taxonomies, thesauri and ontologies can be used to help categorise and structure concepts [30,22,41,42]. Ontologies and related knowledge bases can be applied in a variety of different ways, possibly including the ability of incrementally improving and refining the knowledge data, either automatically during the analysis of new patents, either by collecting and storing user feedback.
- Automatically inferred ontologies built for a particular component, device or technology (see the case of RFID in Trappey et al. [43]) and aimed at deeply investigate a targeted field.

The functionalities provided by patent search tool can be similarly grouped into the following main families:

- Research of similar patents for patent documents or single text passages.
- Automatic patent classification, intended as a helper tool for patent offices.
- Clustering of patents, in general, or with respect to a particular aspect.
- High-level description of the patented device, via automated functional analysis, currently attempted via different solutions derived from TRIZ [30,45,47]. This is the aspect closest to the goal of the present paper.

2.4. Automatic analyses of patents for FBS elicitation

For what concerns the automatic extraction of key engineering information from patents it is necessary to cite the old works by Cascini et al. in [30,44] and [45]. There, more than 10 years ago, the authors started extracting structural information such as the architecture of a product (super-system, assemblies, parts) and their relationship in a tree like structure (BOM).

An example of the methodology used for such extraction is the following. The routine starts with the search for the numeric characters in the text; once a number has been found the five preceding or following words are considered candidate for becoming a component [[30] sect. 0030]; moreover, if the first word following the numeric character is "of", the words on the left side are candidate components, while the words on the right side are used for creating the components classification tree. [[30] sect. 0038 n1]

On top of that, under the correct assumption that in a technical description, the subjects and the objects of the sentences are usually the components of the system itself, subjects, verbs and objects are classified as Tools, Fields, Artifacts (according with Altshuller's TRIZ theory). If the verb has not a functional meaning, the corresponding Tools and Artifacts are discarded as candidate components of the system [[30] sect. 0032].

The component classification tree is also reinforced by using descriptive verbs like "comprise", "to be made of", "to be constituted by" or spies as "it consists in three sub-modules, respectively..." where the components preceding the phrases are supersystems while the components following them are parts of said supersystems [[30] sect. 0038 n2].

Cascini also described a method to disambiguate conflicting hierarchies among components and provided an algorithmic way for labelling component as "assembly", "part" or "portion". [[30] sect. 0040].

Methods for collapsing similarities (e.g. for what concern functions) are claimed in the patent but not explicitly explained.

Later, Russo and Montecchi continued the investigation focusing on both automatic patent search [46] and in key information extraction [22]. Even if the first task could seem out of scope in this paper, Russo's approach is oriented to a conceptual design search, therefore he finds for functions and physical effects (behaviours). The search for similar behaviours is based on physical effect extraction and makes use of a commercial database [41], a large repository of physical effects with description in natural and physical terms. By using such base of information Russo and Montecchi can extract all the patents where a given behaviour (e.g. increase temperature) and the related physical phenomena are used and clusterize the most significant patents according to the implemented physical effect [46].

The evolution between [22] and [47] demonstrated a strong interest in FBS extraction; indeed the authors explicitly cited a forthcoming step towards the extraction of structural/state information as well.

The approaches based on co-occurrencies proposed by Curran et al. [48] seem to be more precise but they need for a series of preliminary detailed analysis in the chosen domain before they can produce results. Indeed, an automatic tool for detecting semantic relationships has to be run at least once for each patent class (assuming that an IPC class can be considered an homogeneous domain). Moreover, being based on statistical evidences, in our opinion the results need to be refined later manually by an expert of the field.

3. Material and methods

3.1. Challenges and solutions

The automatic analysis of patents implies two categories of challenges.

The first type is the one common to the processing of all documents written in natural language that we have seen in Section 2.3.1. This first class of issues is addressed in our approach using Computational Linguistic tools; for many tasks it is possible to use standard, open source software, while for some critical steps of the chain we are using tailored variants.

As for the more specific task of information extraction, the last few years have witnessed a growing body of research and practice aimed at developing domain specific ontologies for application in several fields (e.g. legal domain, biomedical domain, etc.). Also in the field of automatic patent analysis a number of ontologies have been successfully proposed in a variety of research projects, mostly focusing on upper level concepts hand-crafted by domain experts (some examples and references have been listed in Section 2). It goes without saying that realistically large knowledge-based applications in the specific domain will need more and more comprehensive ontologies that should be moreover continuously updated.

To avoid this problem, various techniques for automatically acquiring knowledge from text using information extraction methods have been proposed in the Natural LP research community. In this work we usedT2K (TexttoKnowledge) [49], a system to automatically induce ontological knowledge from texts with an ontology learning system. The system offers a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning, which are dynamically integrated to provide an accurate representation of the content of vast repositories of unstructured documents in technical domains. In particular, we used the term extraction module of T2k specifically developed to acquire domain specific terminology relying on the new contrastive approach described in Bonin et al. [55]. The system has been successfully exploited to distinguish between common words and domain-specific terminology in different domains, such as legal domain [56,57], scientific articles concerning environmental domain [58], product catalogues [59].

The system has not been applied to the engineering domain before, and part of the research work has been devolved to testing and adapting the tool to the specific context.

The linguistic part of the procedure is further described in Section 3.2.

The second category of challenges is more linked to the particular field of technical documents, and to patents in particular. The natural language is already quite ambiguous and fuzzy in itself when describing objects of common life, as already mentioned in Section 2.3.1: polysemy is common, synonymy is a not unique and not precise relationship, and so on. Things may get even worse when it comes to technical artefacts and phenomena, since natural language was not "designed" to describe technology.

To quote Galileo Galilei, we cannot understand the book of nature "if we do not first learn the language and grasp the symbols in which it is written. This book is written in the mathematical language, and the symbols are triangles, circles and other geometrical figures, without whose help it is impossible to comprehend a single word of it."

However unfortunately patents are written in English, not in mathematical language; moreover the focus is often on legal aspects rather than on technical ones; actually as already mentioned, the writer usually takes care of removing all references to equations and other physical/mathematical information as much as possible.

Of course we are not saying that there are no such things as specific and precise technical jargons or dictionaries that can complement standard natural language to better describe scientific entities, or that patents are not written using the appropriate terminology. The above-mentioned ambiguity manifests itself at the language level rather than at the lexical level.

For example electrical resistance is a very precise concept, but the related physics can be expressed in many different ways: *to resist*, resistivity but also through the inverse concept of conductivity, resistor, conductance, ohmic resistor, and in particular conditions (i.e. AC) impedance and admittance, and the computer must be able to recognise they are referring to the same physical effect. The software must also be able to understand whether the reference is to the actual behaviour, to a potential one, to a property of a component, and so on.

As another example, consider the function *to absorb;* it is in principle a very precise technical word, but it can refer to at least four totally different physical effects: absorbing a liquid, absorbing a shock or an impact, absorbing a sound or absorbing an electromagnetic wave.

The strategy we have used to address this second category of challenges was to build a vast knowledge base, in which all the possible technical terms and their declinations have been correlated and enriched of additional information. This step is further described in Section 3.2.

A remark is important at this point. Even having available a knowledge database, Information Extraction is in general a complex task, with many challenges and open issues. On the other hand in the present case the extraction of meaningful entities proved to be more manageable because the language used in patents has a very rigid structure, that allows simple rules to be effective in a large majority of cases, starting from part-of-speech tagged text. Manually-encoded rules are used to detect the nominal chunks and detect when two or more chunks are in a logical relation, such are S-(verb)-O, or O-is-(passive-verb)-by-S, or S-in-(relation)-with-O. In many cases it is possible to detect when two different expressions have same meaning, such as in active/passive expressions, and a few collapsing rules where also manually encoded. While a more elaborate information extraction infrastructure could certainly improve the quality of the result, it is outside the purpose of the present paper where a real-life problem has been faced, and it is remarkable how the rigidity of the language (in claims overall) allows a limited number of expression to capture all important connections between the entities appearing in the analysed patents.

The overall procedure we adopted is therefore shown and summarised in Fig. 2; the developed system is composed of three parts: (i) a linguistic chain, (ii) a knowledge base and (iii) a visualisation interface.

Once the text of the document has been processed using the Computational Linguistic tools and transformed in an internal representation, the Knowledge Base is used to disambiguate the lexicon, find all the correlated concepts, and extract the relationship between such concepts. At the end of the process the natural language description has been transformed into a unique, more formal representation of the artefact.

Finally visualisation software uses the information about concepts and their relationships to generate concise diagrams and maps that can be used by the human expert/user for further analyses.

3.2. The process of patent text analysis

Generally speaking, the term extraction process consists of a few fundamental steps. The first three are related to the Linguistic Annotation: (1) sentence splitting and token identification, (2) part-of-speech tagging, and (3) stemming. After these necessary processes the extraction of domain specific terminology starts by (4) identifying term candidates (either single or multi–word terms) from text, and (5) filtering them from non-terms.

With reference to Fig. 2, Term Extraction process already provides partial information about the Significancy Score, through the values of relative frequency, contrastive relevance and associative strength. However the last two phases of the Linguistic Chain are finalised only after the additional knowledge coming from the theory of engineering design and from the Database has been brought in. Therefore we postpone the description of the final part of the chain after the introduction of the Knowledge Base.

Let's see briefly what each step of the first two phases consists in and which software module is used.

3.2.1. Linguistic annotation tools

3.2.1.1. Sentence splitting and Tokenizer. These two software modules split the text into sentences and then segment each sentence in orthographic units called tokens.

3.2.1.2. POS tagging. The Part-Of-Speech tagging (or POS tagging) is the process of assigning unambiguous grammatical categories (or morphological interpretation) to words in context. It plays a key role in natural language processing and in most advanced language technology systems.

Although the high accuracy scores can reach 97% (in English standard newspaper, see [50]), POS tagging remains a central problem because a POS tagging error may affect all the following steps of natural language processing [51,52].

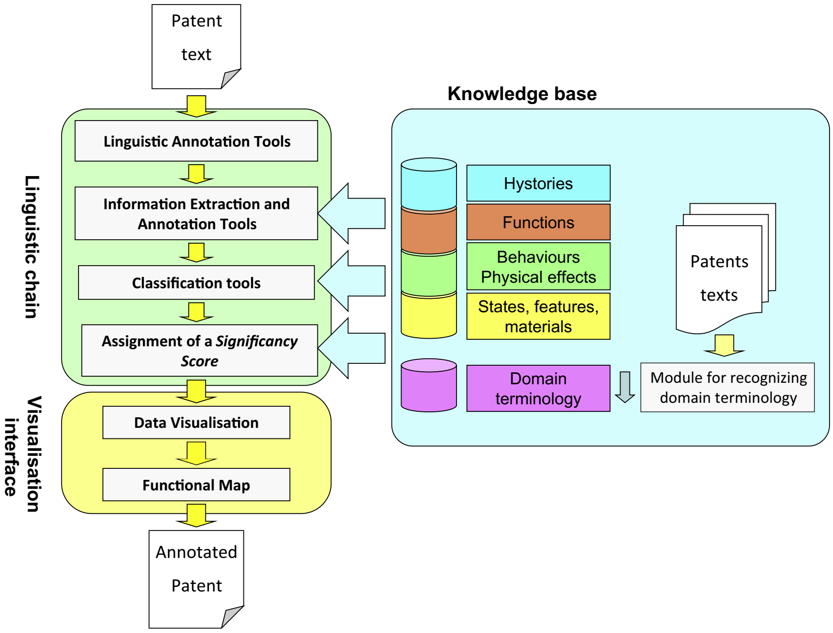

Patent text
Linguistic chain
Linguistic Annotation Tools
Information Extraction and Annotation Tools
Classification tools
Assignment of a *Significancy Score*
Visualisation interface
Data Visualisation
Functional Map
Annotated Patent
Knowledge base
Hystories
Functions
Behaviours Physical effects
States, features, materials
Patents texts
Domain terminology
Module for recognizing domain terminology

Fig. 2. Procedure of analysis, extraction and visualisation of FBS information.

For the present application we make use of the most recent version of the Felice-POS-tagger described in Dell'Orletta [32], which has been tailored for the present application.

3.2.1.3. Stemming. For each word in the text, the POS tagger also detects its root form. This is especially important for inflected (or sometimes derived) words and the following analyses (statistical analysis of single word and multi-word) can usually take advantage of such reduction.

3.2.2. Information extraction and annotation tools

3.2.2.1. Domain-specific terminology extraction. Domain terminology plays a key role in general and in patents in particular. Both single and, even more important, multi–word terms are fundamental to detect crucial concepts [1] in any technical document. Examples of single words are brake, clutch, pump etc. while multi-words are pneumatic brake, electrical clutch, hydraulic pump, etc. The value added by multi-word is higher than single ones since they are more specific and less ambiguous.

The automatic extraction of a corpus of terms characterising the domain has been done by automatically crawling patents belonging to specified patent classes and subclasses. We followed an approach to terminology extraction similar to that proposed in Bonin et al. [55] where, firstly, candidate terms are identified using state-of-the-art statistical measures and, secondly, a shortlist of well-formed and relevant candidate terms is re-ranked by applying a contrastive method.

The term extraction method we followed combines NLP techniques, linguistic and statistical filters. The input text is firstly tokenized, morphologically analysed (i.e. PoS-tagged) and lemmatized passing through a pipeline of state-of-the-art NLP tools for the analysis of English texts. The PoS-tagged text, obtained with the tagger described in Dell'Orletta [32], is searched for on the basis of linguistic filters aimed at identifying a) nouns, expressing **candidate single terms** and b) PoS patterns covering the main nominal modification types which express **candidate complex terms**. It is the case of morpho-syntactic templates such as *adjective + noun* (e.g. piezoelectric actuator), *noun + preposition + noun* (e.g. coefficient of friction), etc..

At this stage, linguistically filtered candidate multi-word terms are screened by using a multi-word preposition stop-list (i.e. a sequence of two or more prepositions, such as 'as well as').

Subsequently, the candidate single terms are ranked on the basis of their frequency of occurrence in the input text, while the candidate complex terms are ranked on the score of a different statistical filter. For this purpose, the C-NC Value measure is used as described in Frantzi et al. [53] and Vintar [54].

Afterwards, the contrastive method is applied against the list of ranked candidate single and multi-word terms. It should be noted that the contrastive function is only applied to a **top list** of these pre-selected multi-word terms, which can be customised through empirically defined thresholds. This procedure allows focusing, firstly, on the retrieval of valid technical terms, thanks to the statistical filters, and secondly on domain pertinence, in two distinct but consequent moments. The top-list of single and multi-word terms are contrasted firstly against the term list extracted from an open-domain corpus and secondly against a top list of terms acquired from a corpus at the level of the different regulated-domain. In both contrastive phases, the contrastive function (**CSmw**) newly introduced in Bonin et al. [55] is used. The **CSmw** score is oriented to prune common words from the list of domain-relevant terms.

As for the maximum number of words of which a complex term can be made, it seems to be domain-dependent (being related to the linguistic peculiarities of the specialised language) [55]. Moehrle presents some results obtained by an analysis performed over different classes where five seems to be the upper limit for having a significant recall (since the flex in the curve frequency vs. multi-word length [for more detail see [1] Fig. 6] is between 4 and 5). The same number of words (five) was used by Cascini [30] sect. 0030], therefore we set the window length to five.

3.2.3. Stop words and stop sentences

A role of particular interest in the analysis is played by what we call standard legal sentences, i.e. those expressions that characterise the way an individual representative usually write patents. Of course they may differ from representative to representative, but are nevertheless recurrent elements in many patents. Even if such sentences play a secondary role in the analysis we preferred to remove them.

Examples from the class B66B9/08 concerning stair lifts and from the same representative are: "An embodiment of this invention will now be described by way of example only and with reference to the accompanying drawings" (132 exact matches in Google Patent Search [60] in 2012-08-11), "Wherever possible, a description of a specific element should be deemed to include any and all equivalents thereof whether in existence now or in the future." (35 exact matches in Google Patent Search in 2012-08-11). Shorter sentences or phases belonging to the legal jargon can be also found extensively in patent literature e.g. "including but not limited to" (5,330,000 exact matches in Google Patent Search in 2012-08-12) or the phrases "Alternate and equivalent embodiments", "Alternate (and for purposes of claim construction, equivalent)", "for purposes of claim construction", "substantially equivalent" and their variations are even more frequent.

At this stage, linguistically filtered candidate multi–word terms are screened by using a contrastive analysis among very different classes (e.g. A43B Characteristic features of footwear parts of footwear, A61B Diagnosis surgery identification E05F Devices for moving wings into open or closed position, E06B Fixed or movable closures for openings in buildings, vehicles, fences, [...], A42B Hats head coverings, F16C Shafts flexible shafts mechanical means for transmitting movement in a flexible sheathing [...], B63H Marine propulsion or steering, B05B Spraying apparatusatomising apparatusnozzles[...], F03D Wind motors [...], B23Q details, components, or accessories for machine tools).

The result of the contrastive analysis is a stoplist of single and multi-word terms (e.g. The present invention, according to claim,...) which appear in all the classes. Such multi-word terms, being commons to so different domain fields, are characterizing the patents with respect to standard newspaper texts, but are too common among patents, and they can be filtered.

The generation of stoplist by contrastive analysis can benefit also from a multilevel approach: first of all the stoplist#1 is generated by contrasting the documents belonging to several different classes with newspaper texts, then a single class is contrasted with the same contrasting set obtaining stoplist#2 and finally the analysis is repeated for each subclass (if the numerosity of documents and contained words is appropriate: words >1,000,000). That seems a correct way (even if computationally demanding) for capturing the three list of stopwords and phrases.

3.3. The knowledge base to extract FBS information

As discussed at the beginning of Section 3, a well-structured and rich Knowledge Base is a key ingredient to disambiguate, gather, select and organise information from technical documents.

Of course the information of interest here is the one related to the FBS design framework, and therefore the construction of the KB is divided into three main branches, that of Functional concepts (which include the user goals), that of physical Behaviours and properties, and that of Structures (including the subcategories of components, materials, geometries etc.).

Each of the three classes of design entities presents its own characteristics and its own particular embedding on the language side. That is, the linguistic counterparts (either single words or periphrases) of functions (as abstract, design concepts) are qualitatively different from the linguistic expressions used to describe physical behaviours, and both are in turn different from the description of structures. For example, active Functions are mainly expressed using verbs, or couples *Verb + Object*, while behaviours can involve *Noun + Adjective* pairs as well; in the same way the prepositions and other particles acting as marker of a relationship will be static, position-oriented for structural relationship and dynamical, time-oriented for physical behaviours.

Thus every part of the KB must be carefully built according to the specific characteristics of the design entity; moreover for the KB to be complete it is not only necessary to have a complete list of all possible functions, behaviours and structures from the design side, but also, and operatively even more important, it is necessary to merge together the sub-databases of all the possible variants that each entity can present in the linguistic counterpart.

The construction of such database is not a straightforward task.

First of all, there is no unique and complete description already of the design entities themselves. Various databases exist in Internet or in the literature, but they have to be merged and enriched.

Moreover, as we have seen in Section 2, the same definition of each FBS entity is multi-faceted when coming to the practical realisation.

For example functions can be expressed by active actions, but can also be implicit in the particular design of the artefact, embedded in the concept of affordance.

A similar distinction between actual and potential applies to behaviours as well, when comparing effects and properties.

Structures are even more articulated, since it is possible to distinguish between subsystems, components, single parts and down to individual features, material properties or geometrical elements.

The second source of complexity in the construction of the KB is that the possible linguistic (syntactical) structures that can be used to represent the design entity in natural language are not known a priori, but have to be determined case by case, often with a bottom up procedure.

Functions are usually represented by verbs, especially when the functioning is described explicitly, in operating condition; however when the focus is on the goal periphrases can appear, such as *to achieve + object*, as well as it happens when describing implicit actions as in the case of affordances where other part of speech can be used.

Behaviours can be expressed with a verb but also by referring to a plurality of physical effects, (for example there is no single verb in the English language to describe what happens during the realisation of a piezoelectric effect), properties and so on.

Structures are mainly substantives, (though structural relationships are more varied), but since they can be considered at different levels: that of BOM, that of materials, that of geometrical elements, etc., for each type of level it is necessary to perform a dedicated machine learning procedure, since every level has its own grammatical structure, and information is to be found, and therefore extracted, from different sources.

To sum up, the main point is that the FBS formalism of engineering design is powerful because abstract and concise, but the natural language is not, and the linguistic expressions that can be used to represent the FBS entities are not in one to one relationship with them.

As already said the human language has not been developed to explain technology, and even technical jargons are far from having the needed precision and uniqueness of formal languages or of mathematics. Sometimes there is no single word for a technical concept (think again at the piezoelectric effect), or the same phenomenon can have different alternative descriptions (consider for example *to heat*, *to get hot*, *to increase temperature* and all other linguistic variants of the same physics), or the differences between related concepts are ambiguous (e.g. the distinction between *to warm* and *to heat*).

It is not a surprise then that the procedure to extract FBS information automatically from corpora of technical documents is not a linear one.

The procedure we have followed is summarised in Fig. 3.

For each of the various design categories we have selected a group of suitable sources, on which to perform the semiautomatic extraction of keywords or word chunks and the subsequent reorganization.

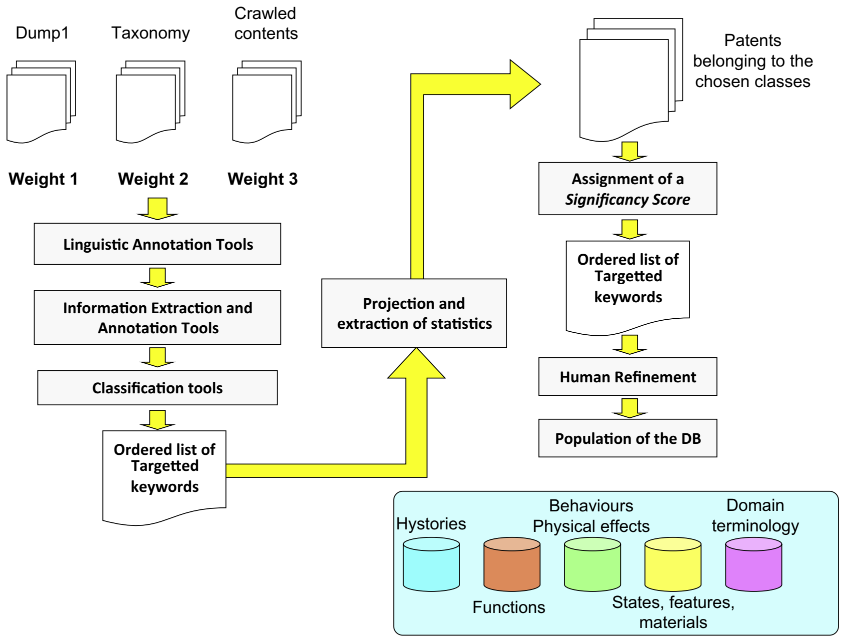

Dump1
Taxonomy
Crawled contents
Weight 1
Weight 2
Weight 3
Linguistic Annotation Tools
Information Extraction and Annotation Tools
Classification tools
Ordered list of Targetted keywords
Projection and extraction of statistics
Patents belonging to the chosen classes
Assignment of a *Significancy Score*
Ordered list of Targetted keywords
Human Refinement
Population of the DB
Hystories
Behaviours Physical effects
Domain terminology
Functions
States, features, materials

Fig. 3. Procedure adopted for the creation of a part of the DB.

Depending on the particular type of entity, on the available sources and corpora, and depending on the quality of such sources and their correspondence with our aims, we had in various cases to widen the database with human intervention and processing.

In the following we review case by case the problems and the results for each FBS category.

Since there are neither free-FBS databases (see for es. Goldfire [41]), nor complete ontologies suitable for our purposes [42] such database has to be constructed almost from scratch. Moreover, the construction procedure cannot be totally automated but needs for human input and supervision. The ideal case is to find (e.g. through a complex search) all and only those elements of interest, but given the statistical nature of certain steps this is rarely the case with the first run.

Of course the need for manual intervention is a hindering factor, as well as having to construct the whole database from the start implies its non-optimality at the first iteration. It is clear that the accuracy in a particular domain can be increased equipping the system with a knowledge base specifically extracted and edited for a specific domain. The creation of ontologies (either automatically or with human supervision) is a well-studied problem, and state of the art techniques can be applied to create a new knowledge base or extend an existing one. The authors decided not to make use of such sources of structured knowledge in the present work. For sure they can provide even better results, but the idea is to avoid the use of proprietary or commercial ontologies and provide a rather simple, although not perfect, tool that any interested designer can replicate in a reasonable amount of time.

There is however another very important reason not to rely on dedicated ontologies, and is the changing nature of the technical domain: new technologies arise almost every year, bringing with them new concepts and keywords. Ontologies are very sophisticated constructions and require time and effort to be adapted to new environmental conditions. Therefore it is more efficient to use as a starting point a database that can be quickly reconstructed every time is needed, and support it with tools that are able to extract relevant and meaningful terminology directly from the patent text, without needing the new term to be already known from ontology. This way it is possible to tackle even the evolution and drifting of a certain domain. An example of this double-pronged approach will be presented in section 4.

3.3.1. Functions

During the last 10 years the research group developed a knowledge database where verbs have been automatically extracted from patents and then organised in 4 classes: functional, structural, logical and non technical [61].

For the **functional**, **structural** and **logical** verbs standard synonym, hyperonym and antonym relationships have been introduced automatically but checked manually. Indeed, in our preliminary tests the automatic generation of lexical relationships from Wordnet 3.0 [36] introduced too much noise (for example the process of synonyms identification could cause the inclusion of too many non technical elements, not suitable for patent analysis). Even machine learning techniques after manual refinement of a 10% of the entire corpus were not sufficient to skip completely the human contribution of final revision.

Such final revision led to a set of about 4,000 meanings of technical verbs. These technical verbs (*to rubber*, *to agglomerate*, *to electrify*, *to decoke,* etc.) are very specific and can be considered more descriptive of behaviours than functions. However they have been related to the functional verbs of the FBS [64] and located in between behaviours and functions [69].

Moreover for what concerns functional information also nouns, adjectives and adverbs having a functional meaning have been automatically detected and linked to the functional verb they belong to: for example motion, movement, movable, movability, etc. have been related to the verb *to move*. Similarly, the database has been extended by inserting also nouns and adjectives related to affordances, again using automatic extraction [62,63].

3.3.2. Behaviour

For what concerns the **behaviours**, the database has been constructed as follows: physical effects have been extracted from different sources but mainly from Wikipedia [66] and, whatever the source (e.g. [67]), organised by using the Wikipedia categories themselves [68]. The union of the results coming from top-down searches in Wikipedia categories and bottom-up searches by using the keywords “effect”; “phenomenon”, etc. allowed us to extract about 5000 pages. We cannot claim that such extraction is either complete or error-free, since the categories are poorly organised, but the total number of effects and the double procedure (top-down and bottom-up) allowed a wide recall of physical effects and behaviours.

Unfortunately, while in standard technical documents behaviours are described in formal ways through equations, physical laws, etc. mathematical analysis is normally not included in a patent [6] since the inventor wants to reduce the level of disclosure and to limit potential developments of the invention by competitors.

Moreover, as explained in Fantoni et al. [69], the boundaries between functions and behaviours, even if theoretically clear, in natural language are often impossible to distinguish (e.g. is the verb *to move* describing a behaviour or a function?). Sometimes the boundaries between property (structure) and behaviour are not clear as well: take for example the case of an elastic beam, where “elastic” is a property of a material but it is strictly related to the equations of stress and strain and to elastic behaviour.

However, since the practical utility of such information, a suitable database has been organised and the multiple inheritances of a term have been solved through multiple attributions to more than one FBS class.

3.3.3. States/structures

Following the definition in Umeda and the recent detailed proposals in Srinivasan [21] and Fantoni et al. [31], **state/structure** information have been detected by the automatic extraction of elements, features and some aggregate properties (as for example in the case of a material that implies different properties depending on the context of the patent, or of a surface quality which depends on the goal of such property, etc.).

Two different problems emerge at the level of state: (i) to distinguish between features (e.g. piezoelectric material) and the corresponding behaviour (e.g. direct piezoelectric effect) is often impossible; (ii) some features, when expressed in natural language, present multiple inheritance or polysemy (e.g. resistance that could be referred both to heat, current, mechanical stress, etc.). The two problems can be partly solved with a disambiguation process based on the context in which the feature appears (in analogy with Ferragina [68]).

For finding elements belonging to states we can adhere to Umeda’s definition that identifies a state through a triplet of entities, attribute and relations. Entity is the label assigned to a real object in the world which has attributes “physical, chemical, mechanical, geometrical, or other properties” and relations with other “entities, attributes and relations” [13].

Therefore we must (i) detect the label of each entity in a patent, (ii) understand if some domain concepts have been used, and then extract: (iii) the material used for the entity, (iv) the material properties and object characteristics (physical, chemical, manufacturing, etc.), (v) its attributes (light, heavy, bulky, thin, thick, etc.), (vi) its physical states (solid, liquid, gas, plasma, gel, mixture, col-

Table 1
Table of the functions.

Class	Examples	Source	Automatic; semiautomatic manual
Functional verbs	Branch, distribute, diffuse, dispel, disperse, dissipate, diverge, scatter, import, allow, input, capture, export, dispose, eject, emit, empty, remove, destroy, eliminate, etc.	Functions from FBS [64,65]	M
		Dictionary and taxonomy of about 4000 meanings [61]	S
Functional nouns and adjectives	Position, contact, alignment, abutment, deletion; partition, ignition; motion; movement, resolution; derivation; adhesiveness, junction, connection, adhesion, leakage, removal	Wikipedia [66]	A
Affordances: nouns and adjectives	Stability, motility, reachable, storability, graspability, prick-able, deformability, climbability, etc.	[62,63]	A
Number of entries in the database	3513		

loid, etc.) and (vii) its material state (martensite, austenite, triple point, etc.), (viii) its geometry and (ix) the basic features it is composed of (see Tables 1, 2 and 6).

No source of structured information was available; therefore, since the heterogeneity of the data to be gathered and classified, several sources have been chosen and different strategies have been adopted for each source and for each task. For each type of structural information, Table 5 shows some example, the sources used to gather the data and how they have been collected.

Moreover, in general, the wider was the source, the lower was the precision and the higher the impact of the manual/human filtering. Conversely, the higher was the precision of the source, the lower was usually its recall (i.e. extension of the collection) and then higher the impact of manual extension.

For the extraction of the materials we decided to use Freebase [70], since of the high quality of its definitions. Unfortunately the scientific session in Freebase is not as extended as in Wikipedia and additional resources (engineeringdraft.freebase.com) are not yet good and homogeneous. Therefore its use, even if very interesting, has been very limited. State information have been splitted into Tables 3–5 which present materials, properties and attributes, and structures, geometrical features and interactions, respectively.

3.3.4. Scenes

Some additional information can be extracted from patents to detect the **scenes** or **histories**[16]. The aim is to identify the logical and causal chains and temporal sequence of behaviours happening in a product for making it to function in a proper way. The extraction of time, causal and logical elements from the patent text can help in organizing the functional representation in a more intelligible manner.

Examples of relevant indicators are the logical verbs, such as *to cause* and *to provoke*, the time related verbs such as *to wait* and *to delay*, (time) adverbs such as *then*, *later*, *after*, but also (logical) adverbs as *hence*, *therefore*, adjectives as *instantaneous* (and is adverb *instantaneously*) etc. Such elements have been automatically extracted and manually checked. In the analysis they have been used to create causal chains of related functions/behaviours.

Even if the markers able to reveal the presence of possible transitions between different histories have been detected, there is not a biunivocal correspondence between the presence of a marker and the corresponding transition. The use of such markers to automatically build functional flow diagrams (as those in Wie [23]) is far to be fully implemented and will be object of future works.

4. Results

The developed system has been used in several industrial cases in the automotive, biomedical, industrial automation and robotics sectors, generally completing the analysis in less then 10 s, and never requiring more than 1 min. Since the system and the databases have been built starting from classes belonging to the automotive and biomedical field, here we propose the analysis of a very new mechanical device for handling objects [79], to illustrate the results that can be obtained through an example, and to provide at the same time evidence for the flexibility of the system. The mentioned device [74] qualifies as a good test case because it is really recent, it does not belong to the training classes and is of interest for the existing EU project Roblog (about robotics in manufacturing and logistics – FP7 ICT-270350).

We note that although the case is sufficiently distant from the training classes to constitute a good test of flexibility, it is also sufficiently homogeneous to be reliable: the class of the chosen device shares with the training ones a lot of electric, mechanical, physical and structural information with them.

Table 2
Table of the Behaviours.

Targeted extraction	Examples
Behaviours	Magnetic field, magnetisation, intensity of magnetisation, magnetic polarisation, diffraction, interference, polarisation, astigmatism, refraction, reflection, transmission coefficient, defocus, coma, astigmatism, field curvature, image distortion, aberration, magnetic flux density, induction coercive force, coercivity, saturation, polarisation, hysteresis, permeability, conduction, resistance, capacitance, impedance, displacement, tension, stress, ideal conductor, etc.
Physical effects	Eddy Current, Joule's effect, Moiré effect, Faraday effect, Doppler effect, Coriolis force, Hall effect, Skin effect, Venturi effect, stick–slip phenomenon, triboelectrification, Speckle pattern, Kaye effect, shear-thinning behaviour, Knudsen layer, Rankine vortex, turbulent jet breakup, rheopecty
Models: equations, laws, principles, etc.	Searches like: 's law OR law (es. Ampere's law, Biot-Savart law) 's principle OR principle (es. Bernoulli's principle, Hamilton's principle, etc.); 's theorem and theorem (es. Gauss–Bonnet theorem, Pythagorean theorem, etc.); 's equation OR equation (es. Navier–Stokes equations, Maxwell's equations, etc.); coefficient (es. Activity coefficient, temperature coefficient); constant (es. Gravitational constant); number (es. Reynolds number, Chandrasekhar number, etc.)
Number of entries in the database	5089

Table 3
Materials.

Class	Type and number of entries in the DB	Examples	Source	Automatic; semiautomatic; manual
Materials	Material names, element symbols, chemical element **1741 entries**	All the elements and all the element symbols in **Mendeley periodic table**, but also: steel; wood; iron; polyurethane; aluminium oxide; polycarbonate; invar; macor; stainless steel alloy; monel400; granite; sandstone; cast iron; stucco; metal; bauxite; coal; lignite; fluorite; slate; pyrite; barite; ironore; wolframite; flint; cassiterite; lime; xsilver, etc.	[70,71]	A

Table 4
Properties and attributes.

Class	Type	Examples	Source	Automatic semiautomatic manual
Physical characteristics	Chemical	Acidity; activity; basic; diffusion; effervescence; hydrophobicity; reactivity; solubility; valency	[72]	A
	Mechanical	Brinell hardness; elongation at fracture; fatigue limit; fatigue strength; hardness; load frequency; Poisson’s ratio; resilience; shear yield stress; tensile elasticity modulus; Vickers hardness; Young’s modulus	[66] [73] [71] [72] [76]	A
	Magnetic	Diamagnetic; ferromagnetic; magnetic; paramagnetic; paramagnetic; permeability; saturation		A
	Optical	Absorptivity; colour; dispersion; fluorescence; index of refraction; luminescence; luminosity; luster; photosensitivity; reflectivity; refractive index; scattering; streak; transmittance		A
	...	...		...
	Manufacturing	Castability; extruding temperature and pressure; hardness; machinability rating; machining feed; machining speed; material removal rate;		A
Unit of measurement	In form of a triple: Name–Symbol–Property	ampere-A-electric_current; candela-cd-luminous_intensity; kelvin-K-thermodynamic_temperature; kilogram-kg-mass; metre-m-length; mole-mol-amount_of_substance; second-s-time	[66]	A
Attributes	Other frequently occurring adjectives extracted from patents	Able; abrasiveness; bulk; capable; capacity; connectivity; eccentric; elastic; fast; plastic; porosity; quick; rapid; resistivity; sensitivity; short; strength; thickness; viscosity		S
Flows attribute	Material flows attributes (energy and signal are shared with behaviour)	Aerosol; aggregate; alloy; body; colloid; composite; elastic-body; foam; gas; gel; glass; liquid; liquid crystal; misture; mixture; object; particulate; plasma; rigid-body; sol–gel;	[64] [65] [66]	A
	States of matter	Aluminides; austenite; eutectic; ferrite; ferritic; heteroazeotrope; liquidus; martensite; martensitic; peritectic; solidus; supercritical fluid; superfluid; superglass; supersolid; symplectite; zeotropic mixture	[66]	A
Number of entries in the database	11490			

The analysed patent concerns a novel Bernoulli gripper for automatic handling of flat thin objects like silicon wafer or solar cells. It exploits the well-known Bernoulli’s principle for generating a negative pressure at the gripping face, thus lifting two-dimensional components in a contactless way (for more detail see [80]). The Bernoulli gripper is also provided with a damping device (located circumferentially) that reduces the shocks between the gripper and the wafers.

4.1. Preliminary processing

The procedure for the analysis of the patent proceeded as follows: the linguistic chain has been applied to the entire document text, the words from the stop list were hidden and the key concepts detected using the various software tools described in Section 3.

First, the following technical compound expressions had been automatically recognised and extracted since they were already in the database: rubber ring (#10); negative pressure (#6); vacuum pump (#2); atmospheric pressure (#2); solar cell (#1). Afterwards, the automatic linguistic multiword extractor detected the meaningful chunks in the patent, according to predefined POS patterns such as *adjective + noun, noun + noun, adjective + verb + noun* and so on (using the Penn TreeBank notation, detailed in the glossary, the above become JJ + NN; NN + NN, JJ + VB*+NN, etc.). The most common chunks are shown in Table 7.

4.2. Extraction and visualisation of structural information

For what concerns the extraction of information on the Structure, the only component missed by the automatic chunks extractor is “component 15 attached bristle ends”, while other really interesting parts or features as damping device, bearing surface and bearing ring (belonging to structure) have been properly detected.

After the extraction, the numbered components (in bold in Table 7) were recognised through the use of simple regular expressions (e.g. chunk + ”said” + CD, chunk + “said component” + CD, chunk + CD, NN + ”said” + CD, NN + “said component” + CD, NN + CD, etc., where CD = Cardinal Number [78]). Using such subsequent analysis also the missing component “bristle ends 15” has been detected.

The other remaining chunks, i.e. those not indicating proper components, usually still indicate fundamental concepts in the patent, therefore they are immediately linked (if possible) to the components and coloured according to their position in the database (functional, behavioural or structural relationships).

Table 5
Table of conceptual structures, geometrical features, interactions, etc.

Class	Type	Examples	Source	Automatic semiautomatic manual
Structure concepts	Standard Structures, devices and general concepts	Shaft, rod, crank, disk, winch, handle, cam, sun gear, pinion, ring, cable, tube, pipe, shell, wheel, chain, helical gear, belt, worm gear, screw, bolt, bearing, magnetic bearings, capacitor, coil, resistor, switch, etc.	[66]	A
Feature	Name	Pocket, hole, socket, groove, step, edge, pitch, tooth, etc.	[74,75]	A
Geometrical Features	2D elements	Points, lines, arcs, circles and axes, vertex, side, etc.		S
	Geometrical properties	Tangency, endpoint, perpendicular, middle, parallel, offset, centre, thickness, corner, coaxial, concentric, eccentric, cross, correspondence, etc.		S
	3D elements	Pad, pocket, socket, hole, hollow, shaft, groove, rib, slot, edge, plane, surface, helical, curve, straight line, face, shell etc.		S
	Solids	Cube, parallelepiped, cone, *-hedron, etc.		S
	Operations to create new geometries	Chamfers, fillets, or to trim and to mirror, to create a pattern (rectangular pattern or circular pattern), to scale up/down, to sweep, to split a solid, etc.		S
	Geometrical operations	Orientation, alignment, mate, translation, rotation, roto-translation, etc.		S
	Constraints			S
	(a) Dimensional,	(a) Length, angle, radius, diameter, etc.		
	(b) Geometrical	(b) Coincidence, concentric, horizontal, vertical, symmetric, etc.		
Interactions	Often they belong also to the class of functions and structure concepts	Junction, conjunction, attraction and repulsion, contact, electromagnetic coupling, magnetic interference, assembly, relationship, interface, interconnected, locations, tolerance, gap, boundary, proximity, connectivity, etc.	[77]	M
Number of entries in the database	2047			

Table 6
Table of the markers for the detection of information concerning scenes.

Targeted extraction	Type	Examples
Phases	Logical or time verbs	To cause, to provoke, to generate, to start, to wait, to delay,
	Time adverbs	Then, later, after, hence, therefore, continuously, when, once, after, while, meanwhile, as a consequence, whilst, etc.
	Sentence structure	**If** sentence, **thus** sentence, **As** sentence, **thus** sentence, **If** sentence, **then** sentence, sentence **to enable** sentence

Table 7
Table of the most common chunks (JJ + NN; NN + NN, etc.) in the patent US8172288.

Chunk	#	Chunk	#
Gripping face	**40**	Lateral view	4
Baffle plate	**27**	Shock-free manner	4
Damping device	23	Silicon-based wafer	4
Bearing ring	18	**Circumferential brush edge**	**3**
Rubberized bearing surface	**17**	**Circumferential clamping groove**	**3**
Clamping ring	**14**	**Clamping groove**	**3**
Tunnel-shaped component	**14**	Suction-induced bearing	3
Bernoulli gripper	**13**	Open ring	3
Two-dimensional component	**12**	**Bristle-free annular section**	**3**
Flow system	**12**	Holding two-dimensional component	3
Bearing surface	10	Air consumption	2
Plan view	9	Slip-resistant manner	2
Suction-induced approach	9	Complete sealing circle	2
Damping resistance	7	**Longitudinal axis**	**2**
Capacitive sensor	**5**	Still achieving reliable holding	2
Excess pressure	5	Same time	2
Other component	5	Air connection	2
Controllable robot arm	**5**	Suction-induced placement	2
Slip-resistant movement	4	Suitable elastic material	2
Second air connection	4	Sealing circe	2

Then, verbs and adverbs, prepositions and structural phrases are used to connect components, while adjectives and nouns are used to assign properties to the components.

We note that the problem of resolving pronouns (=coreferences) such as ‘which’ or ‘that’ is very complex and far from being solved properly for this application. Actually a series of wrong or missing connections among components are due to the errors done by the Dependency Parser [82]. The algorithms used in the applications solve only a limited set of cases and they are not reliable enough to deal with different writing styles (or complicated references).

The detailed description of the patent has been analysed and the result visualised (through Graphviz [81]) in the graph of Fig. 4, where the FBS relationships between components are shown.

It has to be noted that all relationships shown in Fig. 4 have been determined by the software, although in three subsequent steps (chunk extraction, components allocation, properties and relationships assignation). The only human intervention has been in mediating the transition from a step's output to become input for the next stage. The full automatisation of the procedure may be desirable but on the other hand the supervision of the intermediate outputs by the user guarantees a better control and understanding of the final results.

4.3. Extraction of functional–behavioural information

Visualising the network of relationships between the various components of a device is surely important, especially if such information can be integrated with the knowledge coming from the drawings of the patent.

However, the extraction of functional–behavioural information and of state properties is even more interesting. This is done by attempting to cluster relevant chunks and single words around a common FBS concept. Table 8 shows the result of such clustering. The first four clusters correctly belong to functional categories present in the knowledge DB; on the other hand the following two entries, **view** and **brush**, are labelled as functional information while in the present context they are not. In fact, view is related to drawing specific jargon, and the meaning of 'brush' in the patent is referred to a device, rather than being related to the behaviour "to brush" and to the functional verb (*to separate*).

The ambiguity of the two words can be resolved using the additional Part-of-speech information; indeed a human reader would recognise the noun-like usage of both view and brush.

The preliminary co-occurency analysis is very simple and already gives much useful information but the above discussed ambiguity suggests the need to adopt a more sophisticated approach.

In order to better understand the syntax of sentences a dependency parser [82] has therefore been used. Dependency parsing produces a labelled tree-like representation of the syntax of a sentence, attaching each word to the word it is referring to (unless it is

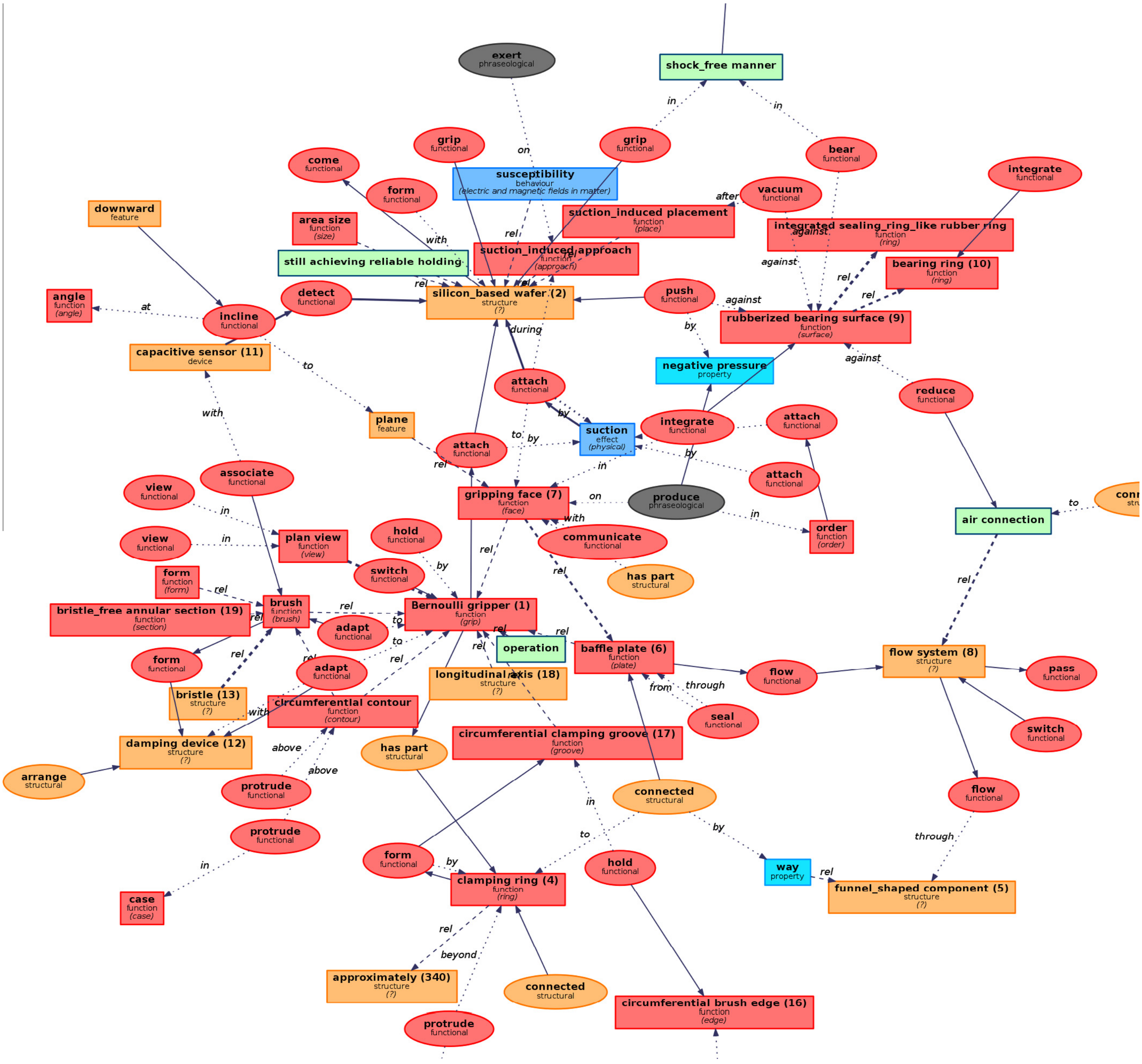
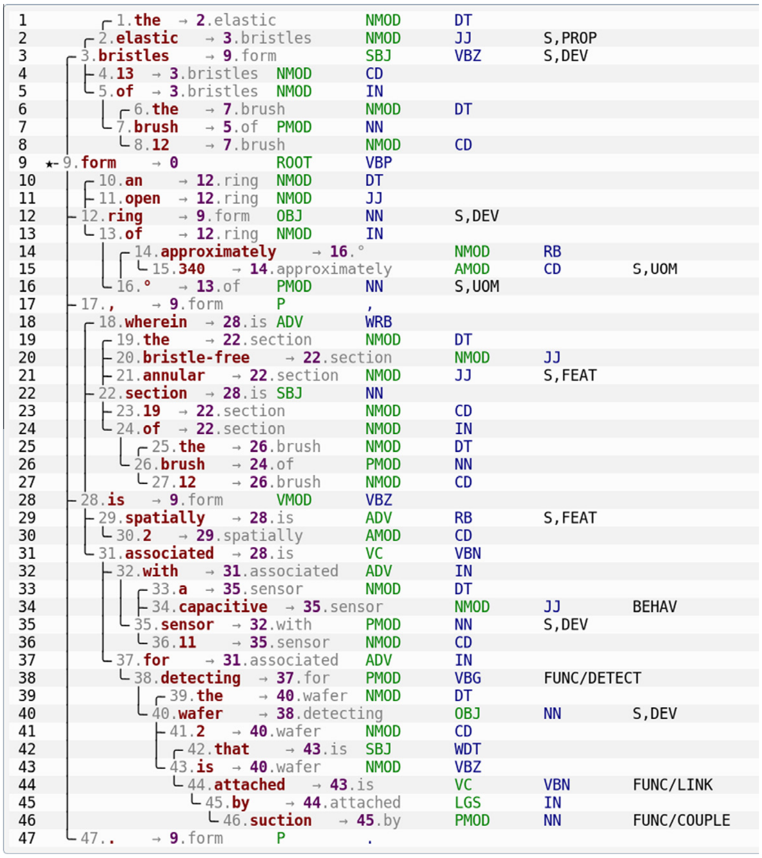

exert phraseological
shock_free manner
grip functional
grip functional
bear functional
come functional
form functional
susceptibility behaviour (electric and magnetic fields in matter)
vacuum functional
integrate functional
downward feature
area size function (size)
suction_induced placement function (place)
integrated sealing_ring_like rubber ring function (ring)
still achieving reliable holding
suction_induced approach function (approach)
bearing ring (10) function (ring)
angle function (angle)
detect functional
silicon_based wafer (2) structure (?)
push functional
rubberized bearing surface (9) function (surface)
incline functional
capacitive sensor (11) device
negative pressure property
reduce functional
attach functional
attach functional
suction effect (physical)
integrate functional
plane feature
attach functional
associate functional
gripping face (7) function (face)
produce phraseological
attach functional
air connection
view functional
order function (order)
view functional
plan view function (view)
hold functional
communicate functional
form function (form)
switch functional
has part structural
bristle_free annular section (19) function (section)
brush function (brush)
Bernoulli gripper (1) function (grip)
adapt functional
operation
baffle plate (6) function (plate)
flow system (8) structure (?)
form functional
flow functional
pass functional
adapt functional
longitudinal axis (18) structure (?)
bristle (13) structure (?)
circumferential contour function (contour)
seal functional
switch functional
damping device (12) structure (?)
has part structural
circumferential clamping groove (17) function (groove)
arrange structural
protrude functional
connected structural
flow functional
protrude functional
form functional
hold functional
way property
case function (case)
clamping ring (4) function (ring)
funnel_shaped component (5) structure (?)
approximately (340) structure (?)
connected structural
circumferential brush edge (16) function (edge)
protrude functional
in
on
after
with
rel
against
at
during
by
to
from
through
above
beyond

Fig. 4. Graphical representation of the patent as obtained through Graphviz.

Table 8
Key concepts as clusterized by the system according to FBS characteristic.

Lemma # DB source	Type	Chunk or verb	#
Grip #110 **function**	Chunk	Bernoulli gripper	25
	Chunk	Gripping face	42
	Chunk	Gripper	24
	Verb	Grip	19
Bear #57 **function**	Chunk	Rubberized bearing surface	19
	Chunk	Bearing surface	11
	Chunk	Bearing ring	19
	Chunk	Suction-induced bearing	3
	Verb	Bear	5
Suction #35 **function**	Chunk	Suction_induced approach	9
	Chunk	By suction	20
	Chunk	Suction_induced placement	2
	Chunk	Suction-induced bearing	3
	Chunk	Suction_induced non_decelerated impact	1
Damp #33 **Function**	Chunk	Damping device	25
	Verb	Damping resistance	7
	Verb	Damp	1
View #26 **Function**	Chunk	Plan view	9
	Chunk	Lateral view	3
	Chunk	Perspective bottom view	3
	Chunk	Lateral section view	2
	Verb	View	9
Attach #26 **Function**	Verb	Attach	24
	Chunk	Extremely shock_resistant attachment	2
Brush #24 **Function**	Chunk	Circumferential brush edge	23
	Chunk	Brush	1
Pressure #10 **behaviour**	Chunk	Negative pressure	5
	Chunk	External atmospheric pressure	4
	Chunk	Pressure	1
... #...	...	...	...



```
      ┌1.the   → 2.elastic         NMOD   DT
    ┌2.elastic   → 3.bristles      NMOD   JJ        S,PROP
  ┌3.bristles    → 9.form          SBJ    VBZ       S,DEV
  │ ├4.13   → 3.bristles  NMOD   CD
  │ └5.of   → 3.bristles  NMOD   IN
  │   │┌6.the    → 7.brush         NMOD   DT
  │   └7.brush   → 5.of   PMOD     NN
  │     └8.12    → 7.brush         NMOD   CD
★-9.form    → 0           ROOT     VBP
  │ ┌10.an      → 12.ring  NMOD    DT
  │ ├11.open    → 12.ring  NMOD    JJ
  ├12.ring      → 9.form   OBJ     NN       S,DEV
  │ └13.of      → 12.ring  NMOD    IN
  │   │┌14.approximately    → 16.°           NMOD     RB
  │   ││ └15.340    → 14.approximately       AMOD     CD        S,UOM
  │   └16.°    → 13.of     PMOD    NN       S,UOM
  ├17.,       → 9.form     P       ,
  │ ┌18.wherein  → 28.is ADV       WRB
  │ │ ┌19.the      → 22.section    NMOD     DT
  │ │ ├20.bristle-free      → 22.section    NMOD      JJ
  │ │ ├21.annular    → 22.section  NMOD     JJ        S,FEAT
  │ ├22.section  → 28.is SBJ       NN
  │ │ ├23.19   → 22.section        NMOD     CD
  │ │ └24.of   → 22.section        NMOD     IN
  │ │   │┌25.the     → 26.brush    NMOD     DT
  │ │   └26.brush    → 24.of       PMOD     NN
  │ │     └27.12     → 26.brush    NMOD     CD
  ├28.is     → 9.form      VMOD    VBZ
  │ ├29.spatially    → 28.is       ADV      RB        S,FEAT
  │ │ └30.2    → 29.spatially      AMOD     CD
  │ └31.associated   → 28.is       VC       VBN
  │   ├32.with     → 31.associated ADV      IN
  │   │ │ ┌33.a    → 35.sensor     NMOD     DT
  │   │ │ ├34.capacitive   → 35.sensor      NMOD      JJ        BEHAV
  │   │ └35.sensor   → 32.with     PMOD     NN        S,DEV
  │   │   └36.11     → 35.sensor   NMOD     CD
  │   └37.for      → 31.associated ADV      IN
  │     └38.detecting   → 37.for   PMOD     VBG       FUNC/DETECT
  │       │ ┌39.the      → 40.wafer  NMOD   DT
  │       └40.wafer      → 38.detecting     OBJ       NN        S,DEV
  │         ├41.2    → 40.wafer    NMOD     CD
  │         │ ┌42.that     → 43.is  SBJ     WDT
  │         └43.is   → 40.wafer    NMOD     VBZ
  │           └44.attached    → 43.is       VC        VBN       FUNC/LINK
  │             └45.by      → 44.attached   LGS       IN
  │               └46.suction    → 45.by    PMOD      NN        FUNC/COUPLE
  └47..      → 9.form      P       .
```

Fig. 5. Tree of dependencies with DRL, POS and FBS tags.

the root), and labelling it with a tag, called Dependency Relation Label - DLR, clarifying the syntactic role (subject, object, etc.). While yielding less structure than a full-blown (constituency) parsing of syntactic constituents, dependency parsing can be achieved relatively quickly and with a considerably smaller error-rate (about 7%) in modern Shift-Reduce dependency parsers. In the present case, the results of the dependency parsing have been further enhanced by adding FBS information to the classical output.

The result of such enhanced parsing for the sentence "*The gripper according to claim 1, characterised in that* the elastic bristles (**13**) of the brush (**12**), *in plan view of the gripper (**1**),* form an open ring of approximately 340°, wherein the bristle-free annular section (**19**) of the brush (**12**) is spatially (**2**) associated with a capacitive sensor (**11**) for detecting the wafer (**2**) that is attached by suction" (where phrases in italic have been removed) can be seen in Fig. 5.

Of particular interest is the last column where FBS related words are properly tagged. For example "elastic" is a property belonging to the States/Structure set (indicated by the label S,PROP), while "attach" belongs to the function *to link*, (label FUNC/LINK), or "suction" is related to the functional verb *to couple*. The unit of measurement 340° has also been properly detected and labelled (as S,UOM).

Once extracted, Functional information can be re-projected on the structure, in order to complete the graph of relationship between the components. In the theory of engineering design, the subset of the graph limited to functional relationships is called a Functional Analysis Diagram, a very useful graphical representation of the way an artefact really works that can be generated in a semi automatic way using the procedure outlined.

In the same way, Functional information can be used alone, according to the theory of Functional Analysis, to perform various analyses on the single artefact, within a whole class of products or even between different typologies of products and technologies.

5. Discussion and conclusions

The overall system, composed of the engine and the database, seems able to provide reliable FBS information to engineers. Moreover, it demonstrated a quite good flexibility of use in different contexts. The possibility of changing the granularity of the map and of providing different views (structural or functional or behavioural views) allows users to analyse a patent from different perspectives and at different levels.

Several examples (additionally to the present example) have been used as for validation, comparing the output of the system against the human analysis to extract the FBS information. However to construct a detailed FBS tree is tedious and require knowledge of the model, and it's easer listing the connections between different components, and compare such connections with the set of connections automatically extracted. The non-expert reader can still have a proof of correctness.

Considering how quicker a human user can get a glimpse of a patent from a graph with the labelled connections of the most important elements, it is also possible to verify empirically the accuracy of the extracted relations: indeed, the verification work adds a small burden to the everyday's task of people working with patents, that can collect judgments' on outputs from the system applied to the patents they are already working with. With a sufficiently wide spectrum of test cases, confidence on the correctness of the extraction can be progressively gained, as well as evaluation of the versatility of the approach when applying the system to new domains.

Of course mistakes still occur during the analysis; however they have only a few possible causes that can be tackled in future. Such causes can be grouped in the following families.

5.1. Pronouns

Pronouns (that, which, etc.) are in general very difficult to resolve properly; moreover the subordinate sentences that pronouns originate may not be correctly linked to their real subject.

5.2. Database structure

Many words have multiple attributions to different classes in the DB. At the present moment the only way for discriminating them is to completely disambiguate each sentence, applying statistical learning to infer the most probable meaning and attribution of each word.

5.3. Missing words

Many elements are still missing in all the three FBS classes of the DB.

5.4. POS tagging

While being generally computed correctly, the errors about the Part of Speech can have a devastating effect when they are propagated in the linguistic chain, in a sort of domino effect. Such problems affect not only the following steps of the linguistic chain but sometimes interfere also with the correct attribution to the FBS class (see the case of the noun "brush" described above).

Since the run time of our system is quite fast for the context of most analyses that involve a small number of documents, it could be advisable trading some speed for better accuracy.

To improve the accuracy and flexibility of the software application various actions can be hypothesised and will be implemented in the future release of the system.

First of all, the database can be automatically expanded by using machine learning techniques. In particular, typical domain words can be extracted in each class and the database populated by using such conceptual words/chunks. Furthermore, all the unknown relevant words processed by the system have to be continuously detected and added to the database in a supervised manner.

Conflicts among words having a multiple attribution (e.g. *to weld* belongs to Functions and Behaviours while *to position* belongs to Functions and Structures) can be solved either (i) at patent level via a disambiguation procedure or (ii) at database level through the development of a more suitable ontology.

Finally, the use of the system by the community of technicians and engineers will allow detecting wrong words, or incorrect attributions to the various FBS categories and their subcategories. This activity will help in the refinement of the tool and of the knowledge base.

Acknowledgements

The financial supports of RobLog Project (FP7 ICT-270350) and LILIT Project (PAR FAS REGIONE TOSCANA Linea di Azione 1.1.a.3), AMDS project (A1.6 POR "CreO" FESR 2007-13 and A1.1 del PRSE), Magna Closures and Erre Quadro s.r.l. are kindly acknowledged. Moreover the authors are in debt to Miss. Del Tevere A., Mr. Lasi M., Mr. Tazzini G., Mr. Campolmi R. and Dr. Manenti M. who manually tagged and revised the data automatically gathered by the software application. The quality of the databases is highly due to their efforts.

References

[1] M.G. Moehrle, J.M. Gerken, Measuring textual patent similarity on the basis of combined concepts: design decisions and their consequences, Scientometrics 91 (2012) 805–826.

[2] I. Bergmann, D. Butzke, L. Walter, J.P. Fuerste, M.G. Moehrle, V.A. Erdmann, Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips, R&D Management 38 (5) (2008) 550–562.

[3] P.F. Burke, M. Reitzig, Measuring patent assessment quality—analyzing the degree and kind of (in)consistency in patent offices' decision making, Research Policy 36 (9) (2007) 1404–1430.

[4] M. Philipp, Patent filing and searching: Is deflation in quality the inevitable consequence of hyperinflation in quantity?, World Patent Information 28 (2) (2006) 117–121

[5] L. Yanhong, T.T. Runhua, A text-mining-bases patent analysis in product innovative process, in: N. Léon-Rvira (Ed.), Trends in Computer Aided Innovation, Springer, New York, 2007, pp. 89–96.

[6] D. Golzio, WHOW (Why, When, Who, Where, What, How) Read a Patent!.

[7] R. Apreda, A. Bonaccorsi, G. Fantoni, Functional Technology Foresight. A New Methodology for the Era of Societal Challenges, Technological Forecasting and Societal Change, submitted for publication.

[8] R. Apreda, A. Bonaccorsi, G. Fantoni, D. Gabelloni, Functions and failures. How to manage technological promises for societal challenges, Technology Analysis and Strategic Management, submitted for publication.

[9] M.S. Erden, H. Komoto, T.J. van Beek, V. D'Amelio, V. Echavarria, T. Tomiyama, A review of function modeling: approaches and applications, Artificial Intelligence for Engineering Design, Analysis and Manufacturing 22 (2008) 147–169.

[10] Y. Umeda, H. Takeda, T. Tomiyama, H. Yoshikawa, Function, behaviour, and structure, in: J.S. Gero (Ed.), Applications of Artificial Intelligence in Engineering V, Proceedings of the Fifth International Conference, Boston, USA, July 1990, vol. 1, Computational Mechanics Publications Southampton Boston Co-published with Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, 1990, pp. 177–194.

[11] T. Tomiyama, Y. Umeda, H. Yoshikawa, A CAD for functional design, CIRP Annals – Manufacturing Technology 42 (1) (1993) 143–146.

[12] Y. Umeda, M. Ishii, M. Yoshioka, Y. Shimomura, T. Tomiyama, Supporting conceptual design based on the function–behavior–state modeler, Artificial Intelligence for Engineering Design, Analysis and Manufacturing 10 (4) (1996) 275–288.

[13] Y. Umeda, T. Tomiyama, H. Yoshikawa, FBS modeling: modeling scheme of function for conceptual design, in: Proceedings of the 9th International Workshop on Qualitative Reasoning. Amsterdam, Netherlands, 11–19 May 1995, pp. 271–278.

[14] J.S. Gero, Design prototypes: a knowledge representation schema for design, AI Magazine 11 (4) (1990) 26–36.

[15] J.S. Gero, U. Kannengiesser, The Situated Function–Behaviour–Structure Framework, Artificial Intelligence in Design, Kluwer, Dordrecht, 2002. pp. 89–104.

[16] P.J. Hayes, The naive physics manifesto, in: D. Michie (Ed.), Expert Systems in the Micro-Electronic Age, Edinburgh University Press, Edinburgh, 1979, pp. 242–270.

[17] K.D. Forbus, Qualitative process theory, Artificial Intelligence 24 (3) (1984) 85–168.

[18] G. Cascini, G. Fantoni, F. Montagna, 2012. Situating needs and requirements in the FBS framework, Design Studies,http://dx.doi.org/10.1016/j.destud.2012.12.001 .

[19] R.T. Somasekhara, A.Chakrabarti, Analysing modifications in the synthesis of multiple state mechanical devices using configuration space and topology graphs, in: International Conference on Engineering Design, ICED11, 15–18 August 2011, Technical University of Denmark, Copenhagen, 2011.

[20] G. Fantoni, G. Tosello, D. Gabelloni, H.N. Hansen, Modelling injection moulding machines for micro manufacture applications through functional analysis, in: 1st CIRP Global Web Conference on Interdisciplinary Research in Production Engineering, 2012.

[21] V. Srinivasan, A. Chakrabarti, SAPPhIRE: Sapphire – an approach to analysis and synthesis, in: International Conference on Engineering Design, ICED'09, Stanford, August 2009.

[22] D. Russo, T. Montecchi, A function–behaviour oriented search for patent digging, in: International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, ASME, 2011.

[23] M. Michael Van Wie, C.R. Bryant, M.R. Bohm, D. Mcadams, R.B. Stone, A model of function-based representations, Artificial Intelligence for Engineering Design, Analysis and Manufacturing 19 (2) (2005) 89–111, http://dx.doi.org/10.1017/S0890060405050092.

[24] D. Brown, Functional, behavioral and structural features, in: ASME Design Engineering Technical Conference Proceedings, 2003, DET2003/DTM-48684.

[25] T. Jensen, Function integration explained by allocation and activation of wirk elements, in: Proceedings of ASME Design Theory and Methodology Conference, Baltimore, MD, Paper No. DETC2000/DTM-14551, 2000.

[26] D. Ullman, The Mechanical Design Process, second ed., McGraw Hill, 1997.

[27] N.P. Suh, Axiomatic Design: Advances and Applications, Oxford University Press, New York, 2001.

[28] D. Gabelloni, G. Fantoni, R. Apreda, A. Bonaccorsi, On the link between features and functions, in: Proceeding of the International Conference on Engineering Design, ICED11, 15–18 August 2011, Technical University of Denmark, Copenhagen.

[29] G. Cascini, G. Fantoni, F. Montagna, Reflections on the FBS model: proposal for an extension to needs and requirements modelling, in: International Design Conference – Design 2010, Dubrovnik – Croatia, May 17–20, 2010.

[30] G. Cascini, System and Method for Automatically Performing Functional Analyses of Technical Texts, European Patent EP1351156, 2002.

[31] G. Fantoni, R. Apreda, A. Bonaccorsi, State machines for functional reasoning,Submitted to AIEDAM .

[32] F. Dell'Orletta, Ensemble system for part-of-speech tagging, in: Evaluation of NLP and Speech Tools for Italian, 2009, Reggio Emilia, Italy, December 2009, Evalita 2009.

[33] D. Bonino, A. Ciaramella, F. Corno, Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics, World Patent Information 32 (2010) 30–38.

[34] J.M. Gerken, M.G. Moehrle, A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis, Scientometrics 91 (2012) 645–670.

[35] E. Ruffaldi, E. Sani, M. Bergamasco, Visualizing Perspectives and Trends in Robotics based on Patent Mining in IEEE International Conference on Robotics and Automation, 2010.

[36] WordNet: <http://wordnet.princeton.edu/> Princeton University "About WordNet." WordNet. Princeton University. 2010, last access date 12th April 2013. See also C. Fellbaum, WordNet: An Electronic Lexical Database, Bradford Books, 2010.

[37] A. Gibbs, Semetric: Conceptual Search and Discovery, Tech. Rep., Patent Cafe, 2005.

[38] B. Yoon, Y. Park, A text-mining-based patent network: analytical tool for high-technology trend, Journal of High Technology Management Research 15 (2004) 37–50.

[39] Hyunseok. Park, Janghyeok. Yoon, Kwangsoo. Kim, Identifying patent infringement using SAO based semantic technological similarities, Scientometrics 90 (2) (2012) 515.

[40] Amy J.C. Trappey, Charles V. Trappey, Chun-Yi Wu, Chi-Wei Lin, A patent quality analysis for innovative technology and product development, Advanced Engineering Informatics 26 (1) (2012) 26–34.

[41] Goldfire Innovator. <http://inventionmachine.com/products-and-services/innovation-software/goldfire-innovator/> (accessed 30.01.13). Invention Machine Goldfire.

[42] S. Borgo, M. Carrara, P. Garbacz, P.E. Vermaas, Towards the ontological representation of functional basis in DOLCE, in: M. Okada, B. Smith (Eds.), Interdisciplinary Ontology, vol. 2, Proceedings of the 2nd Interdisciplinary ontology Meeting, February 28th–March 1st, 2009, Tokyo, Japan.

[43] C.V. Trappey, H.-Y. Wu, F. Taghaboni-Dutta, A.J.C. Trappey, Using patent data for technology forecasting: China RFID patent analysis, Advanced Engineering Informatics 25 (1) (2011) 53–64.

[44] G. Cascini, F. Neri, Natural language processing for patents analysis and classification, in: Proceedings of the TRIZ Future 4th World Conference Florence, 3–5 November 2004, 88-8453-221-3, Firenze University Press.

[45] G. Cascini, D. Russo, Computer-aided analysis of patents and search for TRIZ contradictions, International Journal of Product Development 4 (1–2) (2007) 52–67.

[46] D. Russo, T. Montecchi, Creativity Techniques for a Computer Aided Inventing System, ICED 2011, 2011.

[47] D. Russo, T. Montecchi, L. Ying, Functional-based search for patent technology transfer, in: International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, ASME, 2012.

[48] J.R. Curran, M. Moens, Improvements in Automatic Thesaurus Extraction, 2002, pp. 59–66.

[49] F. Dell'Orletta, A. Lenci, S. Marchi, S. Montemagni, V. Pirrelli, Text-2-knowledge: a computational linguistic platform for the extraction of knowledge from texts, in: Proceedings of the SLI-2006 Conference: 20–28, Vercelli 2006.

[50] <http://aclweb.org/aclwiki/index.php?title=POS_Tagging_%28State_of_the_art%29> (accessed 12.04.13) Wiki Running Under the Auspices of The Association for Computational Linguistic.

[51] R. Watson, Part-of-speech tagging models for parsing, in: Proceedings of the 9th Annual CLUK Colloquium, Open University, Milton Keynes, UK, 2006.

[52] K. Yoshida, Y. Tsuruoka, Y. Miyao, J. Tsujii, Ambiguous part-of-speech tagging for improving accuracy and domain portability of syntactic parsers, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), Hyderabad, India, 2007.

[53] K. Frantzi, S. Ananiadou, The C-value/NC value domain independent method for multi-word term extraction, Journal of Natural Language Processing 6 (3) (1999) 145–179.

[54] S. Vintar, Comparative evaluation of C-value in the treatment of nested terms, in: Proceedings of "Memura 2004 – Methodologies and Evaluation of Multi-word Units in Real-World Applications (LREC 2004 Workshop), 2004, pp. 54–57.

[55] F. Bonin, F. Dell'Orletta, G. Venturi, S. Montemagni, A contrastive in approach to multi-word term extraction from domain corpora, in: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), La Valletta, Malta, 19–21 May 2010, pp. 3222–3229.

[56] A. Lenci, S. Montemagni, V. Pirrelli, G. Venturi, Ontology learning from Italian legal texts, in: Joost Breuker, Pompeu Casanovas, Michel C.A. Klein, Enrico Francesconi (Eds.), Law, Ontologies and the Semantic Web Channelling the Legal Information Flood, Frontiers in Artificial Intelligence and Applications, Springer, vol. 188, ISBN:978-1-58603-942-4, 2009, pp. 75–94.

[57] F. Bonin, F. Dell'Orletta, G. Venturi, S. Montemagni, Singling out legal knowledge from world knowledge, in: Proceedings of the IV Workshop on

"Legal Ontologies and Artificial Intelligence Techniques" (LOAIT'10), Fiesole, 7th July 2010.

[58] F. Bonin, F. Dell'Orletta, G. Venturi, S. Montemagni, Contrastive filtering of domain-specific multi-word terms from different types of corpora, in: Proceedings of the Workshop "Multiword Expressions: from Theory to Applications" (MWE 2010), 23rd International Conference on Computational Linguistics (COLING2010), Beijing, China, August 28, 2010, pp. 76–79.

[59] E. Giovannetti, S. Marchi, S. Montemagni, R. Bartolini, Ontology-based semantic annotation of product catalogues, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, (RANLP-2007), 27–29 September 2007, pp. 235–239.

[60] Google Patent Search. <https://www.google.com/?tbm=pts> (accessed 10.04.13) Google Inc.

[61] A. Bonaccorsi, G. Fantoni, Expanding the functional ontology in conceptual design, in: International Conference on Engineering Design, Iced'07, 28–31 August 2007, Cité des Sciences et de l'Industrie, Paris, France, 2007.

[62] G. Cascini, L. Del Frate, G. Fantoni, F. Montagna, Beyond the design perspective of Gero's FBS framework, in: Paper Presented to the DCC2010, Stuttgart, 2010.

[63] D.C. Brown, L. Blessing, The relationship between function and affordance, in: ASME 2005 Design Theory and Methodology Conference, Long Beach, CA, 2005, Paper No. DETC2005-85017.

[64] J. Hirtz, R. Stone, D. McAdams, S. Szykman, K. Wood, A functional basis for engineering design: reconciling and evolving previous efforts, Research in Engineering Design 13 (2002) 65–82.

[65] G. Pahl, W. Beitz, J. Feldhusen, K.H. Grote, Engineering Design: A Systematic Approach, third ed., Springer, London, 2007.

[66] Wikipedia. <http://en.wikipedia.org/> (accessed 12.04.13) WikiMedia Foundation Inc.

[67] Physical Effect DB. <http://function.creax.com/> (accessed 01.03.12) CREAX NV.

[68] P. Ferragina, U. Scaiella, Fast and accurate annotation of short texts with Wikipedia pages, IEEE Software 29 (1) (2012) 70–75.

[69] G. Fantoni, R. Apreda, D. Gabelloni, A. Bonaccorsi, Do functions exist?, in: International Conference on Engineering Design, ICED11, 15–18 August 2011, Technical University of Denmark, Copenhagen.

[70] Freebase. <http://www.freebase.com/> (accessed 17.09.12) Google Inc.

[71] Comsol. <http://www.matweb.com/search/PropertySearch.aspx> (accessed 24.11.12) MatWeb LLC.

[72] AccessScience. <http://www.accessscience.com/> (accessed 13.11.12) McGraw-Hill Education.

[73] Ansys. <http://www.ansys.com> (accessed 01.03.12) ANSYS Inc., Canonsburg PA, USA.

[74] Catia. <http://www.3ds.com/products/catia/> (accessed 01.03.12) Dassault Systèmes.

[75] PTC. <http://www.ptc.com/product/creo/parametric> (accessed 21.04.12) PTC Inc., USA.

[76] <http://www.utexas.edu/tmm/npl/mineralogy/science_of_minerals/optical_properties.html#Luster> (accessed 02.09.12) Texas Natural Sciences Center, Austin, TX, USA.

[77] Niloy J. Mitra, Yong-Liang Yang, Dong-Ming Yan, Wilmot Li, Maneesh Agrawala, Illustrating How Mechanical Assemblies Work, SIGGRAPH, 2010.

[78] J. Pettibone, Penn Treebank Tags, 2002. <http://bulba.sdsu.edu/jeanette/thesis/PennTags.html> (accessed 13.04.13).

[79] S. Jonas, L. Redmann, Gripper, in Particular a Bernoulli Gripper, Patent US8172288, 2012.

[80] G. Dini, G. Fantoni, F. Failli, Grasping Leather Plies by Bernoulli Grippers, CIRP Annals, 2009.

[81] Graphviz. <http://www.graphviz.org/> (accessed 03.12.09) Open Source, Initially AT&T Inc. .

[82] J. Nivre, M. Scholz, Deterministic dependency parsing of English text, in: Proc. Of COLING 2004, Geneva, Switzerland, 2004, pp. 64–70.

Glossary

DB: Database
DRL: Dependency Relation Label
FBS: Function–Behaviour–State
FBS: Function–Behaviour–Structure
KB: Knowledge base
NLP: Natural Language Processing
POS: Part of Speech
RFB: Reconciled Functional Base
CD: Cardinal number
IN: Preposition or subordinating conjunction
JJ: Adjective
NN: Noun, singular or mass
NNP: Proper noun, singular
RB: Adverb
VB: Verb, base form
VBD: Verb, past tense
VBG: Verb, gerund or present participle
VBN: Verb, past participle