Università degli Studi di Salerno

Dottorato in Scienze della Comunicazione

XV° Ciclo

2013-2016

Advisor Prof. Annibale Elia

Semantic Technologies for Business Decision Support

Discovering meaning with NLP Applications

FRANCESCA ESPOSITO

UNIVERSITÀ DEGLI STUDI DI SALERNO
Dottorato in Scienze della Comunicazione

Semantic Technologies

for Business Decision Support

Discovering meaning with NLP Applications

Francesca Esposito

Advisor
Prof. Annibale Elia

XV Ciclo
2013-2016

A Giuseppe,

passato, presente e futuro della mia vita.

TABLE OF CONTENTS

INDEX OF ACRONYMS

Artificial Intelligence	AI
Business Intelligence	BI
Cloud Computing	CC
Computational Linguistics	CL
Content Management	CM
Decision Support System	DSS
Finite State Automata	FSA
Information and Communication Technology	ICT
Knowledge Discovery	KD
Knowledge Discovery from Text	KDT
Knowledge Management	KM
Lexicon-Grammar	LG
Linguistic Resources	LR
Machine Learning	ML
Machine Translation	MT
Multi-Word Atomic Linguistic Unit	MWALU
Natural Language Processing	NLP
Opinion Mining	OM
Predictive Analysis	PA

Semantic Technology	ST
Semantic Technology Enterprise	STE
Sentiment Analysis	SA
Social Network Analysis	SNA
Text Mining	TM

PREFACE

He who is seeking the truth, bumps into the unable to learn it from those who have already achieved it, since it is the result of pre-conceptual and direct experience, on which every individual can state everything, or its opposite. The greatest masters of all time have challenged, albeit with an act of love, the limit of non-communicability of things, giving us, in exchange, precious pearls to spend along our path to knowledge.

In the moment we undertake this journey, our mind is prone to seek the extraordinary, something special that no one has ever seen before. The nature of our ego drives us to do some immense laps looking for something, that even selves, we know what it is and that unfortunately we cannot find. The pursuit of extraordinary confuses us to the point of not allowing us to understand that what we need is right inside of us. The same mechanism comes into motion when we are going to deal with something, which seems to be far from our perception: we strive to understand, that sometimes, new objects taking advantage from the vision of others, triggering a continuous motion towards the new place. However, is it necessary to become estranged from itself to have a feeling for the new? In the end, we have travelled a tortuous and unsatisfactory way; we begin slowly to understand that to be able to know we must make a journey into ourselves. This thesis summarizes in many ways the

last three years of my life. It contains the intensions, results and potential upgrades of my research, and comes from the combination of my two areas of interest: Computational Linguistics and Economics. Despite they have different roots, they find in Communication their common field of application. I believe that Economics may be useful in Computational Linguistics to better understand languages and language mechanisms, while Computational Linguistics may provide useful insights.

I am very glad I had the chance to meet so many academics and people who influenced the course of my research, who guided me and provided with resources and encouragement. I would like to say thank you to my advisor, Annibale Elia, for his help, support and the directions provided during my Ph.D. Thanks to Maddalena della Volpe for conveying to me the love for research and for giving me the most precious thing she has, that is her time. Finally, thanks to Mario Monteleone for his invaluable comments and patience in reviewing this thesis.

INTRODUCTION

The world is full of data. In order to improve and to be competitive, enterprises should know how to grab the opportunities hidden inside these data. This strategic vision implies a high level of communication sharing and the integration of practices across every business level. This does not mean that enterprises need a disruptive change in their informative systems; rather, they have to converse them, reusing existent business data and integrating new ones. However, data are heterogeneous, so to maximise their value it is necessary to extract meaning from them considering the communication context in which they are used. Inside enterprises, during data analysis and integration phases, the use of Semantic Technology (ST) solutions, and more specifically the adoption of Computational Linguistic tools, from our point of view are necessary pre-requisites.

Due to the increasing competition in Europe and all over the World, nowadays industries are under pressure. Still, not many enterprises have yet recognized the opportunities provided by ST, or once identified them, they are not yet be able to implement them, due to high costs. Consequently, enterprises set-up collaboration and cooperation with other actors of the global market, to enhance internal information system adopting innovative solutions of ST. This

way proves to be efficient to supply competition and, at the same time, to be attractive with reference to potential collaboration, creating a network of knowledge sharing and enduring improvement.

The relationship between actors who revolve around the enterprise, and the collaboration with the same, aim to the fulfilment of several activities, require a "communicative connection" as a necessary condition to understand and share mutually goals, strategic orientations and actionable criteria. Language is a relevant aspect to build solid relations between actors that lead interdependent activities; it becomes important, due to the higher level of autonomy of the companies. If the power is split, it will be important to express a shared and unambiguous language. The different nature of enterprises generate a variety of language, based on peculiarities of relation between actors, although the need of unitary move us to recognize and record a strategic language (Vaccà, 1986).

Even if the innovation mind-set could be the same for everyone, technical solutions change their settings in each enterprise. Such a change depends on different cultures, resources, languages and tools that are part of the enterprise itself. The form of the elements that constitute the architecture of an enterprise derives essentially from two variables: the phase of development in an enterprise life cycle; and the business sectors in which it operates. Thus, every technology introduced must be calibrated on the needs which the enterprise has at that time.

Usually, when we refer to an enterprise, we believe to employ only the typical language of Economy, but this vision reduces the possibility to understand the real language used. As we said, the features of enterprise derive

from their life cycle and sector(s): consequently, even the language used to express themselves and collaborate with others concerns and involves different knowledge domains.

The language used by enterprise in the early phases of its life is very special: we could define it as a sublanguage of Economy, since terms are associated clearly to specific meanings that overcome the general language. Moreover, this has an effect above all on the efficiency of that communicative connection: actually, a strategic and general language may not be sufficient, but these kind of relations require a deep knowledge level of the process and terms utilised to express them.

This study aims at being a first approach to the development of a document-driven Decision Support System (DSS) based on NLP technologies within the theoretical framework of Lexicon-Grammar (LG) established by Maurice Gross. Our research project has two main objectives: the first is to recognize and codify the innovative language with which the companies express and describe their business, in order to standardize it and make it actionable by machine. The second is to use information resulting from automatic text analysis to support strategic decisions, considering that through Text Mining (TM) analysis we can capture the hidden meaning inside documents. Language used in business is not always recognized by traditional analysis systems; the issue concerns on one hand the language of a startup and its activities in general and on the other hand the fact that there is a terminology that describes the innovative products relating to the field of reference. The automatic analysis of natural language makes possible the comprehension of the meaning and the retrieval of information hidden in a large amount of

business documents. This approach contributes to the growth and improvement of a single segment (such as sales or communication) or involves the entire business. The development of this task would contribute to the optimization of the execution phase; also, it would respond to the needs of:

- *entrepreneurs*, as it allows them to position themselves in the market understanding which aspects have to improve in business communication;
- *investors*, who may assess in a much more efficient way business projects and how to fund them;
- *sectors*, since this approach increases the competitiveness and spreads the attitude to improvement and continuous innovation.

This proliferation of new terms linked to the growth of STs creates problems in conducting research and communicating with decision-makers about decision support systems. The best solution to seek is developing an expanded and well-defined framework for categorizing terms used in DSS. A DSS is not a new concept, but it is complex and in evolution (Power, 2001). It can be typically approached from two major disciplinary perspectives, the one of Information systems science and the one of Artificial Intelligence (AI). In our work, we added a linguistic approach for a DSS, which is based on business document analysis. With this extension, the result of business document analysis could acquire benefits with reference to statistical methods based on keywords or single terms.

In the first chapter, we examine the concept, characteristics and different types of DSS, with particular reference to the changes that these systems have experienced due to Web developments, and consequently the information system perception within companies. The decision-making process becomes more and more complex: the aim is to make decisions in the shortest time possible, at a reduced cost and knowing the most of the system in which decision making operates. Indeed, the attention will focus on the ill-structured decisions typical of complexity, in which companies operate, not only based on objective and measurable data but also on predictions, insights that especially attract the management's ability to grasp the challenges of the future. Afterwards, we describe the characteristics of a particular type of DSS of our interest: the document-driven system. It is in this system of analysis that the language support regains its dimension. Managers need to transform documents into strategic knowledge, thus it is essential that they take a large amount of information sufficient to evaluate the actions to be undertaken; moreover, the less visible data (not recognized by the most of statistical systems) can only be understood through a systematic study of the linguistic properties of the language used. Companies still find considerable difficulties in relying on linguistic analysis solutions, even introducing them in everyday data analysis procedures. Considered a distant discipline from Economics, Business Organization and Strategic Management, Linguistics, discovers its extension and his way for supporting these processes through the collaboration with computer technologies. Then, through the description of three basic elements (Data, Technology and Enterprise), we identify a decision-making model based on the introduction of new information, which are sorted,

processed and incorporated into the company structure, with the support of linguistic rules on the intelligence phase.

In the second chapter, we start with a brief review on Computational Linguistics, paying particular attention to goals, resources and applications. Computational Linguistics is closely linked to computer engineering development, as it takes advantage of other disciplines as Machine Learning, Artificial Intelligence and Cognitive Science. Computers, at the service of human labour, were increasingly also used in business. In particular, we propose a Computational Linguistics tool useful to analyse large collections of documents taken from different areas of business knowledge. In order to identify the theoretical and methodological framework in which natural language formalization develops, we focus on LG method, a linguistic theory established during 60's by Maurice Gross. After having reviewed the main influences that the lexical-grammatical theory has suffered in its development, we will highlight its predisposition for automatic analysis, shown by the efficiency of specific linguistic software packages such Intex, Unitex, NooJ, and Cataloga. As the volume of textual data increases, we need technology that can cope efficiently with this immense variety of linguistic accomplishments. Moreover, mark-up languages like XML and the increasing mutual exchange among similar types of texts has allowed giving explanation about the structures of texts. In our research, we do not deal with singular texts, but we refer to the concept of corpus and his annotating phase. Starting from this phase of annotation, we try to explain the concept of TM and how to create an efficient research environment according to LG methods.

In the third chapter, we define STs and their process of diffusion in the innovation market. Leading the enterprise towards a global cognitive approach is the main tool to catch innovation challenges and compete on international markets. Therefore, we describe the main sectors in which STs are used and the typical applications showing the objects that define the usefulness of STs for business. Although integrating semantic technology in enterprises requires big efforts in terms of economic resources, there is a growing tendency to outsource data analysis activities, thus allowing semantic technology market to define itself through the explosion of Semantic Technology Enterprises (STEs). To understand the dynamics underlying this diffusion and the general trends in this market, we conducted a census on approximately 200 international STEs that offer TM or Knowledge Discovery from Text (KDT) services. A data set of STEs has been created with organizational characteristics, geographical recognition, target market, main activities, applications, company development phase and sources of analysed data. The main goal is to understand the characteristics of those enterprises using STs, how they employ these technologies to improve business processes with reference to the difficulties of market, finally how linguistic support can produce benefits. This analysis has evidenced a concentration of STEs in the United States and approximately 65% of the total number of them are emerging companies in their early stages, using semantic technologies especially to improve their business process.

In the fourth chapter, we propose a model of linguistic analysis, according to LG, in order to create an enriched solution for document-driven decision (DDS) systems. Specific features of business language are shown in

the model through some results on experimental research work. The language of business deliver the complexity of language, and the problems linked to its formalization, with particular reference to the specialty languages and Multi-Word Atomic Linguistic Units (MWALUs) treatment. Business documents are full of technical terms that describe and define processes, actors, activities. Language support analysis of documents is not functional if it cannot locate this terminology items, linking them immediately to a specific meaning. Then, language of business is extremely varied: in these documents, we find formal and technical expressions, together with other jargons. Dialects belonging to disciplines related to the Economics, such as Finance, Business Organization, not considering that many of these terms are in English or American. For this reason, before proceeding with the linguist analyses, we have reviewed the documents that companies typically use in their activities or that contain information that is not negligible during the decision-making processes. Using electronic dictionaries, it is possible classify these terms with reference to specific domains. In this research, we deal with the language used in the start-up ecosystem. Besides containing the typical business language, the language of start-ups undergoes a much more accelerated continuous innovation process of mature businesses. The characteristics of this business entity flock in the way it expresses and communicates with its stakeholders.

Finally, discussion and conclusion end this disquisition. In the last section, we briefly evaluate our results, presenting possible future works that open up new scenarios.

1

SEMANTIC TECHNOLOGIES TO SUPPORT BUSINESS DECISION

1.1 Decision Support Systems

A DSS is a software system designed to increase the effectiveness of data analysis as it provides support to all those who need to make strategic decisions while coping with problems not solvable by means of operative research models. A DSS is typically a computer program application that analyses business data and presents it so that managers can make business decisions more easily. In fact, the main function of a DSS is extracting, from a significant amount of data, in a short time and in a flexible way, the information useful to the decision-making processes. It relies on data stored in a database or knowledge base, and it is not only a computer application, because it also contains business intelligence tools and expert systems technologies, such as decision support models.

DDSs find their roots in the IR field, may also include an expert system or AI routines, and may present information graphically. The continuous changes and the exponential growth of the amount of data that the companies

have to analyse in order to compete in the market, impose them the need to develop DSSs, integrating new functions and new applications. In this chapter, we examine the basic concepts, features and types of general DSSs, then focusing on document-driven DSSs. Also, we will add a deep natural language level to DSSs: the use of linguistics in DSS models is not expected, and cannot be based only on statistical analysis methods. Due to particular characteristics of the enterprise, the use of linguistics has to be recognised and suitably sustained.

1.1.1 Concept

DSSs are generally defined as a collection of computerized data system that supports some decision-making actions (Power, 2000). In the 1960s, researchers began systematically studying the use of computerized quantitative models to assist in decision-making and planning, at any level of organization. Sprague and Carlson, quoted by Power (2008), defined a DSS as "a class of information system that draws on transaction processing systems and interacts with the other parts of the overall information system to support the decision-making activities of managers and other knowledge workers in organizations" (1982). A properly designed DSS can play an important role in compiling useful information from raw data, documents, personal knowledge, and business models to solve problems (Niu, Lu, & Zhang, 2009). Decision-making is the process of developing and analysing alternatives to make a decision, a choice from the available alternatives. Decision problems can be

classified according to their characters. Turban *et al.* (2005) identified three types of decisions referred to a given problem structure: structured, semi-structured and unstructured.

A *structured decision* can be descripted and solved by traditional mathematical and statistics methods. These types of solution are known standard solution methods: an example is the choosing of a product among others basing on price evaluation.

An *unstructured decision* problem is uncertain and ambiguous: differently from the previous one, for this type of decision there is not a standard solution method. Typical unstructured problems contain planning new services, hiring an executive, and selecting a set of projects.

Semi-structured decision problems include structured and unstructured solutions; it is a combination of both standard optimized solution procedures and human intuition or judgments.

Thus, DSSs can be either fully computerized, human-powered or a combination of both. Niu, Lu, and Zhang (2009) called *ill-structured solutions* the last two (unstructured and semi-structured), because for the problems linked to these there are not clear procedural or predetermined solving path, therefore solutions may vary greatly. Actually, the hardest challenges for businesses are to make decisions in our changing environment: this implies the cooperation between Computer Science and Human Science to develop mutually sophisticated solutions satisfying the need to make the best decisions in the shortest possible time and at the lowest cost. Managers come across fuzzy interrogatives that they must answer. Most decisions are taken to solve problems, or generally improve business performances, based on uncertain

solutions: with respect to the first type of decision (structured), it is important for actors pore on the improvement of *ill-structured solutions*, advantaging the relation and collaboration between humans and computers.

Making decisions is now even more difficult for companies, which have to deal with the complexity of the business system: it means doing in-depth analysis of each asset and reconsidering the rapid change of business equilibria. Moreover, the collaboration between computer and linguistic tools in decision systems allows the pondering of opportunities useful to understand and explain the facts contained in business documents: choosing not only between 1 and 0, but also among an infinite number of possibilities other than 1 and 0.

An interested upgrade in Physics, related to Majorana fermion, could bring important progresses in the field of IR. Recently, the researchers of Oak Ridge National Laboratory have demonstrated the presence of this fermion and the possibility to generate it. By using these particles as elementary units, actually quantum supercomputer could be realized capable of dwarfing current ones coming to performances simply unthinkable until now. Traditional processors, based on electronics and semiconductors (those found in our laptops, smartphones and tablets, for instance), store and process data in the form of bits, small units of information that can take on the values 0 and 1 and that encode, respectively, the passage or a power outage. Quantum computers, however, make use of so-called "qubits" (quantum bits, or, to be precise), which encode the quantum state of a particle and can store much more information than the only two possibilities of traditional bits. The idea of a quantum computer that behaves such as quantum physics phenomenon for the

treatment and processing of information is not new and allows explaining better the language phenomena. Already in 1982, the famous physicist Richard Feynman expounded his idea of this kind of machines. Since then, with the passage of time, many discoveries and innovations have followed which could make feasible this device, and many fields of studies have been involved, from nanotechnology and spintronic, to quantum cryptography to the logic. Several technological solutions already exist to build this kind of computers, going from carbon nanotubes for memories to quantum correlation for communication, and up to superconducting materials and self-assembling. The Majorana fermion could be the spark which gives life to all this and leads to a new relation between Human Sciences and Computer Sciences.

1.1.2 Features

DSSs are designed to help support decisions that are formulated mostly as semi-structured problems. These problems remain resistant to complete computerization, and require human intervention. The features of a general DSS change by different types of analysis or applications fields. Alter (1980) identified three major characteristics of DSSs:

1) DSSs that are designed specifically to facilitate decision processes;
2) DSSs that should support rather than automate decision making;
3) DSSs that should be able to respond quickly to the changing needs of decision makers.

Each of this type is expressly designed for either a particular decision-maker or a group of decision-makers. This allows the system designer to customize important system features, in order to adapt them to the type of representations. Thus, Power (2003) identifies seven supplementary features common to different type of DSSs:

a) *Facilitation.* DSSs facilitate and support specific decision-making activities and/or decision processes.
b) *Interaction.* DSSs are computer-based systems designed for interactive use by decision makers or staff users who control the sequence of interaction and the operations performed.
c) *Ancillary*. DSSs can support decision makers at any level in an organization. They are not intended to replace decision makers.
d) *Repeated Use.* DSSs are proposed for repeated use. A specific DSS may be used routinely or used as needed ad hoc.
e) *Task-oriented.* DSSs provide specific capabilities that support one or more tasks related to decision-making, including intelligence and data analysis, identification and design of alternatives, choice among alternatives and decision implementation.
f) *Identifiable.* DSSs may be independent systems that collect or replicate data from other information systems or subsystems of a larger, more integrated information system.
g) *Decision Impact.* DSSs are intended to improve the accuracy, timeliness, quality and overall effectiveness of a specific decision or a set of related decisions.

Many DSSs are designed to support specific business functions or types of businesses and industries. For this, nowadays the features of DSSs are less defined, because the continue changes which companies are subjects require the interrelation between different parts, the application of various disciplines and the consequent generation of hybrid models of DSS which are configured to the needs of the individual company.

1.1.3 Types

DSSs have been classified in different ways, according to the opportunities offered by technologies and as the results the encounter of different fields. According to Donovan and Madnick (1977), DSSs can be classified as:

- *Institutional*, when they support continuing and recurring decisions;
- *Ad hoc*, when they support a one off kind of decision.

Hackathorn and Keen (1981) classify DSSs with reference to organizational strategies, identifying three levels of increasing interdependence:

a) Personal DSSs, focused on an independent user or class of users;
b) Group DSSs, focused on a group of individuals, each one engaged separately but strictly interrelated (i.e. office work)

c) Organizational DSSs, based on an organizational task or activities involving a sequence of operations and actors.

Alter (1980) states that decision support systems could be classified into seven types based on their generic nature of operations. The following Table 1 shows his description and classification of these seven types of DSSs:

File drawer systems	This type of DSS primarily provides access to data stores/data related items
Data analysis systems	This type of DSS supports the manipulation of data with specific or generic computerized settings or tools.
Analysis information systems	This type of DSS provides access to sets of decision oriented databases and simple small models
Accounting and financial models	This type of DSS can perform 'what if analysis' and calculate the outcomes of different decision paths.
Representational models	This type of DSS can also perform 'what if analysis' and calculate the outcomes of

	different decision paths, based on simulated models
Optimization models	This kind of DSS provides solutions using optimization models, which have mathematical solutions.

Table 1. The seven type of DSSs by Alter (1980)

In the last decade, upgrades in technology created new computerized decision support applications in many disciplines. They consist of interactively computer-based systems that emphasise manipulating quantitative models, accessing and analysing large databases, and affect decision-making structures. According to Power 2002, the applications are of the following types: model-driven DSSs, data-driven DSSs, communication-driven DSSs, document-driven DSSs, knowledge-driven DSSs and web-driven DSSs.

A *model-driven* DSS emphasizes access to and manipulation of a statistical, financial, optimization, or simulation model. Model-driven DSSs use data and parameters provided by users to assist decision makers in analysing a situation; they are not necessarily data-intensive (Gachet, 2004).

A *data-driven DSS* refers to the access and manipulation of a time-series data, both internal to a company, and external. Simple file systems accessed by query and retrieval tools provide the most elementary level of decision support functionality. Data warehouse systems often provide additional functionalities. Analytical processing provides the highest level of functionality and decision support linked to the analysis of large collections of

historical data. Some data-driven DSSs use real-time data to assist during operational performance monitoring.

A *communication-driven* DSS supports decision making within a group of decision makers, through the facilitation of efficient information exchange. Communications technologies are central to supporting decision-making and include LANs, WANs, Internet, ADSL, and Virtual Private Networks; moreover whiteboards, Video conferencing, and Bulletin Boards.

A *knowledge-driven DSS* is normally a person-computer system with specialized problem-solving expertise. The expertise consists of knowledge about a specific domain, understanding of problems within that domain, and "skills" at solving some of these problems.

A *document-driven DSS* integrates a variety of storage and processing technologies to provide complete document retrieval and analysis to assist during decision-making. This type of DSS allows managing unstructured documents and Web pages. The Web provides access to large amounts of documents[1]. In this study, we deal with documents presented in the form of written texts. Examples of these documents used in the document-driven analysis are policies and procedures, plans, product specifications, catalogues, and corporate historical documents, including minutes of meetings, corporate records, and important correspondence.

These different models are not mutually exclusive, and they can be used in a mix version, depending on the needs of enterprises. The DSSs are generally capable of collecting and representing several types of information,

[1] Documents included are hypertext documents, images, sounds and video. In fact, Power (2001) defined document such as not only written documents.

such as comparative data, accessing information assets that include relational and legacy data sources, providing past experiences as well as projecting future choices, according to assumptions or new data, consequences of decisions. This turns out to be necessary for managers and companies that are transforming the way in which they are designing and constructing new information management capabilities. Companies are increasingly looking for new data to use as strategic resources, developing their analysis capabilities and refining tools that support concretely and help top managers to take better decisions: they fit with a huge amount of data that often appear unstructured, not categorized in traditional data warehouses. Table 2 is a survey proposed by Power (2001) on the characteristics of DSS classification framework. At the left column list, we find the five generic categories of DSSs that differ in terms of the dominant technology component used; subsequently, we can observe other elements as target users (internal/external), purpose of technology (general/specific), and typical deployment (technology). According to Power, document base systems concern specifically internal users, with the rapid expansion of the group, due to the complexity of the analysis. In the analytical process, such systems involve intra-organizational structure and many professionals (such as linguists, data scientists). They affect different level of enterprise, after the taking-decisions time. The purpose refers to Web documents or business documents: it can cope with general or specific topics. Then, the deployment technology focuses on the Web or client/server.

Dominant DSS Component	Target Users: Internal →External	Purpose: General → Specific	Deployment Technology
Communications **Communications-Driven DSS**	Internal teams, now expanding to external partners	Conduct a meeting or Help users collaborate	Web or Client/ Server
Database **Data-Driven DSS**	Managers, staff, now suppliers	Query a Data Warehouse	Main Frame, Client/Server, Web
Document base **Document-Driven DSS**	Internal users, but the user group is expanding	Search Web pages or Find documents	Web or Client/ Server
Knowledge base **Knowledge-Driven DSS**	Internal users, now customers	Management Advice or Choose products	Client/Server, Web, Stand-alone PC
Models **Model-Driven DSS**	Managers and staff, now customers	Crew Scheduling or Decision Analysis	Stand-alone PC or Client/Server or Web

Table 2. Adapting of DSS Framework by Power 2001

1.1 Support business with document-driven analysis

For our research, we will focus on some specific types of DSSs mentioned above. The document-driven DSS is a relatively new field in decision support, nowadays focused on the information retrieval and management of unstructured documents. Particularly, decision support driven by documents aims to manage, retrieve and manipulate unstructured information in a variety of electronic formats (Power, 2000). It is the case of business documents: companies use them to describe their activities, characteristics, or strategies. We can divide such documents into two

categories: oral and written. Oral documents consist in conversations transcribed[2], while written documents can be written reports, strategic plans, catalogues, memos and even e-mails. These unstructured forms of knowledge require the use of technologies that are able to recognize the hidden meaning in the texts. They assist manager and IT unit of companies in knowledge categorization, deployment, inquiry, discovery and improving communication.

Back in 1996, Fedorowicz estimated that American businesses store almost 1.3 trillion documents and only 5 to 10 percent of these documents were available to managers for use in decision-making, just because they were not standardized in a uniform pattern or structure. Managers need a way to transform these documents into usable formats that can be matched and processed to support decision-making (Peterson, 2016). Moreover, in 1998 Merrill Lynch sustained that somewhere around 80-90% of all potentially usable business information may come from unstructured form. This rule is not demonstrated scientifically, but it expresses a principle, often inferred from experience, that it is indicated as valid in most cases: as increasing contents recovered by the Web, unstructured data, such as texts, will become the predominant data type stored online. In recent past, data grew too fast and exceeded human capacity to retrieve and utilize. As already mentioned, managerial decision-making process can be highly dependent on hidden information in text documents. However, careful reading and sorting of large

[2] Kowal and O'Connel (2004) argue about transcription conversations intending them as understanding of graphic representation of individuals behaviour in a typical conversation. They recognized several elements that must be involved: transcribers, a system of notation, the product of transcription and the transcription readers (p.248). Social media emphasize studies in this area, with frantic production of large number of conversation.

amount of documents is a time-consuming work. This type of activity wastes working time of managers and, in the end, can even cause wrong decisions. Companies could gather innovation challenges using this type of technologies that combine different textual data and create a new generation of business functions, with the support of specific knowledge. The document-driven type is the most common support application targeted among a wide base of users. Its main function generally is to search Web pages and look for documents based on precise set of keywords or terms. Besides, it is possible to use it to convert documents into important business data.

Before analysing the computational linguistic approach to document-driven DSSs, we must understand the role that decision-making process have in the organizational structure of the enterprise.

1.3 Decision-making model: enterprise, data, technology

The Decision-making process involves different level of the enterprise. It could operate at the *strategic level* in which a small group of managers determinate the objectives, resources and policies of the organisation, with reference to complex problems. Major problems at this level of decision-making are predicting the future of the organisation and its environment, and matching the characteristics of the organisation to the environment. Then, the process of decision-making takes an *operative level* when we have to determine a specific tasks setting, due to the strategic decisions of the management, that is: determining which units in the organisation will carry out the task, be spending resource and evaluating requires decisions. These two

levels of decisions involve the whole enterprise, from human to financial resources, technologies to processes, internal to external dimension, lowest to highest ground of management. We will see how in the document-driven DSSs every decisions that implies and innovation or changes in the structure operate on three elements: *enterprise* intended as corporate culture; *data* useful to take decisions and which come from the process of retrieval and selection; *technology* used, which have an impact on the time spent for data processing and its effectiveness.

Decision-making identifies the process useful to concluding which decisions need to be made and how to find alternatives for each decision. In general, there are several methods or tools available for solving a particular decision problem. Different methods share the same process. The recognition of a DSS develops in four main phases:

- *intelligence phase*: the phase in which we search for and collect data and information from the internal, identifying the real problem to be addressed;
- *design phase*: the phase in which we develop the decision-making plan, analysing and generating possible alternative actions;
- *choice phase*: the phase in which we select a path of actions among those available, evaluating the best choice and optimal solutions;
- *implementation phase:* the phase in which the DSS is achieved by implementing and adopting the selected solutions.

Finally, there is the feedback for the evaluation of the results, in order to change the decision, if necessary.

The creation of a DSS must meet specific requirements, related to the characteristics of the decision-making and to the user needs. Flexibility must be a fundamental requirement of the DSS, since there are various types of problems, decisions, data or users that represent different ways of processing.

Documents provide a package of information, which people and business use, while interacting. Common business documents are catalogues, invoice, orders, reports and plans. The ability to recognize linguistic features in texts, even concepts and sentiments, and to extract them to databases, is now an important feature for these tools, and also an opportunity for companies. The collaboration between ML and linguistic tools, that support functions such as semantic disambiguation, is the only way to transform texts into manageable knowledge (Grimes, 2016). In the major part of cases, the ability to understand the semantic aspect of texts is not yet developed, therefore the meaning contained in a document is hidden. The complexity of texts and the potential knowledge hiding require a much more comprehensive study of natural language characteristics, terminology and resulting semantic annotation; such a study should describe the use of language and connect the various terms in specific meanings sectors, as we will see subsequently with the semantic expansion concept. This is partly because of a lack of understanding by companies with respect to the situations in which semantic technologies can add value to their business. This stands as a valuable reason for introducing a new framework aiming at enabling industries to use semantic technologies in their business (Benjamins, 2011). Innovation challenges is disregarding the traditional functions of information management, focused mainly on analysis of internal data storage: companies must consider many variables and sources.

Even if companies recognize that new information systems are required, in practice this requirement has not yet been fulfilled. The continuous flow of data, produced by enterprises, requires a new sampling method to manage information: an appropriate way to analyse and collect each single datum regarding its features.

Introducing DSSs based on this strategic orientation is much easier to apply with startup companies[3], while it requires a profound change in the mentality of ICT departments of big company activities. The learning culture and the willingness to revise opinions on management systems is not immediate inside certain organizations. Many companies offering semantic technology solutions focused on external data analysis, as can be conversations on the Web, opinions and data about economic trends. However, we have to consider the role of internal information. They usually appear in unstructured

[3] The term *startup* identifies the operation and the period during which a business starts. In a startup, there can be acquisitions of current technical resources, defining hierarchies and methods of production, of personal research, but also market research with which we try to define the activities and business addresses. As Graham (2012) explains, a startup is a company designed to scale very quickly. It is this focus on growth unconstrained by geography, which differentiates startups from small businesses. A simple restaurant in one town is not a startup, nor a franchise is. In recent years, popular lexicon has begun equating startups with tech companies, as though the two are inherently intertwined. The one of startups in not a new concept but in the last years the development of technologies has defined and evaluated the role of startups in the economic growth of a country.
In Italy, startups have been introduced juridically in 2012, with the Decree n. 179, the so-called "Decreto Crescita 2.0", in addition to a series of urgent measures for the development of the country. It introduced into Italian law the notion of innovative startups, providing important benefits and simplifications for those people planning to undertake this entrepreneurial path. The aim of the measure is to make Italy an attractive and hospitable country for the creation and development of innovative companies, by introducing a framework for intervening on several fronts. By the enactment of the decree, today in Italy there are about 6,500 startups, acting in many different sectors.

form, but contain a huge potential: they belong to the company, have already been targeted, and should not go lost in the sea of information that nowadays crowd the Web. The collaboration between internal and external data allow the development of a series of tools and multidimensional analysis models. In a general DSS, various factors are taken into consideration for making managerial decisions:

1) Data
2) Decision rules
3) Mathematical models
4) Managerial knowledge
5) Human judgement

Decisions are based on set of rules that depend on data: however, it is up to the manager to decide which rules to apply. A selection of the most appropriate decision rule depends on the manager's judgement; as well, he establishes the goals for determining it. The set of decision rules is considerably variable, because it requires a forecast of solutions. The data contained are of historical and current kind, thus, forecasts are made with high-technology systems, based on mathematical or statistical models. The modelling component of the DSS provides the forecasting model. Mathematically based forecasts are not without error, however, and they may not consider all relevant factors. Thus, it is necessary to use human judgement to interpret results and correct forecasts. In the case of document-driven DSSs, the process of decision-making requires a closely collaboration between humans and computer: linguists work both before and after the analysis of texts. Managerial judgements are based on the features of the organisation and

its environment, so the role of enterprise knowledge is fundamental. This knowledge is not included in a database and is rarely kept in computer systems.

Our approach tries to capture new data and information to integrate them with already existent knowledge, with reference to specific problem area. Data tend to be numeric, while information are much qualitative or textual in nature, and interference engines may use qualitative reasoning rather than quantitative models to reach decisions. The development of DSSs integrates some of these components, although we identified three element that deeply interconnect to the introduction of document-driven analysis in a DSS: enterprise, data and technology. In the following sections, we will examine in depth the role of each element.

1.3.1 Enterprise

The theme of corporate culture is the nerve in business strategy. Corporate culture indicates the goals to reach, favourites decisional processes and contributes to determinate coherent and unambiguous behaviours. It represents the knowledge and organizational resources shared by each members but applied with different benefits in the enterprise. Corporate culture consists in common rules, ways of thinking, ethic codes of behaviour, beliefs, experiences, languages and procedures that establish values and develop working methods (Hodgson, 1996). For this reason, the factors mentioned here are closely correlated with the process of decision-making: they lead actions and strategies. Traditional analytics should consider the whole enterprise, trying to understand how the sharing of technology and human resources data

across all organizational units can lead to the achievement of general analytical goals. For innovative startup and early adopter companies, it is necessary to coordinate data collection at all company levels (Davenport, 2007). In large organizations, it is necessary to introduce the analysis of data to assist all various departments: marketing, finance, product development, strategy and information systems. Establishing where to place this type of technology within the enterprise depends on the goals. An aspect not to be underestimated is the human one, which must always be more specialized in data management and transfer knowledge: it is from the combination of man and machine that we can get the best results. Furthermore, each sector has a different way to communicate its own business. For example, if we analyse a Business Plan written by an Agri-food company, we may find in the text many words referring not only to Agriculture knowledge field, but also to others as Biology, Chemistry, Botany, Ecology and more. Therefore, it is necessary knowing the company features, its market, its products, and its development phase.

1.3.2 Data

Every day around the world, millions of digital data are created. The term used to describe this large volume of data is Big Data[4]. Although the amount is considerable and companies are inundated of data, what really

[4] In 2001, the analyst firm MetaGroup (now Gartner) introduced us to the three V of Big Data: Volume, Velocity and Variety. Updates and announcements of new products and solutions push organizations to use their arms around these three Vs. At these three Vs, they are added other two new Vs: Value and Veracity, which have entered the equation, making the Big Data ecosystem even more complex, and more important to solve.

matters is the way in which organizations may utilise these data. It is common opinion that the improvements made possible by Big Data are a function of the data collected. Actually, the growth of the amount of data without the ability to process them is useless in itself. Big data can be analysed especially for insights that lead to better decisions and strategic business moves.

Big Data processing requires considerable computing power, appropriate technology and well-defined resources, which go beyond current management systems and data storage capabilities. These data have aroused a great interest from the academic and business world: an impetuous demand for services has generated an uncontrollable range of analytical Big Data solutions. This market was characterized in 2014 by an increase of + 25%, due not just to the maturity of the analytical tools used as a storage services and low-cost clustering that have attracted the interest of small businesses and large enterprises. Big Data are high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing enabling enhanced insight, decision-making, and process automation (Gartner, 2013). There are two categories of textual data: unstructured or structured. Structured data are the typical elements stored in a database, carefully selected and categorized. These data are usually the result of years of work, in which they are collected, sorted and stored. Unstructured data usually refer to data that do not reside in a traditional row and column database. For example, unstructured textual data include text and multimedia content such as emails, documents and more. This type of data is more difficult to analyse, especially when we are in front of a very large number of documents. Making correct decisions often requires analysing large volumes

of textual information and in spite of this it has more potential to recover crucial information for businesses (della Volpe, 2013). In fact, our approach aims to extract useful knowledge from document collections helping then managers to make better decisions. Regarding data, one of the most important challenges is developing the ability to collect them in real time. Businesses today are increasingly more likely to analyse data coming from the outside: documents, conversations on the Web, reports, opinions and many other types of sources. External data are important because, other than providing data in support of governance, they make us never losing touch with reality. These useful data must be integrated with already existing data in the enterprise, past or present: only in this way, it is possible to sketch a complete sight of a business system.

1.3.3 Technology

TM is new and exciting area of Computer Science that tries to solve the problem of information overload, combines techniques of Data Mining, Machine Learning, Information Retrieval, and Knowledge Management (Feldman & Sanger, 2008). TM is proposed to extract meaning from data, making visible their hidden meaning (Radovanovic & Ivanovic, 2008; Feldman & Sanger, 2007). The attention for opportunities that could arise from the analysis of these type of data, attracts businesses, institutions and society never before in the history of internet. Especially, businesses play a central role in this scenario, since they are usually looking for new information to be used

as a strategic resource. However, one reason that is having more interest in business education is the ability for companies to monetize this information by selling them to other companies. It is a very interesting phenomenon: companies that produce data analysis software, in the last years innovated and converted their systems offering online and offline solutions; at the same time, we are witnessing an expansion of market, due to the birth of startups.

It is opening a new sector of data analysis dedicated to enterprises that need to be adept in transforming information into knowledge to improve their performances. The flexible and continuous reconnaissance of resources, based on this new knowledge, is the only way to resist to the continuous change of the surrounding environment. The term *semantic technology* is very broad and encompasses a number of techniques used to extract meaning and applications. In this context, we refer to the semantic technologies that deal with data analysis of Web texts, and then we will deal with the drawing of the boundaries of TM (Blomqvist, 2014). Both NLP and TM offer tools to extract non-trivial knowledge from free or unstructured texts, typically using Part-Of-Speech (noun, verb, adjective, etc.) and grammatical structures (Kao & Poteet, 2007). Actually, a large part of business information appears in unstructured form: emails, letters, reports, transactional documents and financial documents. Business documents are drawn in a specific format and with specific lexicon, based on the different type of business coped with. The lexicon used by each enterprise to express itself is very singular, so it is necessary that researchers have a good knowledge of the semantic field pertaining to such lexicon before proceeding to POS-tagging, remembering that terminological words are

semantically univocal for all specialists in their specific sector of use (De Bueriis & Elia, 2008).

The process of support based on documents analysis showed in Figure 1, and which influences decision-making process, can be developed only evaluating three basic elements: enterprise, data and technology. Cooperation between these three elements generate a process that lead managers to and during their decision phase. Subsequently, decisions have to be communicated to and shared with each organizational level of an enterprise. In this way, the data retrieved travel within the enterprise, influencing performances. In fact, the last step process, that we called embedding phase, indicates the integration and transfer of document-driven analysis findings, in order to influence and improve business performances.

We try to correlate the embedding data process, together with the typical phases of DSSs: the intelligent phase is located between data management (collect, select and identify document) and technology performing (apply technologies to extract data from texts). This phase is very important, because it is the one in which we can apply linguistic rules and algorithms to support the analysis and to have best results. Then, we proceed with understanding: once retrieved information from texts, we recognize how they can be employ in decision-making process and design the useful actions. The assessment of choices requires top management interventions, not only to take decisions but also to develop the strategy of implementation and integration of these results and decisions at every level of a company. This process triggers a cycle of continuous improvements based on knowledge sharing: the implementation of solutions coming from document analysis

creates new assets inside the company and with external stakeholders. After the embedding of decisions, this releases a special feedback, which can identify new needs and thus activate a new process of document analysis.

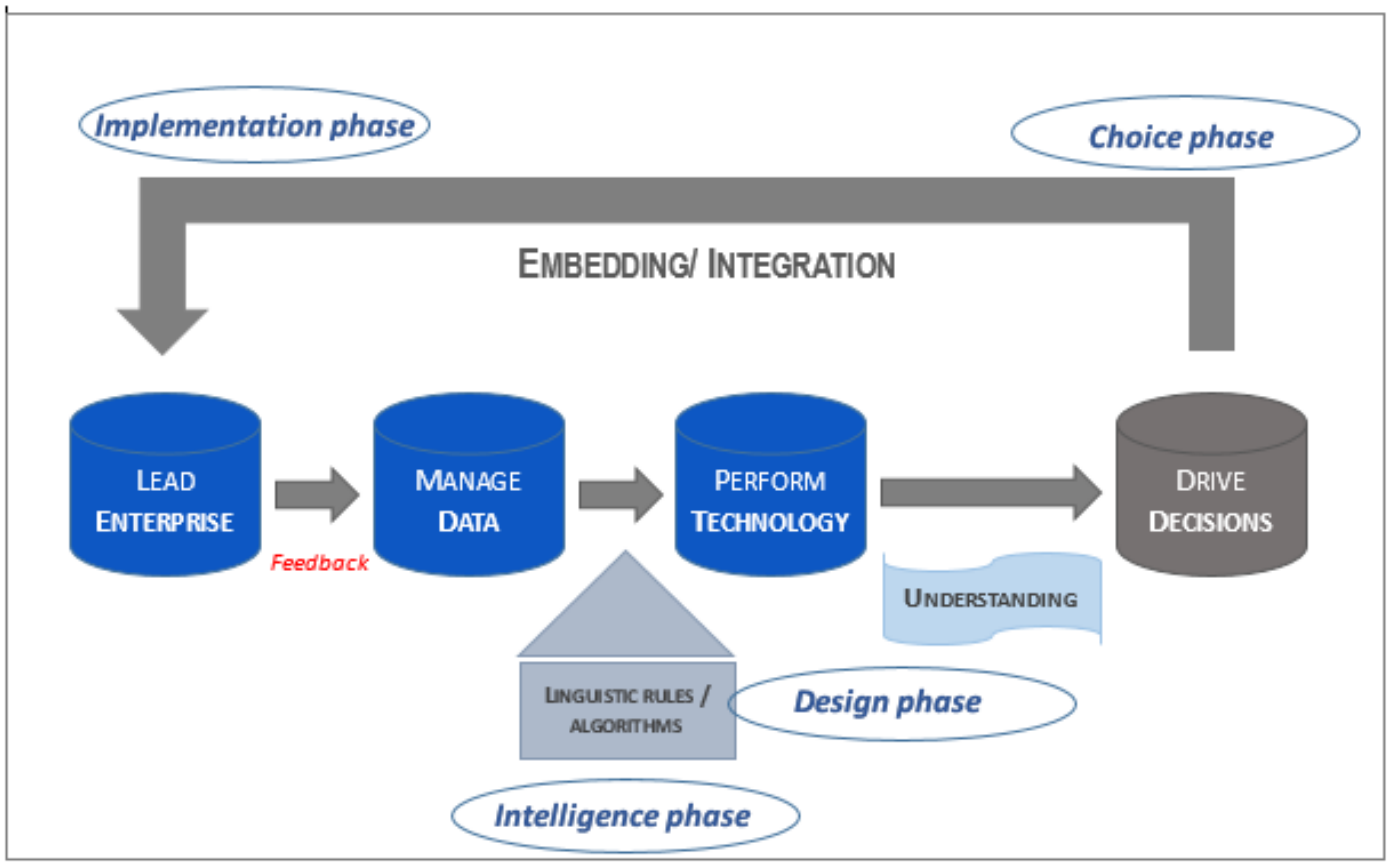

Figure 1. Decision-making model: integrating information in the enterprise

2

COMPUTATIONAL LINGUISTICS: GOALS, RESOURCES, APPLICATIONS

2.1 Introduction

Since ancient times, while always expressing its nature, communication has undergone profound changes, mainly with the advent of Internet expanding and deepening communication processes. The question is to recognize that the Web is a great revolution, almost like fire invention, and it takes time to learn how to correctly manage it, that is how to govern something that today is bigger than anything we can expect. For this reason, it is necessary to develop communication forms suitable to acting, not only to produce effective and consistent communication, but also to ensure that all this effort can improve processes of social, cultural and economic development.

The basis of human intelligent behaviour is certainly the capacity to process symbols by loading the objects of meanings not entirely recognizable in their intrinsic nature. Natural language, which is a system of communication among individuals, is certainly an example of symbolic elaboration that most characterizes the human being: through it, we can transmit information

conveyed by grammar rules application. Starting from the Second Post-War period, linguistic research have had a rapid evolution and expansion, thanks to the analytical methods introduced, such as the use of statistical or quantitative methods in the study of languages and literary works. Those studies advantaged interdisciplinary, that became a necessary transition, since the birth of Computational Linguistics (CL), i.e. the study of language mediated by machines.

CL is the first discipline that has associated Human Sciences and Computer Sciences. It deals with studying, analysing, processing natural language by computers (Grimshaw, 1986). The main goal of CL is to make the computer able to simulate the behaviour of humans in the use of language. Computational Linguists are interested in providing computational models of various kinds of linguistic phenomena. CL includes the formulation of grammatical and semantic frameworks for characterizing languages so to enabling computationally tractable implementations of syntactic and semantic analysis. It is the analysis of language with the help of the computers, allowing data processing in large amounts and in a short time. The history of CL is closely linked to computer development. In fact, it takes advantage of other disciplines as Machine Learning, Artificial Intelligence and Cognitive Science. Other disciplines are included in CL applications like Mathematics, Logics, Psycholinguistics and Engineering. This cooperation among different professionals expresses the complexity of CL field. Thus, CL is primarily the study of language and its components with the support of computer science tools and techniques.

Computers, at the service of human labour, were increasingly also used for the construction of data banks, electronic information containers, able to retrieve relevant elements starting from queries sent by users by means of a dedicated software. In particular, we quote the creation of textual databases, consisting of large collections of documents for different areas of knowledge (Pardelli & Biagioni, 2013).

The association between textual resources and the opportunity given by the Web created several large corpora: they are paradoxically brought together to go to form a single large corpus, which can be observed, investigated and studied, before being interrogated to obtain relevant information. Kilgarriff & Grefenstette affirm that "the Web is vast, free and contains hundreds of billions of words of text and can be used for all manner of language research" (2003). CL can be divided into several areas, depending upon the medium of the language processed. Some of these areas of research that are studied by CL include:

- *Computational complexity* of natural language largely modelled on automata theory, with the application of context-sensitive grammar and linearly bounded Turing machines.
- *Computational semantics*, which is focused on suitable logics for linguistic meaning representation, automatically constructing them and reasoning with them.
- Computer-aided *corpus linguistics*[5].

[5] We defined corpus linguistics in the following paragraphs.

- Design of *parsers* or *chunkers*[6] for natural languages
- Design of taggers like *POS-taggers* (part-of-speech taggers).
- *MT*, one of the earliest and most difficult applications of computational linguistics, draws on many subfields.
- Simulation and study of language evolution *in historical linguistics/glottochronology*.

During the '50s and '60s, due to the political situation at the time, these tools were used almost exclusively in Machine Translation (MT), on the translation from Russian into English. Significant resources were dedicated to this task, both in the U.S.A. and in Great Britain, for political aims. Other countries, mainly in the continental Europe, joined the enterprise, and the first systems became operational at the end of that period. However, the limited performance of these systems made it clear the difficulty in the development of these tasks, and in the subsequent years and decades, many efforts were spent on basic research in formal linguistics. Today, a number of MT systems are available commercially, although there still is no system that produces fully automatic high-quality translations without human intervention.

Another application that has attracted the attention of many companies in the last years is the analysis and synthesis of spoken language, as speech understanding and speech generation. These type of applications are used to help handicapped like blinds, in telemarketing or for several vendors in telephony. An application that will become important is the creation and management of texts by computer. Even reliable access to written texts is the

[6] *Phrase chunking* is a natural language process that separates and segments a sentence into its subconstituents, such as noun, verb, and prepositional phrases.

main block in science and business. The amount of textual information is enormous (and growing constantly), and the traditional, word-based, IR methods are getting increasingly insufficient.

As for CL and Natural Language Processing (NLP) formalization methods, in the following paragraphs, we will explore the Lexicon-Grammar framework by Maurice Gross, which is our theoretical and methodological outline. Thanks to its peculiarities, LG has found in computational linguistics an area where readapt, rework and automatically apply the language resources collected in the years of linguistic research in which CL applications, software and routines where not yet existing. Subsequently, we will cope with the original phases of this theory, then passing to the recalibration of its structure and aims, which was due to the inclusion of computers and the creation of software for automatic linguistic analysis.

2.2 About the origins: Tesnière and Harris

Since the early seventies, in an attempt to verify the applicability of Generative and Transformational Grammar by Chomsky[7], Maurice Gross begins to describe 3000 French verbs that hold a completive, followed afterwards by his collaborators. Testing the components of transformational large amounts of data, Gross and his team discover that exceptions exceed the same rules. In addition, noting the strong irregularities and idiosyncrasies of

[7]Chomsky (1957) define a new model of Generative Grammar (TGG) in which for the first time the lexicon makes its entry into the formal description of languages, alongside the reiterated primacy of syntax.

lexicon, Gross (1975) moves away from Chomsky's model, entering into open conflict with it. The conventional notation for the presentation of grammatical information is as simple and transparent as possible, in line with the intention of Zellig Harris, whose theory is strongly oriented toward the surface analysis of directly observable linguistic data. In this, it differs from Generative Grammar, avoiding the usual recourse to abstract structures such as the deep structure.

LG represents the theoretical and methodological framework adopted in this paper, and this is why we consider it essential to describe the cardinal principles of our research, before presenting their results. Maurice Gross succeeded in showing the limits of the rules-based approach to language, being such an approach an almost insurmountable obstacle to an exhaustive representation of many phenomena of language. To understand better the context in which Maurice Gross' theories developed, we will review the research studies of Lucien Tesnière firstly (1953, 1959) and then examine the influence of Zellig Sabbetai Harris (1954, 1981, 1991) on LG.

Tesnière was a prominent and influential French linguist. His main contribution to linguistics is the systematisation of specific syntactic characteristics common to almost all natural languages, and widely known as *dependency grammar*. He presented a fancy formalization of syntactic structures, providing a primary distinction between arguments (actants) and adjuncts (circonstants).

"Every word in a sentence is not isolated as it is in the dictionary. The mind perceives connections between a word and its neighbours. The totality of

these connections forms the scaffold of the sentence"[8]. The connections between the elements in a sentence, as described by Tesnière, represent the relations of dependency that connote the different hierarchical levels of words (governor or subordinates) and contribute to create a concrete syntactic structure. Tesnière also contributed to understand the nature of the lexicon. He compared verbs to molecules. As an oxygen atom O attracts two hydrogen atoms H to create an H2O molecule, verbs attract actants to create clauses: verbs are central inside sentences, and each of them has a specific valency. The simple sentence involves a central verb from which some extensions depend, called actants, while the peripherals expansion identify the circumstances. Tesnière distinguished between verbs that are avalent (no actant), monovalent (one actant), divalent (two actants), and trivalent (three actants). The features of valency allow identifying and understanding the syntactic structure of a sentence. Moreover, even not being directly interested in separating syntax from semantics, Tesnière (1988) affirms that given a sentence, its syntactic structure also directly reflects its semantic one. According to Radimský (2012), we may affirm that Tesnière enhanced the role of syntax out of other levels of analysis, implicitly providing the impetus to consider a second level of analysis as could be the semantic one. Although seemingly complementary, reasoning about levels of analysis leads us to recognize a difference between Tesnière's and Gross' studies , since the lexical-grammatical structure, as we shall see, defines the syntax and semantics of the simple sentence, exceeding the central role given to syntax by Tesnière.

[8] The passage cited here is taken from the first page of the *Éléments de syntaxe structurale* (1959).

Zellig Harris' theory represents a solid basis for our research program, leading Maurice Gross through the principles of distribution and transformation, referring to the analysis of the sentence structure and discourse analysis, and the principles of grammar in Operators and Arguments. Thus, the classification made by Harris on operators is a necessary starting point for LG studies, since it is based on the recognition of the essential role played by elementary sentences within a linguistic system. The concept of distribution by Harris consists in the fact that an element in a sentence is constituted by the sum of all its possible contexts, namely by all those members that can freely co-occur with it in the same position, regardless of the greater or lesser probability of co-occurrence. In such sense, starting from the notion of morpheme and from the method of commutation or equivalence between different morphemes (Bloomfield, 1933), transformational rules can notice mutual relationship between simple sentences (as active/passive or positive/interrogative) having different structure but similar meanings (Harris, 1964). In this way, Harris shows that starting from words combinatorial predispositions it is possible to define specific word subclasses having similar semantic features. Entire sequences come into correspondence, in reciprocal transformation. On such basis, Harris also identifies the existence of elementary (or nuclear) structures of sentences consisting of Operators (verbs) and Arguments (complements).

Actually, the study of operators may be achieved only within the special relations that bind them to their arguments, whether these are simple nouns (i.e. elementary subjects) or complements (i.e. non-elementary arguments, propositions or speeches). The Grammar of Operators and

Arguments is an important theoretical and methodological reference useful to better understand the lexical-grammatical perspective and incorporate it inside NLP tools. Harris developed this framework in the second half of the seventies, delineating it as a consistent continuation to his distributional and transformational studies. According to Harris (1952), we could develop a formalized procedure for the analysis of the speech: this method can be applied directly to a text without the need of any linguistic knowledge. The aim is to collocate inside a single class all the elements that have the same distributional properties, or that appear in the same linguistic contexts, so being able to discuss about the distribution of the whole class and not about the one of its single elements. From a distributional point of view, are equivalent not only the linguistic elements that appear in identical contexts, but also those that appear in equivalent contexts[9]. For the theory of Arguments and Operators, every speech can be interpreted as a concatenation of sequences, describable according to the relationship between a central element (Operator), and peripheral elements (Arguments). According to Elia et al. (2011), in LG framework, Gross gives a concrete form to these methodological bases, classifying verbal predicates on distributional and transformational similarity, and identifying sets of verbs with same formal and semantic features.

[9] Harris (1952) argues that two elements (morphemes or sequences of morphemes) are equivalent if they appear in the same contexts, or in equivalent contexts. Each set of mutually equivalent elements is said equivalence class. So, according to Harris, each sentence of a text can be represented as a sequence of equivalence classes.

2.3 Maurice Gross and Lexicon Grammar Framework

Maurice Gross was a French linguist and a pioneer in NLP. His theories are very usefully applied in CL theoretical and practical investigation fields. His major contributions are the description of the idiosyncratic properties of lexicon and represent a fundamental part of syntax and semantics descriptions of natural language. The summa of his works and research, i.e. Lexicon-Grammar theory, represents an empirical approach, which aims to obtain a recording of linguistic data starting from the observation of all linguistic combinatorial phenomena. Maurice Gross advocates a subjective method with a collective control: a team of native-speakers linguists performs observations. This means that LG does not apply a hypothetical reasoning, but it achieves the empirical observations of linguistic acts evaluated in their concrete contexts of production and usage. LG methodology is inspired by experimental sciences, and places emphasis on the collection of linguistic facts, subsequently comparing them with the reality of language uses, in quantitative and qualitative terms. Quantitatively, LG involves a systematic description of lexicon combinatorial features. Qualitatively, LG adopts methodological precautions to ensure good reproducibility of the observations taken, and in particular, to address the risks to the examples built. The first precaution is to consider the elementary sentence as the minimum semantic unit to analyse. Actually, a word acquires a certain meaning and function only in a specific context; this means that when inserting a word in a sentence, one has the advantage of manipulating a sequence that can be found to be semantically acceptable or unacceptable. This is the necessary condition for establishing that

the syntactic-semantic properties are defined with a degree of precision that makes it possible to relate them to a lexicon, considered in its totality. In this framework, based on the transformational theory Z.S. Harris (1976), the basic meaning unit is analysed in terms of operators and arguments, in the following format:

Operator (argument 1, argument 2, argument 3)

Operator is the central element and arguments turn around it, being this mechanism essential for any grammatical structure.

In the second half of the ‘70s, Maurice Gross starts to apply LG theoretical-methodological framework for NLP to French. Subsequently, a same application for Italian starts with the studies of Annibale Elia, Maurizio Martinelli and Emilio D’Agostino (Elia, Martinelli & D’Agostino, 1981). Today, LG stands as one of the most profitable and consistent methods for natural language formal description, and for several languages. For LG, every word may encapsulate inside the structure of an elementary or simple sentence, which represents the best environment for the study of semantic and syntactic uses. Therefore, LG presents a formalized grammar and a lexicon-syntactic dictionary. The basic assumption is that lexicon items are not severable from the grammar features that each of them brings into a sentence context, while mutually combining on the base of their properties and according to the specific rules of co-occurrence and selection restriction.

One primary need of LG is formalization. The results of linguistic description should be sufficiently formal to allow the verification by

comparison with actual uses. A specific LG application is the automatic processing of languages, to achieve particularly through the creation of computer parsers. The need of formalization manifests itself through the adoption of a discretized model of syntax and semantics. In this way, acceptability is modelled by means of binary properties: for the needs of the description, a phrase is considered acceptable or unacceptable, in the same way and for the same reasons theorized by generative grammar, but from the point of view of concrete linguistic use. An efficient automatic textual analysis, presumes to lemmatize all these different expressions, not only simple or compound words, but each lexical form having a meaning, increasingly reducing the gap created by unknown words, updating the dictionaries to use for text analysis. According to Elia et al. (2010) the tools necessary for an efficient system of analysis are:

- Electronic and morph-syntactic dictionaries of simple and compound words;
- Lexicon-Grammar tables of verbs, predicative nouns, adjectives and adverbs;
- Local grammar finite state automata (inflectional, morphological, and syntactic grammars).

The construction of the linguistic resources implemented by Gross follows four steps:

- Large coverage of elements of a language;
- Formal representation allowing a computational scouting of natural language;
- Parallel models for all different languages;

- Accurate information, relying on a manual analysis of attested data, trained by linguists, and not coming from rules or corpora.

As for linguistic analysis, this method provides a high quality, especially thanks to the constant manual construction and monitoring (Laporte, 2005) of its resources. Besides, Maurice Gross did not neglect the formal aspect of natural language, adopting the approach of Zellig Harris as for the concepts of distribution and transformation[10]. LG methodology defines the mechanisms of language formalization through the description of elements occurring inside sentence contexts. Nevertheless, LG originates from mathematical models[11] of language (Gross, 1972; Harris, 1991), which implies that the description of linguistic elements is not based on statistic rules and algorithms, but on the analysis of words co-occurrence, distribution and selection restriction observed inside simple sentences by means of predicates syntactic-semantic properties (Gross, 1964).

From a formal, lexical and morph-syntactic point of view, LG separates simple words from compound ones. Actually, in LG simple words are alphabetic or alphanumeric sequences written without any interruption, while compound ones are all sequences of two or more simple words, separated by blank spaces or non-alphanumeric characters, and participating as such in the creation of a unique global meaning. In this sense, both simple and compound words are lemmatized inside dictionaries as Atomic Linguistic Units (ALU).

[10] In Harris, the idea of transformations relates observable sentence forms, and differently from Chomsky, for whom transformations change deep structures into superficial ones.

[11] It is evident a strong correlation between the works of Zellig Harris and those of Maurice Gross. He started from the Harris's theory of syntax, which includes a good deal of semantics, within a well-specific description.

Examples of simple words are prepositions as *per* (for), conjugations as *e* (and), or nouns as *tavolo* (table) or *elefanti* (elephants), in which there is only one lexical morpheme plus all the needed grammatical/inflection/compositional morphemes. Examples of compound words are nouns as *carta di credito* (credit card), functional sequences as "a cavallo di", used as a preposition in the expression *essere a cavallo di una motocicletta* (be riding on a motorcycle).

Thanks to this separation, Gross provides an analytical depth that allows it to go off simple words and free word groups, individualizing and formalizing non-casual agglomerations of words, in which items are related to their arguments by different degrees of variability and cohesion, based on semantic coherence. About this varied and contiguous typology, we can say that a small internal cohesion of a group of words corresponds to a high level of semantic compositionality and to a low level of idiomaticity. On the contrary, a group of words having great internal cohesion presents a low level of semantic compositionality and a high level of idiomaticity.

Today, teams of researchers in different universities, laboratories, and global research facilities apply methodological framework of LG; they are identified with the scientific network that goes by the name of RELEX, which develop language resources and automatic text analysis software. Thanks to this approach, LG allows the international community of linguists to get a complete, empirical and exhaustive description of natural languages by means of a large data set consisting of tables of syntactic and semantic properties of

thousands of lexical entries Among lingwares developed on the basis of LG framework we may mention INNTEX[12], CATALOGA[13] and NooJ[14]

2.4 Semantics and LG approach

Semantics is the branch of linguistics devoted to the investigation of linguistic meaning, and to the interpretation of expressions in a language system (Chierchia & McConnell-Ginet, 2000). Semantics is the linguistic and philosophical study of meaning in natural language, programming languages, formal logics, and semiotics. It focuses on the relationship between a signifier, (like words, phrases, signs, and symbols) and a signified, i.e what a signifier stands for, or its substance. In linguistics, we use the word semantics to intend the interpretation of signs or symbols used by societies within particular communication circumstances and contexts. We could consider different approach to semantics: not all theories developed in semantic studies can be applied to automatic linguistic analysis. Semantic representation and

[12] INTEX is a corpus processing system, based on automata-oriented technology. Its concept was born at the LADL (Laboratoire d'Automatique Documentaire et Linguistique) and developped by Max Silberztein.

[13] Cataloga is an automatic textual analysis software built by Annibale Elia, Alberto Postiglione and Mario Monteleone at the Department of Political, Social and communication Sciences of the University of Salerno (Italy). The automatic textual analysis carried out can determine the semantic content of terminological digitalised texts without any human contribution. The four steps of analysis are automatic reading of a text; computing of all terminological compound words found; complete automatic word indexing, statistics-based listening of all terminological occurrences based on semantic fields.

[14] We will discuss about NooJ in the fourth chapter, with reference to our analysis environment.

interpretation[27] follow a prevalent philosophical, semiological and semiotical line to cope with meaning. In our analysis, we overlooked this approach to semantics, choosing a structuralist one.

From the point of view of LG framework, a set of lexical-syntactic structures defines the value of semantic predicates, while the arguments selected by each semantic predicate are given the value of actants, subjects included (Elia et al., 2010). The features of each verbs are expressed by the application of the rules of co-occurrence and selection restriction, through which verbs select semantically their arguments to construct acceptable simple sentences. According to Vietri and Monteleone (2014), we have semantic predicates expressing the intuitive notion of "exchange" (*Transfer Predicates*), "motion" (*Movement Predicates*) or production (*Creation Predicates*). Each set of semantic predicates assume those arguments whit which they have compatible semantic roles. *Transfer Predicates* have a "giver", an "object to transfer" and a "receiver", as in the sentences:

[27] As for representation, semantics has an autonomous level related to deep structure: we can consider such representation as the formal language in which meaning are described. Opinions differ about whether semantic representation is sufficient or necessary, about its form and about how it relates to syntactic representations. Psychologists affirmed that a mental semantic representation is necessary to account for the fact that language users grasp meanings. Denotational theories of meaning, on the other hand, claim that meaning can only be explicated in terms of denotations in the world. Semantic representation can take the form of a structure of semantic features (Katz & Fodor, 1963; Jackendoff, 1983) or formulas of a logical system. Generative semantics identified it with syntactic deep structures. Finally, we can consider the pragmatic theory, which identifies the meaning of an expression with the use that is made of it by the participants in an interaction. This theory is often named "meaning-is-use theory", after Wittgenstein (1953). It is characteristic for those theories in which speech acts play a central role, following Austin (1962).

1. Mario (giver) gives a cake (object to transfer) to Juliet (receiver)
2. Juliet (receiver) receives a cake (object to transfer) from Mario (giver)

Movement Predicates select an "agent of motion", an "object to move" and a "locative name", as in the following:

1. Mario (agent of motion) goes to Paris (locative name)
2. Mario (agent of motion) moves the books (object to move) from his house to the office (locative name)

Creation Predicates, finally assume a "creator" and a "creation", i.e. an item that did not exist at the beginning of the sentence, and exists at the end of it:

1. Mario (creator) write a novel (creation)
2. Juliet (creator) compose a song (creation)

In this way, entries belonging to electronic dictionaries can be classified presuming their similarity and proximity to semantic predicates, or better inferring on their likeliness of being selected by a specific set of semantic predicates. Even if the list of semantic tags is not simply identifiable, due to simple word polysemy, grammars can be built with NooJ for each single set of semantic predicates. As for Italian LG, descriptions assign correlated predicates and arguments by means of electronic dictionaries. It is also possible

to build grammars that annotate all specific semantic predicates. Textual analysis based on simple words is not sufficient to retrieve all semantic elements inside texts; it is necessary to consider also compound forms, which have a specific meaning. This specifically happens in the case of terminology or specialty languages[28], which largely use compound words. First of all, terminology is the ordered set of specialized terms related to one or more fields of science and technology. In addition, terminology refers to scientific disciplines through a work of acquisition, definition, description and dissemination of concepts belonging to specialist knowledge, aiming at an optimal transfer of knowledge in one or more languages. Linguistic and communication function of terminological compound words are always monosemic (Elia et al., 2010). Thus, terminology is not ambiguous, mainly due to the fact that each terminological compound word is closely linked to a specific concept and/or object. Using electronic dictionaries it is possible classify compound words with reference to each specific domain sector , which they belong to.

[28] A specialty language is a diaphasic a variety of language (words, expressions, technical terms, etc.) used by a "minority of experts" of a given material or the workplace in order to make it clear, fast, accurate and effective communication and collaboration among members of the group. In this sense, the technical language has affinities with professional jargons and trade, which represents an evolution, even if it is distinguished by the exact and in some cases (i.e. the language of mathematics or physics) for the explicit formalization. Therefore, specialty languages are professional jargons, the languages of various academic disciplines, scientific and technical, and all community languages that share some knowledge or some specific activity (Riediger, 2012). Some of these are consolidated, such as the languages of natural science, while others are rapidly converted, for example, the languages of the various technologies.

2.5 Parsing and Syntax

Parsing refers to the process of analysing string of words written in natural language, in order to define if they conform to well-defined grammar rules. Parsing is often performed to understand the meaning of a sentence or a word, sometimes with the assistance of tools such as sentence diagrams. It usually emphasizes the importance of formal grammatical divisions such as subject and predicate.

In Computational Linguistics, the term parsing identifies the analysis of a sentence or string of words by computer, specifically into its components and the relation between them. As well, a *syntactic parsing*, has the task of recognizing a sentence (or a constituent) and assigning a syntactic structure to it. Also, parsing is the process of assigning structural descriptions to word sequences produced from natural languages (De Bueriis et al., 2005), as in the following example (Figure 2):

INPUT. “Google is located in California.”

OUTPUT.

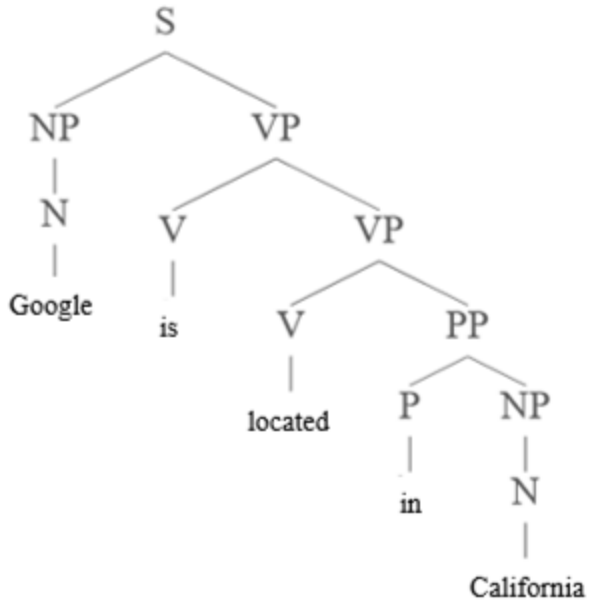

Figure 2. Example of syntactic parsing

With specific reference to parsing activities, the time and procedures of the structural description to be assigned to a language depend on the grammars according to which a parser attempts to analyse the sequence of symbols presented. In other words, a parser takes a sequence of words from natural language and an abstract description of the possible structural relationships, which may exist between words or sequences of words. In basic, we will have

zero input descriptions, if the grammar cannot be represented and recognized, or we will have one or more input descriptions if such an input is ambiguous, or if it has more than one possible correct structural description. In this sense, parsing can contribute to the identification of the meanings of word sequences. In fact, it is believed that grammatical structures contribute so much to meaning creation that today parsing operations go from the study and isolation of a simple sentence until the complete semantic analysis of a corpus for information extraction.

2.6 Acquiring Knowledge: from text to corpora

The word *textus* has Latin origins, as a past participle of *texere*, a verb formed from the ancient root *Teky* indicating the work of loggers and carpenters. With this meaning it occurred in Indo-Iranian areas, Greek (Tekton, "carpenter"), Slavic, Germanic and Celtic (Devoto, 1979). Tiles, therefore, have to be understood as webs of weaved wires, canvas, as well as the woodblocks by a carpenter. The verb to tile is used also figuratively, to indicate the action with which "to weave a plot or a deception." From here, the intertwining and then the complexity in the language of discourse (Segre, 1981), as it appears in the Institutio oratori (IX, 4.13) of Quintilian. In fact, it is precisely in the Latin of the Christian era that Segre sees the triumph of writing.

Thenceforth the concept of text has assumed several roles in social communication, with particularly further processing today in digital

communication. The context is no longer the world of life crystallized in a system of knowledge that allows reconstruction. In digital communication, the context is represented by all other texts, which are related to the first one as in a network. In digital communication, social experience is reticular: centreless, multiple and diversified. Digital society is expressed through hypertext[29]. In fact, on the Web, with the increase in size of textual data likely to be analysed comes the need of a technology capable to cope with such an immense variety of linguistic accomplishments. Moreover, mark-up languages like XML and the increasing mutual exchange among similar types of texts has allowed explaining the text structures. In fact, in CL, we cannot talk about singular texts, but we have to refer to the concept of corpus. A corpus is a collection of texts that have been selected to be functional for a specific linguistic analysis. Also, to use a more suitable definition, a corpus is a systematic collection of naturally occurring texts (Nesselhauf, 2005). In this definition, the adjective systematic indicates that the structure and contents of the corpus follow specific extra-linguistic principles (i.e. varieties of languages, concepts, words). The texts or documents, their sections or simple sentences, are generated either from written texts or from transcriptions of oral speeches.

According to Barbera (2013) a corpus is also collection of texts in electronic format treated in a uniform way (i.e. tokenized and tagged with a suitable mark-up language) so as to be manageable and queried by a computer. If languages are the target (i.e. the description of natural languages or their varieties), words are mostly chosen to be authentic and representative. Thanks

[29] Set of text documents connected by links. In fact, while a traditional text is read linearly, a hypertext allows readers to read by associations of concepts.

to the growing interest in CL, and to the consciousness of how important linguistic data are to their research, corpora are now the main data source for this discipline. The evolution of computers has also played a key role, because a computer allows to store increasing amounts of texts and explore them more quickly and effectively.

Corpora can be classified into various types, depending on the criteria used to select the texts that compose them:

- *general corpora*, the texts of which are selected transversely to from different varieties of a language and then explored as a whole;
- *specialized corpora*, designed to study a specific field of language (Medicine, Bureaucracy etc.);
- *corpora of written language, spoken language*, or mixed;
- *monolingual or multilingual corpora*;
- *synchronic corpora*, i.e. composed by texts which all belong to the same historic period;
- *diachronic corpora*, including texts from different historic periods;
- *annotated and not annotated corpora;* annotated corpora, more and more popular today, include information on the linguistic structure of texts, at various levels (syntactic, semantic etc.).

In order to extract general indications on a given language, it is necessary for the corpora's body to be representative of that language (or of a language sector) at a particular time, or to represent a scale model of the

language under investigation. Corpora observations represent the main activity of Corpus Linguistics, which is a sub-discipline of Linguistics. Corpus analysis allows deriving sets of abstract rules from texts; such rules are hence used to govern a natural language, or to study how to compare such language to another language. Before CL birth, these rules were manually derived from corpora; now they are automatically derived from the same corpora, with the help of computers.

Corpus linguistics suggests that the analysis of a reliable language is more feasible if corpora are collected in a specific field that is in their natural settings. For example, today the Web provides many texts in free form. When Semantic Web was firstly introduced, the important proposal[30] had also born to use the Web as a corpus.

Actually, the fact of coming to an exploration of web resources was historically predictable. In recent years, the quantitative insufficiency of traditional databases to face increasingly complex specific language problems, the rapid aging of materials (considering the continued evolution of language), the new technologies and new means of communication, could only consider the entire Web as a kind of mega-corpus from which to extract information. The proposal, however, clashes with the problem of finiteness (Barbera, 2013). In the next table, we represent corpora classification as established by Bolasco (2012).

[30] For further, see Kilgarriff (2001).

Classification	Tokens	Dimension
Little corpus	15.000	100 KB
Medium corpus	45.000	300 KB
Medium-big corpus	100.000	700 KB
Big corpus	= o > 500.000	3 Megabytes

Table 3. Classification of corpora dimension by Bolasco

This type of classification is more important in statistical automatic linguistic analysis, because in it, the number of occurrences influences results. The fragment of texts may have a very variable length: whether it is an entire document depends on the kind and type of materials gathered in the collection. If each fragment is a record of information attributable to a different speaker, as the answer to a question in a sample of respondents, or as a message of a corporate database, it will be of very limited amplitude (short text).

Corpus linguistics has generated a number of research methods, trying to trace a route from data to theory. Wallis and Nelson (2001) introduced three perspectives of corpora: Annotation, Abstraction and Analysis.

- *Annotation* consists of the application of a scheme to a corpus. Annotations may include structural mark-up, POS tagging, parsing, and numerous other representations.
- *Abstraction* consists of the translation (mapping) of terms in the scheme to terms in a theoretically motivated model or dataset. Abstraction typically includes linguist-directed search but may include e.g., rule-learning for parsers.
- *Analysis* consists of statistically probing, manipulating and generalising from the dataset. Analysis might include statistical evaluations, optimisation of rule-bases or knowledge discovery methods.

In the decision-making process, acquisition of new knowledge is fundamental. The knowledge available is often expressed in textual documents. The quality of decisions depends on the quality of information submitted to decision-makers. To facilitate the acquisition of knowledge, it is expected that open texts are organized and evaluated. The properties and characteristics of single texts must be recognized, starting from the process of annotation.

2.7 Annotating linguistics corpora

Annotating a text means to add information to it in form of labels. This typically manual activity has become an automatic procedure thanks to the emergence of technologies linked to TM. The concept of semantic annotation

is closely related to classification, especially when we refer to the terms contained in a document. Classification and semantic annotation are operations that can be performed on texts, both automatically and manually, using a base of reference data. Before delving into this topic, it is useful to summarize some basic concepts about classification and text annotation, which have specific characteristics and problems.

As for the classification of a text, we intend its assignment to one or more existing classes among given sets of distinct classes, each one of these having its own profile, and described by a number of features. A classification process determines which class a text document belongs to, or which class best describes its characteristics. The criterion for selecting the relevant characteristics to classify a text is crucial and is determined a priori by linguists. For example, we may be interested in the length of documents and therefore classify them according to the number of words they contain (statistical analysis), or to the date in which they were produced, or to the language components (morphology, syntax, vocabulary). This kind of classification cannot be strictly defined semantic: annotated corpora are texts in which linguistic information is coded in association with the same texts. The clarification in the coding levels of information, such as syntactic structure and the semantic roles of a sentence, makes these levels accessible to a computer, an aspect that leads to the importance of linguistic annotation in CL today. As for information representation, each part of the language description levels poses specific problems:

- The *morphological* description presumes a lemmatization (adding quotation marks to words plus their lemmata) of each token of the text; also, to each word, it assigns the respective grammatical category;

- In the *syntactic* description, it is necessary to parse explicitly the sentences of a given text; parsing can be achieved in different ways, depending on the theoretical approaches adopted. For instance, the representation of constituents describes a sentence in terms of dependencies between words, indicating grammatical relations (subject, object, etc.);

- The *pragmatic* description can affect various phenomena related to the communicative function of an utterance, or the relations between the linguistic elements, which go beyond the single phrase. For example, in corpora containing transcriptions of spoken dialogues, it is useful to identify the illocutionary function of utterances (defined as the type of action we take to issue a special statement: question, demand, order, etc.). We may also need to highlight anaphora and cataphora deixis, i.e. phenomena in which the correct interpretation of an element depends on the references it creates inside a linguistic context, respectively backward and forward;

- In *semantic* description, it is required the explicit coding of the meaning of the linguistic expressions in a given text. It is possible to classify lexical words according to a predefined set of conceptual

categories, in a way to capture the most important traits of meaning (person, place, process); also, it is possible to mark semantic roles, describing the semantic function performed by a phrase in the event expressed by the verb.

As for knowledge contained in corpora, it is representable by means of different types of schemes found in three basic types of information. These could be individual or combined, and constitute a backbone for any further scheme:

1) *Categorical information*, expressed as labels that associate categories to units of a given text. The records of the speech, the lemma, or even the semantic roles, are typically presented in the form of categorical information. The most intuitive way to represent this kind of information on XML is the one that makes use of associated attributes to the reference element;
2) *Structural information* concerns the identification of structural units in a given text and their organization into hierarchical structures. The syntactic level is the one most closely related to this type of information; the hierarchical relationship between the constituents are represented in XML, with the inclusion of smaller elements into larger ones;
3) *Relational information* interconnects different language units, thus allowing the possibility to account for mutual relations (dependencies between subject and object, or between pronoun and anaphoric pronoun antecedents).

Of course, it is possible to use multiple types of information for a same level of description. Indeed, it is rare that a level can be described with only one type of information. The annotation schemes tend to intermingle, so that hardly any of these types is present as a "pure" state in the corpus. On the other hand, identifying how they are encoded, and evaluating the descriptive efficacy in which they are expressed, represent fundamental steps for checking the validity of all annotation schemes.

As mentioned, annotation is considered semantic when items of interest for a classification refer to the meaning inside a document. Categorizing a document for its main argument (Text Classification, CAT tools), is an example of semantic classification, but it is equally necessary to divide documents on the base of their main topic, not treating them according to the attitude or opinion of their authors. Classifying documents on the base of the occurring frequency of a particular word in a text requires a necessary association of each word to a numeric value, i.e. to a string that can be more or less ineffective for the purposes of categorization. Semantic assignment instead introduces a more complex issue, which is that of ambiguity. The study of the ambiguity of a word allows the reaching of a more precise level of description in textual analysis. For semantics, ambiguity occurs when a same word covers more than one of possible meanings. Some of these meanings, generally the denotative ones, are partly found in dictionaries, inside the entry relating to the lemma of a given word. Other meanings, mainly the connotative ones, are marked by the context (environment) in which the statement is made, and by the co-text (verbal) in which a given word is located. Is it not superfluous to recall once again that connotative meanings are extremely

unstable, while the denotative meanings of a word in a natural code always differ from those of any other word, even if belonging to the same natural code, or to to another natural one. For example, we can examine the second type of ambiguity in this example:

a. *La matrice a scalini presenta tre **pivot**, contenenti le cifre 3, -1 e 5* (the stepped matrix has three pivot, containing values 3, -1 and 5).

b. *Il **pivot** potrebbe rimanere in panchina per un infortunio* (the pivot could remain on the bench due to injury).

c. *L'ultimo cartone animato l'ho creato con **Pivot*** (I have made the last cartoon with Pivot).

d. *Nella startup lean, il numero di **pivot** può essere elevato* (in the lean startup, the number of pivot could be high[31]).

In order to determine the probability that a certain sentence structure or a certain semantic value are used within a given speech community, it is necessary to resort to textual corpora, as we have seen above, and which contain millions and millions of real sentences, (i.e. spoken or written). All these considerations explain why the automatic translation processes, i.e. computers that translate without human intervention, have still a scarce

[31] A pivot, in linear algebra and computer science, is a concept related to matrices; also, pivot is one of the roles played by football, basketball and handball players. Pivot is also a program to make cartoons; finally, in lean startups, it identifies a correction in the development strategy.

success. In the disambiguation of a sentence, our brain uses not only its grammatical and lexical knowledge, but also statistics, or better the frequency with which certain lexical and grammatical structures recur in our experience. However, since large textual corpora are much more roomy and reliable than our brain, which is currently the best tool possible, a textual analysis work is obtained by combining the flexible human intelligence with the consultation of existing corpora. Textual corpora, like all other necessary tools for natural language processing, will be a major topic in the new technologies' development.

2.8 Text Mining

Linguistics provides methods to understand texts: the elements of a sentence are recognized, processed and placed in relation to each other. Precisely this understanding of text allows for in-depth, automated analysis of extensive document sources. The associated extraction of data from text is described as TM. According to Witten (2005) TM is a "burgeoning new field that attempts to glean meaningful information from natural language texts" (p.1). It characterized the process of analysing text to extract information that is useful for particular purposes. In fact, Kopackova, Komarkova and Sedlak (2008) noticed how TM can be used for pre-processing of textual information, in order to find hidden knowledge and ease the process of decision-making. In each case, the types of document need the context of use and accommodates the cost of producing and processing information. TM routines include broad umbrella

terms describing a range of technologies for analysing and processing semi-structured and unstructured text data (Miner *et al*., 2012). In order to understand the technological context in which TM was developed, we have to consider the challenge posed by a great volume of textual information stored in various forms and generated on the Web. TM has its roots in Linguistics but in recent years, it fulfills a pregnant role in Analytics market, and also for analysis interfaces of term-extractions. Thus, other disciplines are involved in TM, as Computational Linguistics, Statistics, Machine Learning, and Artificial Intelligence, none of these working separately, but all cooperating to produce efficient analysis. It could be different approach in TM as shown the following classification by Miner:

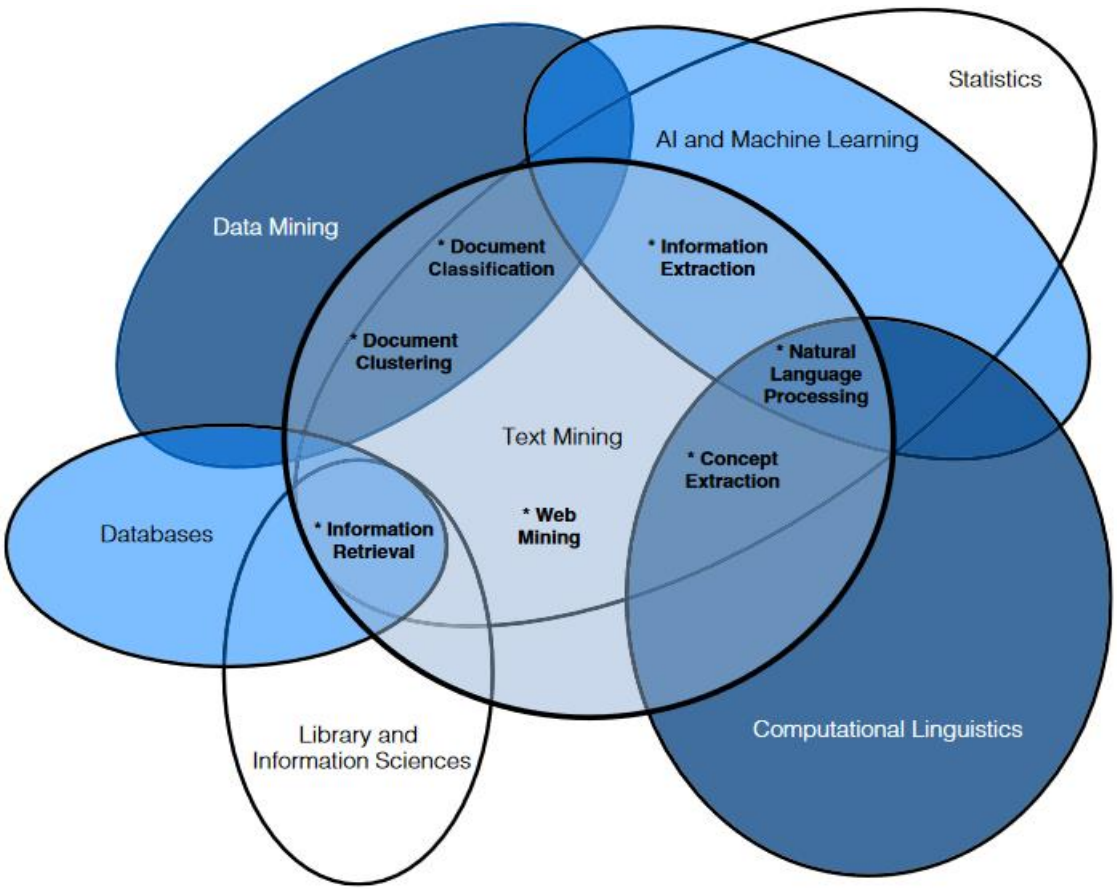

Figure 3. Schema readapted to TM approaches by Miner

If we exceed the theoretical frame in which we are producing our observations on TM, we can identify seven practise areas, each one having different characteristics, but strongly interrelated between them. These guide practitioners are useful to point out that, in text analytics, we need several types of support, and that many projects of document analysis require techniques coming from different areas.

Miner (2012) identifies the approaches and techniques used in the main different applications of TM, which are: IR, Document Clustering, Document Classification, Web Mining, IE, Natural Language Processing, and Concept Extraction.

Information Retrieval. It includes a set of techniques used to manage the representation, storage, organization and access to objects that contain information such as documents, web pages, online catalogues and multimedia objects.

Document Clustering. It is the task of grouping a set of documents in such a way that objects in the same group, called a cluster, are more similar to each other than to those in other clusters.

Document Classification. It is used to assign a document to one or more classes or categories. This may be done manually or algorithmically. The manual or intellectual classification is an activity of library science, while the algorithmic classification of documents is mainly used in information science and computer science.

Web mining. It allows looking for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines. Structure mining is used to examine

data related to the structure of a particular Web site, and usage mining is used to examine data related to a particular user's browser.

Information Extraction. Extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases, this activity concerns processing natural human language texts. A broad goal of information extraction is to allow computation to be done on previously unstructured data. Structured data are semantically well-defined data from a chosen target domain, interpreted with respect to a specific category and context.

Natural Language Processing. It is the automatic process of using an electronic computer to "read" information written or spoken in a natural language. This process is particularly difficult and complex, due to the ambiguity of human language. For this reason, the process is divided into different phases of analysis: lexical, grammatical, morpho-syntactical and semantic. In the semantic analysis, disambiguation is the automatic procedure that matches the form of written or spoken language to all its acceptable and grammatical contents, i.e. to all its possible meanings.

Concept Extraction. This procedure identifies concepts or artefacts exposed inside in documents. Due to the fact that artefacts are typically inaccurately structured sequences of words and other symbols (rather than concepts), concept extraction presents non-trivial problems, but it can provide powerful insights into the meaning, provenance and similarity of documents. The conversion of words to concepts can be performed using thesauruses, which are either specially created for the task, or based on pre-existing language models.

Also, Miner identifies the main topics for each of these applications, as shown in Table 4:

Application	Topics
IR	Keyword search
	Inverted index
	Question answering
Document Clustering	Document Clustering
	Document similarity
Document Classification	Feature selection
	Sentiment Analysis
	Dimensionality reduction
	eDiscovery
Web mining	Web crawling
	Link analytics
	Sentiment Analysis

IE	Entity extraction
	Link extraction
Natural Language Processing	Part of speech tagging
	Tokenization
	Question answering
Concept extraction	Topic modelling
	Synonym identification

Table 4. Application topics by Miner

TM has changed significantly with the merging of new applications and techniques of analysis. These developments allow defining new tasks in Information Retrieval and Extraction. For instance, tokenization, used also to split texts into named entities; morphological and lexical analysis modules, which assign tags to the terms and disambiguates words and phrases.

Despite the importance of statistical approach in TM, retrieved data through keyword search and quantitative evaluation do not seem apt to fulfil all main TM aims, routines, and procedures. Early approaches of TM would treat a text source as a set of words. However, definitions on words, either formally or semantically, have been evolving in more complex forms, as shown by Multi-Word Atomic Linguistic Units (MWALUs), which have been defined starting from basic linguistic forms. Basic lexical and statistical analysis might be used to count frequencies of words, in order to classify

documents by topic, but from our point of view, to build an efficient semantic TM system, it is necessary to create and adopt the following linguistic resources:

- electronic morph-syntactic dictionaries of simple and compound words;
- LG tables of verbs, predicative nouns, adjective and adverbs;
- Local grammars in the form of FSA/FST (inflectional, morphological, and syntactic grammars).

Thus, a change must concern the greater propensity to data discovery, especially as for information discovery and extraction. Through semantic TM systems, we will be able to extract patterns, resources and opportunities from both old and new data sources. In addition, the introduction of deep linguistic analysis of business documentation in DSS will allow reducing the time spent on the measurement and analysis of all given of the project. Also, it will allow better possibilities of massive control on documentation; the non-dispersion of information related to a given company; improving the communication of company documents, while monitoring the quality and adequacy of activities. Nowadays, most computational approaches deal with semantic TM tools, but applying statistical rules, strings of words with a single meaning (as MWALUs) are not always recognised, which induces powerful information loss. Furthermore, languages are in continuous development, both in their grammatical and lexical features, which includes the necessity to update and

create new codified linguistic resources with electronic dictionaries and local grammars, fundamental to obtain careful results within NLP applications.

3

A STATE OF THE ART: SEMANTIC ENTERPRISE APPLICATIONS

3.1 Discovering the meaning with Semantics

So far, we have seen how semantics is concerned with the studying of the meaning of natural language, according to traditional theories. But if we associate semantics theoretical framework to new technologies, developed thanks to the evolution of the Web, the resulting research environment becomes more complex. First, semantic technologies is not only the study of contents in textual form, so when we treat of semantic technologies, we are actually talking about data in general. Piraquive, Aguilar and García (2009), note how new generations have understood that to build knowledge, you have to go gradually with technology growth and web evolution. Organizational structures are concretely supporting collaborative environments, changing their information systems in a way to run all actions related to this new scenario. Web 3.0 and Web 4.0 are the proposals for the future; the Network has given them more meaning and semantic contents (2009: 244). Domingue, Fensel and Hendler define semantics as a set of technologies that "provides machine-understandable (or better machine-processable) descriptions of data,

programs and infrastructure, enabling computers to reflect to these artefacts" (2011:11). Companies today have a growing interest in integrating all the data related to the core components that drive their success. Consolidating information about key concepts − such as employees, customers, competitors, operations, and products − enables them to make decisions based on a comprehensive understanding of their business environment (Stephens, 2011).

3.2 Semantic technologies adoption life cycle

The Semantic Web, as Tim Berners-Lee imagined it (2001), remains partially unfulfilled. However, in recent years, many advances have been in the development of tools to support semantic Web infrastructure. The core technologies have been identified and designed, providing some software environments to be adopted by companies: today, a concrete surfing in the Semantic Web is still a utopia, the available tools are still too sophisticated to serve the crowd and require a support by specialized professionals. The Semantic Web still has a long way to go to achieve a good level of widespread adoption by users. However, research in this field have produced notable results until today, sometimes creating valuable and adoptable tools, especially within companies.

The goal of STs developers is creating a language, a universal code, able to express the data that are present on the Web, and combine them to create new knowledge. This issue is fundamental for enterprises, which must innovate and introduce new skills. In order to adapt themselves to the new technologies, enterprises have to manage the following conditioning variables:

- *External pressures*. The adoption of new technologies requires the existence of external conditions, which would ease their absorption (institutional, market). The external pressures are related to internal resistance to change.

- *Adaptability*. It concerns the availability of quality human resources, to be employed in the process of introduction and adaptation of new technological knowledge in the enterprise.

- *Shared Decisions*. The introduction of new technological knowledge must be shared by all the corporate hierarchy levels (ownership, senior management and operations) through clear and measurable objectives. All parts of the business must feel the need for a common technology vision, and have awareness of the benefits coming from this new process.

- *Absence of external obstacles to change*. Conditions must be avoided that may prevent the launch of the process of introducing new technological knowledge (usually from the stakeholders).

- *Availability of resources*. It is necessary that human resources (Transfer groups) and techniques are properly guided and utilized, in order for the process to happen properly. Also, it is necessary to assess the risk of using human resources in routine tasks and problems.

- *Skills availability*. When transferring a technology with high tacit knowledge component, it is necessary to obtain in-depth expertise[32].

In the literature, great importance has the description of the product life cycle[33], which is the probable performance of a product in its life through four phase: introduction, growth, maturity, and decline. Applying the basic structure of product life cycle in the technology market, we can observe each phase of ST adoption. The curve, which represents the different stages, has a bell shape, going from innovators to laggard users. Rogers (1983) proposes that adopters of any innovations or ideas may fall in the following five categories, resulting from the mathematically-based bell curve by Rogers (1962): Innovators (2.5%), Early Adopters (13.5%), Early Majority (34%), Late Majority (34%) and Laggards (16%).

Adopters	Characteristics
Innovators	Innovators are ready to take risks, have typically higher social status, financial liquidity, closest contact with scientific sources and know how to interact with other innovators. Their

[32] Adapted by Schiavone F., *Gestione dell'innovazione nelle PMI*, retrieved online at http://www.economia.uniparthenope.it/modifica_docente/fschiavone/LA_GESTIONE_DELLE_INNOVAZIONI_NELLE_PMI_FERRETTI_-_CAPITOLO_3.PDF

[33] See Vernon, 1966; Levitt, 1965; Kotler, 1972; Kotler & Armstrong, 2006.

	risk attitude allows them to adopt technologies that may ultimately fail.
Early adopters	These lighthouse customers have the highest degree of opinion leadership among the adopter categories. Early adopters have a higher social status, financial liquidity and advanced education. They are more discreet in adoption choices than innovators. They use cautious choice of adoption to help them maintain a central communication position.
Early Majority	These type of adopters accept an innovation gradually in time, i.e. significantly longer than innovators and early adopters. Early Majority have above average social status, contact with early adopters and rarely hold positions of opinion leadership in a system.
Late Majority	They adopt an innovation after the average. Their personalities approach an innovation with a high degree of scepticism and after the majority of society has adopted its technology. Late Majority are typically disbelieving about an innovation, have below average social status, and little financial liquidity, in contact with others in late majority and early majority and little opinion leadership.
Laggards	They are the last to adopt an innovation. Contrasting some of the earlier categories, individuals in this category have no opinion leadership. Laggards typically tend to be focused on

	traditions, have lowest social status, lowest financial liquidity, are oldest among adopters, and show no mindedness.

Table 5. Adopters of innovation

Taking over Rogers' innovation theory (1962) and applying them to the high-tech technology market, Moore (1991) tracks a limit in the analysis of the different categories of adopters. Rogers intended the technology adoption as if there were a continuous process without interruptions. On the contrary, Moore identified many breaks in the process of adoption due to some different characteristics among adopters. These gaps are present in the technology life cycle, and are generally small; however, as it is showed in Figure 4, the great chasm occurs between the early adopters (i.e. the visionaries) and the early majority (i.e. the pragmatists). Mainly, this is the chasm in which companies are likely to fall.

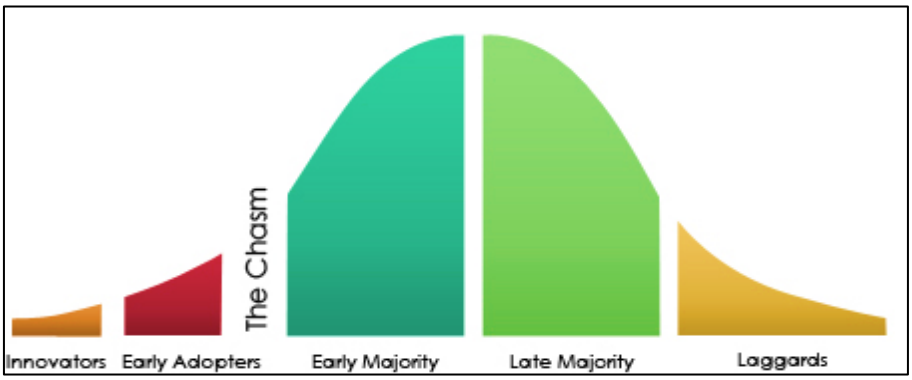

Figure 4. Technology adoption life cycle

The difficulty of intoducing a technology in the market precisely concists in moving from one group of users to another by offering the product always in a different way, since each type of customers has difficulties to accept it in the same way as it was presented to a previous group. The first interval that occurs inside the curve is exactly the one which exists between innovation and visionary, and is a result of the substantial difference between innovators who like the product only for its technology, and visionaries who are not real fanatics of the technology and expect it to be really more competitive than old ones. For example, in this split appear to be falling companies that produce neural networks and video conferencing equipment, i.e. products that for their high level of innovation have failed to win the visionaries market: in fact, their adoption has not been perceived as strategic to gain ground on those competitors conquering the poor innovators market.

Once the first hurdle passed, Moore noted another one, much more difficult to overcome: the crevasse between visionaries and pragmatists. We may find that in this abyss numerous high-tech companies have already fallen among those that possess a highly innovative and potentially successful strategy, but are not fully aware of being right on the brink. Actually, visionaries have very high expectations, are enthusiastic, are bearers of an optimistic and positive view with respect to innovation; on the contrary, pragmatists are, cautious, reluctant to take risks and have a strong system of low expectations. According to Moore, many innovations in the field of technology are beached in the passage where the spread should begin to engage pragmatists. In fact, it is very easy to confuse a sales increase with the onset of a primary market, while in reality you find yourself at the end of the visionaries

market and near the ravine. If we traslate this process of technologies adoption to enterprises, the issue becomes more troublesome.

By the way, STs are not a recent idea. Although enterprises, institutions and society are showing great attention towards the opportunities that could arise from the analysis of these data, this is not a known side of the history. We can affirm that the advancement of these types of technologies is delirious: adopters do not have time to accept a technological innovation that they have to face another one, more recent, exciting, and competitive for business. For example, Analytics market propose frequently new methods to analyse data and obtain profit from them. IBM (www.ibm.com) evaluates that each day in the Web we create 2.5 quintillion bytes of unstructured data, which cannot be ignored. Companies increasingly feel a strong need to adopt knowledge management systems that integrate internal data from traditional storage systems, and external data recovered with the analysis of the resources offered by the Web. Businesses play a central role in this scenario, since they are usually looking for new information to use as a strategic resource. Chaudhuri, Dayal, and Narasayya stress how difficult is to find a successful enterprise that has not leveraged business intelligence technology for its business (2011: 88). However, one aspect that is becoming crucial in business is the ability for companies to monetize data by selling them to other companies (Gartner, 2015).

Even if the time of stability is over, companies find many difficulties to install new technologies inside their systems. The relative stability of organizations no longer exists. The rapid growth makes that innovation is a continuous process: innovation is going on all the time, in almost every

organization. Nevertheless, many enterprises, especially those belonging to the old generation, are not ready to marry this approach. This probably happens because top management does not recognize the value coming from the use of new technologies; or because their adoption requires considerable efforts in terms of economic resources; or simply because the business strategy is not very accustomed to the change.

In the field of STs, a very interesting trend is now developing: on the one hand, companies that had software products innovate and convert their systems, offering online and offline data analysis solutions; on the other hand, we are witnessing the creation of many new enterprises, that we call innovative startups, which crowd into Analytics markets. For an innovative startup, and in comparison to companies already in the market, approaching new STs is much simpler. This depends on the fact that startups are already born with an innovative product or process, and above all, are subject to a very fast growth. If we consider the lean startup model[34], the attitude to the flux is even closer for these entrepreneurial forms: they are able to drive the technology adoption process on innovation markets, and our research on STs shows precisely this.

STs can be used to integrate heterogeneous data sets and formalize the underlying structure of the information they contain, allowing a machine to

[34] The name "lean startup" comes from the lean manufacturing system implemented at Toyota by Taiichi Ohno and Shigeo Shingo. The dogmas of the lean system are based on the creativity of the individual workers, the reduction of the lots, the just-in-time production and the control of inventories. Moreover, these dogmas include the speeding up of cycle life times; the teaching of the difference between activities that create value and waste; instructions on how to build quality into the product from the inside. Lean Startups adapted these ideas to entrepreneurial context, teaching new methods to judging their processes differently from the way other companies do (Ries, 2011).

understand their semantics (Shadbolt et al., 2006). Adopting STs consists in a real revolution for the all levels of companies: this explains why most of them are not ready for such a disruptive change, and refers to other enterprises, which offer them finite products. Specifically, ST adoption implies the:

1) Conversion of traditional data analysis and storage systems;
2) Employment of specialized professionals: data scientists and project manager that devise adoption and implementation of semantic technologies;
3) Embedding of new technologies in each business units.

If we reconsider the bell curve used to describe the technology adoption life cycle, we can observed that STs occur in the most complex phase of realization: between early adopters and majority adopters (Figure 5).

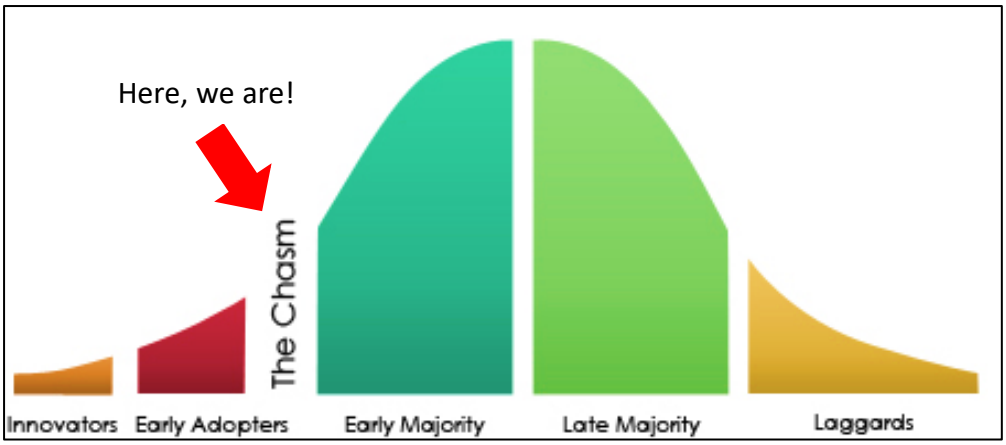

Figure 5. ST adoption stage

The ability to anticipate, address and manage changes is the only way to make proactive enterprise face the needs of a changing market. The fact of not accepting changes implies the refusal to understand the market and ignore the winning balance that brings enterprises to a certain failure (Foglio, 2011).

3.3 Semantic Technology Enterprises

Organizations need more than ever new information to improve processes, innovate products, manage customer relationships, support business decisions and introduce a smart working way. STs are expanding rapidly and an increasing number of Semantic Technology Enterprises (STEs) are developing all over the world. Innovation and new information technologies are changing the way in which companies produce, manage and communicate. Nowadays, companies are looking for much more efficient and creative business process. Carbone et al. (2012) observe that companies need to place better solutions in the market in a less time with less cost. They begin to recognize that information technologies based on analytics can provide a strategic advantage helping them to be competitive and adapt to rapid change (Stephens, 2007).

More than ever, companies need to integrate data, which contribute to improve processes, products and services, and it is not an easy task. The data taken into account are not always structured and contained inside traditional storage systems, but they are especially unstructured and generated by companies (i.e. email, documents and reports) or collected by the Web. Organizations are ever more interested in translating data in the form of

competitive advantage: they have changed their traditional business model, adopting other ones, which are more collaborative and involve an increasing number of strategic partners able to contribute to innovative process. Leading the enterprise towards a global cognitive approach is the main tool to catch innovation challenges and compete on international markets. Besides, this approach has to involve the whole enterprise, trying to understand how knowledge and technology sharing and human resources data across all organizational units can achieve the general analytical goals. It is necessary for innovative startups and early adopter companies to coordinate data collection at all levels of the company (Davenport, 2014). The costs of the required infrastructure are expansive, and the time dedicated by the expert is justified. In scenarios where a fast response is required and different departments provide data in semi-structured formats, a more agile solution is needed (Casas-Bayona & Ceballos, 2014). The race for the retrieval and analysis of strategic data is growing the tendency to outsource data analysis activities, thus allowing the semantic technology market to define itself through the explosion of STEs.

An alternative to integration of ST tools for analysis and data management inside company consists in making firms manage and acquire a finished package of information geared to company needs. In fact, many enterprises are moving toward more collaborative business models, which involves other enterprises to improve these tasks. It allows acquiring competitive advantage in a shorter time, avoiding transaction costs denoted by uncertainty and often non-recovered by investments. Indeed, it is disseminating a tendency to outsource the retrieval and analysis data activities,

and the result is an expansion of Analytics market in specific fields of application, to offer sophisticate and fitting solutions. A growing number of STEs are responding to the critical need to manage and integrate a large number of data sources and application in enterprises. Generally, STEs are organizations that have developed semantic technology tools to extract hidden meaning from data. Specifically, on the one hand market is attending to the conversion of systems of old companies always competing in the global ICT market; on the other hand, startups emergence on the market occurs with innovative services and a high professional profile.

According to the company client, STEs generally propose two types of semantic solutions: a licensed software to integrate into enterprise processes or a report with final data with respect to the analysis carried out. Probably the second solution is more widespread. Although the applications are very numerous and specific to each sector, we will focus on STEs that propose themselves as a support for users or companies in the analysis of large amount of textual data.

The increasingly availability of text data such as reports, mails and documents brings new challenges for business (Chaudhuri et al., 2011: 90). This amount of text is becoming a valuable resource of information and knowledge. The heterogeneity of data and sources complicate knowledge discovery process: it becomes obvious that too much information available are very difficult to manage. For this reason, TM and KDT used to identify any kind of textual analysis (Feldman & Dagan, 1995; Feldman & Hirsh 1997; Kodratoff, 1999, 2000; Loh, Wives & de Oliveira, 2000) can support

companies to extract hidden meaning from texts and help them to take better decision more quickly.

3.4 Sectors and Applications

As already mentioned, many applications are developed in STs field, with specific tools for each use. The collection of all documents on the World Wide Web (several hundred trillion bytes of text) is proving to be a corpus that can be mined and processed in many different ways. Advances in digital sensors, communications, computation, and storage have created huge collections of data, capturing information of value to business, science, government, and society (Bryant, Katz, & Lazowska, 2008). Following this trend, STEs have expanded traditional sectors with innovative applications. In the following Tab. 6, we resumed the major used by companies to support business.

Applications	
Knowledge Management	Any solutions or systems that deal with organizing data into structures that build knowledge within a business. KM includes courses taught in the fields of business administration, information systems, management, library, and information sciences. Many large companies, public institutions, and non-profit organisations have resources dedicated to internal KM efforts, often as a part of their business strategy, information technology, or human resource management departments. Several

	consulting companies provide advice regarding KM to these organisations. Knowledge management efforts typically focus on organizational objectives such as improved performance, competitive advantage, innovation, the sharing of lessons learned, integration, and continuous improvement of the organisation.
Information Retrieval	Information retrieval services are becoming more and more popular, offering help to a full spectrum of businesses with information storage and document retrieval. Whether you need assistance with storing paper documents or computer files, a computer information retrieval system can help you make space or free up memory on your computer.
Content Management	It is a formalized means of organizing and storing an organization's documents, and other content, that relate to the organization's processes. The term encompasses strategies, methods, and tools used throughout the lifecycle of the content. Content Management activities therefore include the writing of the texts, the organization of content, information architecture, optimization of graphics and visual aspects, from the update, the organization of the preparation, to the management through software (Web Content Management System).
Business Intelligence	The process of turning raw data into useful information that drives business decisions. BI has a number of key

processing and analysis capabilities like data extraction and transformation, dashboards, roll-up and drill-down reporting and pivot tables. These systems have always been made with different mix of software tools (Reporting, OLAP analysis, dashboards) and application software that is, containing logical true and application rules, addressed to the performance management, the optimization of a number of operational decisions or finalized forecasts and future predictions, employing statistical functions also very sophisticated. All applications have taken different names but similar meaning, such as analytic applications, analytics, and business analytics.

Predictive Analysis

Predictive analytics is the branch of the advanced analytics, which is used to make predictions about unknown future events. Typical activities are: finding new contacts, prioritize known prospects, and gain insight into buying habits; utilizing time-sensitive data with a more profound predictive engine; tracking prospects and close deals quickly; managing the sales cycle by providing insights for each account; building closed-loop reports to better align sales and marketing.

Sentiment Analysis

Sentiment analysis is the extraction of the views and relates to the processing of natural language, text analysis and computational linguistics to identify subjective information in the sources. The analysis of the sentiment is widely applied to reviews, social media and customer

service. Companies use sentiment analysis to detect if the features of their product have a positive or negative perception. Sentiment analysis is especially useful in understanding social media and product review data.

Social Network Analysis

Social network analysis describes customers' behaviour, but not in terms of their individual attributes. Rather than basing models on static individual profiles, social network analysis depicts behaviour in terms of how individuals relate to each other. In practical terms, this approach highlights connections between individuals and organizations. For business purposes, social network analysis can be employed to avoid churn, diffuse products and services, and detect fraud and abuse, among many other applications.

Machine Learning

Machine learning is a branch of Artificial Intelligence that gives computers the ability to learn without being explicitly programmed. Data are collected to optimize performance, anticipate breakdowns, and streamline maintenance, continued advances in data-processing power, sensors, and predictive algorithms. Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction. These analytical models allow researchers, data scientists, engineers, and analysts to produce reliable, repeatable decisions and results and

	uncover hidden insights through learning from historical relationships and trends in the data.
Cloud Computing	In computer science, it refers to a delivery paradigm of computing resources, such as storage, processing or transmission of data, characterized by the on-demand availability through the Internet from a set of pre-existing and configurable resources. Leveraging the technology of cloud computing users connected to a cloud provider can perform all of these tasks, including via a simple Internet browser. Companies can use remote software not directly installed on your computer and save data to online storage devices prepared by the same provider (using both wired and wireless networks).

Table 6. Main Applications in STs

3.4.1 Healthcare and Life Sciences

Over the past century, technology has played a decisive role in defining, driving, and reinventing procedures, devices, and pharmaceuticals in healthcare. Healthcare organizations are increasingly using data analytics and producing a large amount of data (Raghupathi, 2010). The applications encourage greater care for the patient and support the research. Some undertakings engaged in this type of analysis include medical records, examinations and other types of medical records to understand, for example,

what it is the evolution of a specific disease, the appropriate care and verify the correlations with other diseases. Modern medicine collects huge amounts of information about patients through imaging technology (CAT scans, MRI), genetic analysis (DNA microarrays), and other forms of diagnostic equipment. By applying data mining to data sets for large numbers of patients, medical researchers are gaining fundamental insights into the genetic and environmental causes of diseases, and creating more effectiveness in diagnosis. The analysis of sensitive data of each patient can lead to discovering a predisposition to a disease and prevent its evolution. This type of service is usually offered to clinical, diagnostic and research centres, or to the same patients.

3.4.2 Marketing and Communication

This activity is typically addressed to companies: knowing the market and understanding the needs of its current and potential customers is the critical issue. Major Internet firms such as Google, Amazon, and Facebook, continue to lead the development of web analytics, cloud computing, and social media platforms. The emergence of customer-generated Web 2.0 content on various forums, newsgroups, social media platforms, and crowd-sourcing systems offers another opportunity for researchers and practitioners to "listen" to the voice of the market from a vast number of business constituents that includes customers, employees, investors, and the media (Doan et al. 2011; O'Rielly, 2005). Tools that analyse data from Social Networks enrich Analytics explosion: studying the conversations and the opinions of users is

crucial to customize more and more the offer of products or services. Long-tail marketing accomplished by reaching the millions of niche markets at the shallow end of the product bit stream has become possible via highly targeted searches and personalized recommendations (Anderson, 2004). For example, one of the most used tools to evaluate feedback on products and services offered by social media analytics is Sentiment Analysis. Various analytical techniques have also been developed for product recommender systems, such as association rule mining, database segmentation and clustering anomaly detection, and graph mining. All these techniques define users' opinions expressed with reference to a particular topic. The meaning of the expressions used is extracted with the help of NLP, in order to learn about judgments, evaluations, emotional states and intentions. The classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. A phrase has a positive semantic orientation when it has good associations and a negative semantic orientation when it has bad associations (Turney, 2002).

3.4.3 Technology Integration

This category includes all those software applications based on semantic technology and which aim to improve and further automate the processes of meaning extraction. The enormous volumes of data require automated or semi-automated analysis, techniques to detect patterns, identify anomalies, and extract knowledge. Again, companies need to introduce new software algorithms, new forms of computation, combining statistical analysis,

optimization, and AI, able to construct statistical models from large collections of data, and to infer how the system should respond to new data. Large companies, such as IBM, Oracle and Microsoft working on information technology in general, have developed semantic services as a way to improve Web research results.

3.4.4 Information, Media and Entertainment

Many companies, such as publishing houses, press and media analysis agencies, extract information from various internal and external sources, in real time, therefore providing users with more detailed information. Burger (2008) observes that also in media industry there is a great demand for the reuse of content. However, most multimedia objects are created from scratch, due to insufficient reusability capabilities of the existing tools.

3.4.5 Legal Services

These services include the analysis of the enormous amount of documentation related to legal processes: the creation of a meaning extraction automated system would reduce the review time of court filings. Patent Analysis is proving to be a powerful tool for monitoring businesses and technologies. In this regard, some businesses directly analyse patent texts, rich with not only legal terminology, to verify the availability of technology injunction and identify other companies operating in the same sector.

Translating patent data into competitive intelligence allows firms to gauge their current technical competitiveness, forecast technological trends, and plan for potential competition based on new technologies (Fleisher & Bensoussan, 2002).

3.4.6 Insurance and Safety

Semantic technologies are applied to this sector to get better risk management, thanks to the transparency of its instruments. Predictive Analytics has quickly become an insurance industry best practice. Insurers use predictive analytic techniques to target potential clients, determine more accurate pricing, and identify potentially fraudulent claims (Nyce, 2007). Furthermore, governments use tools like Predictive Analytics studying the historical and present facts to predict future events, and guarantying National Security: social and political phenomena, track terrorist activities, welfare risks.

3.4.7 Manufacturing, Logistics and Utilities

The analysis of several resources, such as business documents, reports, business plans, helps companies to speed up and streamline the decision-making process and strategic planning. Although a large amount of data constitute an invaluable wealth of knowledge, not all of knowledge is useful, especially the one recovered from the Web. Any company has to build its own

knowledge management, based on its business needs system: TM tools reduce search time, matching it to interest resources efficiency. Everything can be shared, but not all knowledge is helpful to business processes.

3.4.8 Customer Relationship Management

Customer Management solutions allow the monitoring and analysing of users' behaviours, sometimes also identifying the optimal actions to better interact with current and potential customers. Customers' experience can be enhanced through greater personalization: the needs, in most cases, are not explicit but one must investigate thoroughly to understand what they wish, how and when.

3.4.9 Education

TM applied in the field of education refers to those systems used to improve and optimize learning processes. Using software to extract meaning from text provides a simplification and an overview of the concepts covered in a specific publication.

3.4.10 Banking and Financial

TM analyses are integrated with other statistical analysis, to support the financial sector. For example, this is the case with the assessments

necessary to grant a mortgage in the event of a bank, or to undertake an investment in the stock market, or even to prevent possible financial crises. A recent application of all this can be found in the collapsing of the subprime mortgage market. “Financial organizations could not quickly identify and quantify the exposure to subprime mortgages that might have existed in their own portfolios” (Cataledo, 2009:2).

3.5 A state of the art of STEs

Adopting semantic technologies and their main TM applications is crucial to understand what happens in the market. Those who offer this type of textual data analysis service, how they fit them to the market, and which tools they use were all basic questions for the present study, mainly carried out in 2016. This survey contributes to understand the level of shared-features diffusion and major market trends of STEs. The objective of our research is providing a state-of-the-art on STEs, which specifically offer TM to help and support business. In a previous study (Esposito & della Volpe, 2016) we detected on the Web 210 international STEs by their visibility degree: 154 have been founded on major search engines (Google, Yahoo) and 56 by secondary sources such as articles, Web forums, and online software providers. A census has been realized in a period of four months: since March 2015 until July 2015. For each STE, we tracked the following features: company name, company website, foundation year and development phase, geographical recognition, main applications, technology used, products or services offered, data sources, sector and target market.

STE	Dev. phase	Foundation	Country	City	Application	Sector	Software	Target	Data Source	Website
Apache Foundation	Company	1999	USA	Los Angeles	public sectors, open c	Healtcare	cTakes	B2B	Combined data	http://www
360pi	Startup	2008	Canada	Ottawa	assortment intelligen	Marketing	360pi online	B2B	Combined data	http://www
Abbyy developers	Company	1989	Germania	Monaco	Imaging&texting Anal	Tech Integration	Abbyy Comprenc	B2C	ED	http://www
Abzooba	Startup	2010	USA	Sunnyvale	automated distillatio	Marketing	Xpresso	B2B	ED	http://www
Acetic	Company	1994	Francia	Parigi	technological watch,	Tech Integration	Tropes	B2B	ED	http://www

Figure 6. STEs data set extract

Subsequently, we have built a data set, shown in Figure 6. The first record of our data set contains the name of the STEs identified. The second and third are interlinked: they identify the year of foundation and subsequently the development phase of the companies. The fourth and five records contain respectively the country and the city in which the company's head quartier is located, specifically only the main office. The sixth one shows the main applications and services offered by the company. In the seventh record, we have identified the sector domains previously classified. The eighth field shows the software used. The ninth field contains the information about target market: B2B or B2C. The tenth record shows data sources: external or internal data provided by company clients or a mix of both that we called combined data. Finally, the company website is indicated.

3.5.1 Geolocalization

STEs are in 32 countries, with a primary focus in the West rather than in the East, in the North rather than in the South of the world (Figure 7). STEs

are located in USA (115); Great Britain (17); Italy (10); Canada (9); Spain (7); Germany and Israel (5); Austria, France and India (4); Sweden (3); Chile, Ireland, Netherlands, Portugal and Switzerland (2); Australia, Belgium, Bulgaria, China, Colombia, Czech Republic, Denmark, Egypt, Estonia, Hungary, Japan, Malta, New Zealand, Poland, Russia, Turkey, Uruguay (1). From data observation, it is evident a higher concentration of STEs in USA (54.3%) and Europe (31.4%), followed by the other countries (14.3%).

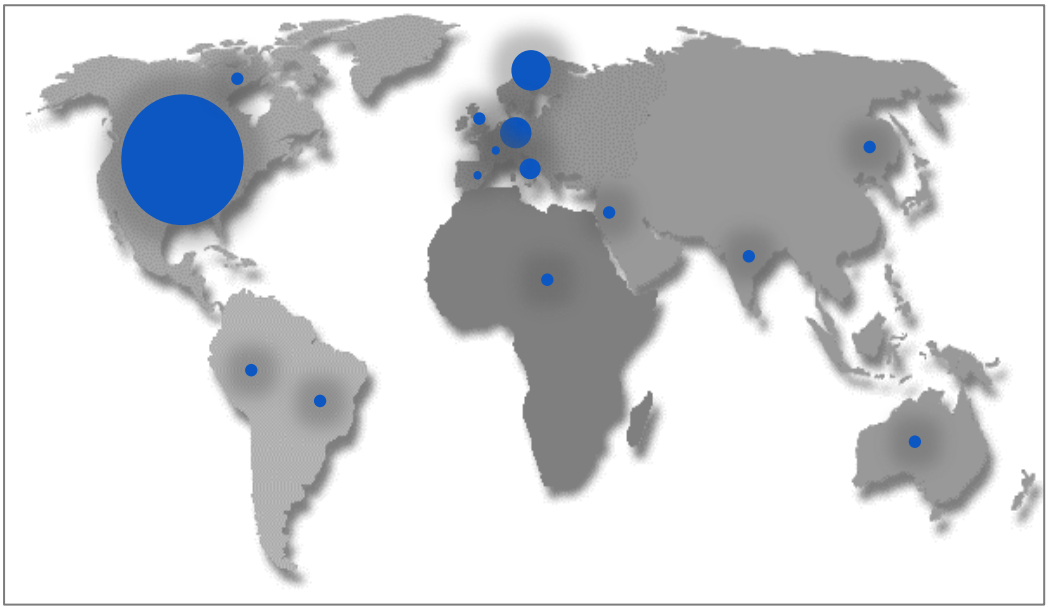

Figure 7. STEs world distribution

Table 7 reports the number of STEs for each sector by three areas: USA, Europe and other countries identified previously. Also, the Global value for each sector is present: Technology Integration (32.9%); Marketing and Communication (26.2%); Manufacturing, Logistics and Utilities (16.7%); Customer Relationship Management (7.2%); Information, Media and Entertainment (4.8%); Healthcare and Life Sciences (3.8%); Education (2.9%); Legal Services (2.4%); Banking and Financial Services (1.9%); Insurance and Safety (1.4%). As shown in Table 7, the most crowded sectors are specifically business insider; this demonstrates the importance that semantic technologies have in the innovation process.

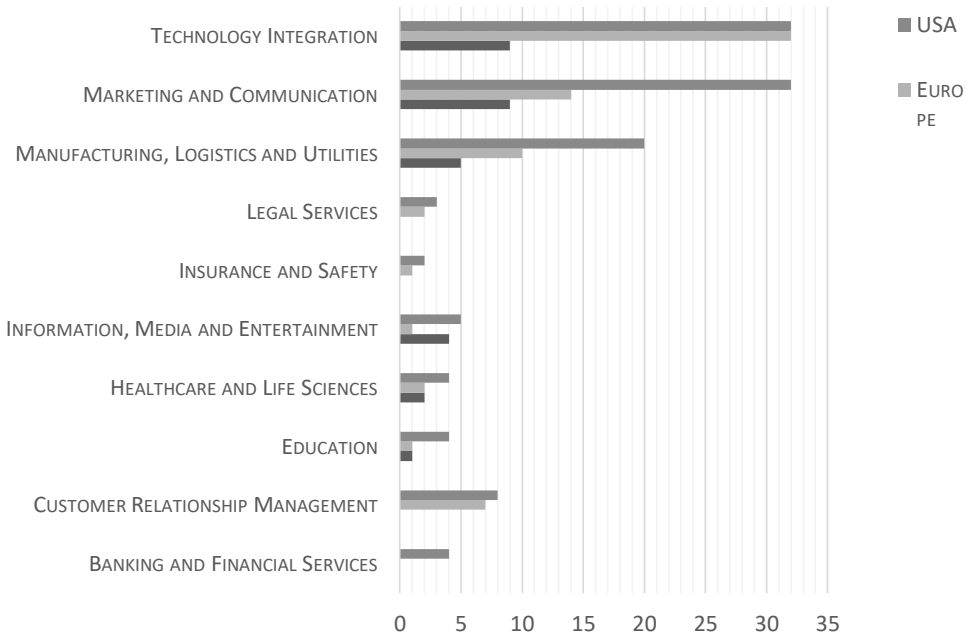

Table 7. STEs sectors by geographical distribution

3.5.2 Applications in different sectors

Focusing on the STEs identified, it is possible to state that they use TM applications to meet the requirements of their reference sector. These type of technologies could be applied to different tools based on specific objective of analysis: Knowledge Management (KM); Information Retrieval (IR); Content Management (CM); Business Intelligence (BI); Predictive Analysis (PA); Sentiment Analysis (SA); Social Network Analysis (SNA); Machine Learning (ML); Cloud Computing (CC); Natural Language Processing (NLP). Globally, the most used applications by STEs are NLP (25.2%), IR (14.3%) and BI (13.8%), while lower values are referred to CM (3.5%), CC (3.8%) and KM (4.5%).

NLP is the only higher-value application present in every sector; this is clear if we consider that TM and KDT tools require a deep study of natural language as a starting point. Considering that, some sectors do not work with every application, and in Table 2 we can see which are the most used ones. In Banking and Financial Services it is ML (33.5%), followed by NLP (33.3%), BI (16.6%) and PA (16.6%); for Customer Relationship Management, the first position is occupied by NLP (26.7%), followed by BI (20%), with the same percentage of SA and ML (13.4%); then we have CM (6.7%), IR (6.6%) and SNA (6.6.%). Instead, Education field is divided between NLP (50%) KM (20%), IR (10%), PA (10%) and SNA (10%). Regarding Healthcare and Life Sciences sectors, the more spread application is PA (30%), followed by KM (24%), IR (12.5%), NLP (12.5%) and finally by ML and CC, which have minor percentages (10%). Information, Media and Entertainment see SNA at the first

position (26%); then NLP (25%), IR (18%) SA (17%) and CM (14%). Insurance and Safety bet on PA (45%), NLP (30%) and BI (25%) while Legal services work whit IR (40%), NLP (21%), CM (20%), CC (13%) and ML (6%). These last sectors are critically more crowded with varied applications: this denote a complexity of the analysis process that requires a cooperation among tools. Thus, Manufacturing, Logistics and Utilities take advantage of NLP (40%), BI (34%), IR (10.3%) PA (10.3%), KM (6.9%), both SA and SNA (3.5%). Marketing and Communication primary focus on SNA (21%), then work together on NLP (19%), ML and SA (15.4%), BI (10.7%), PA and IR (6.2%), CM (4.6%) and finally CC (1.5%). Instead, Technology Integration bet on IR (32.3%), NLP (28%), BI (11.3%), ML (9.4%), KM (7.6%), SA and CC (3.8%), CM and SNA (1.9%). Finally, we omitted Patent Analysis among applications, because just one of these STEs operating in Legal Services works with it.

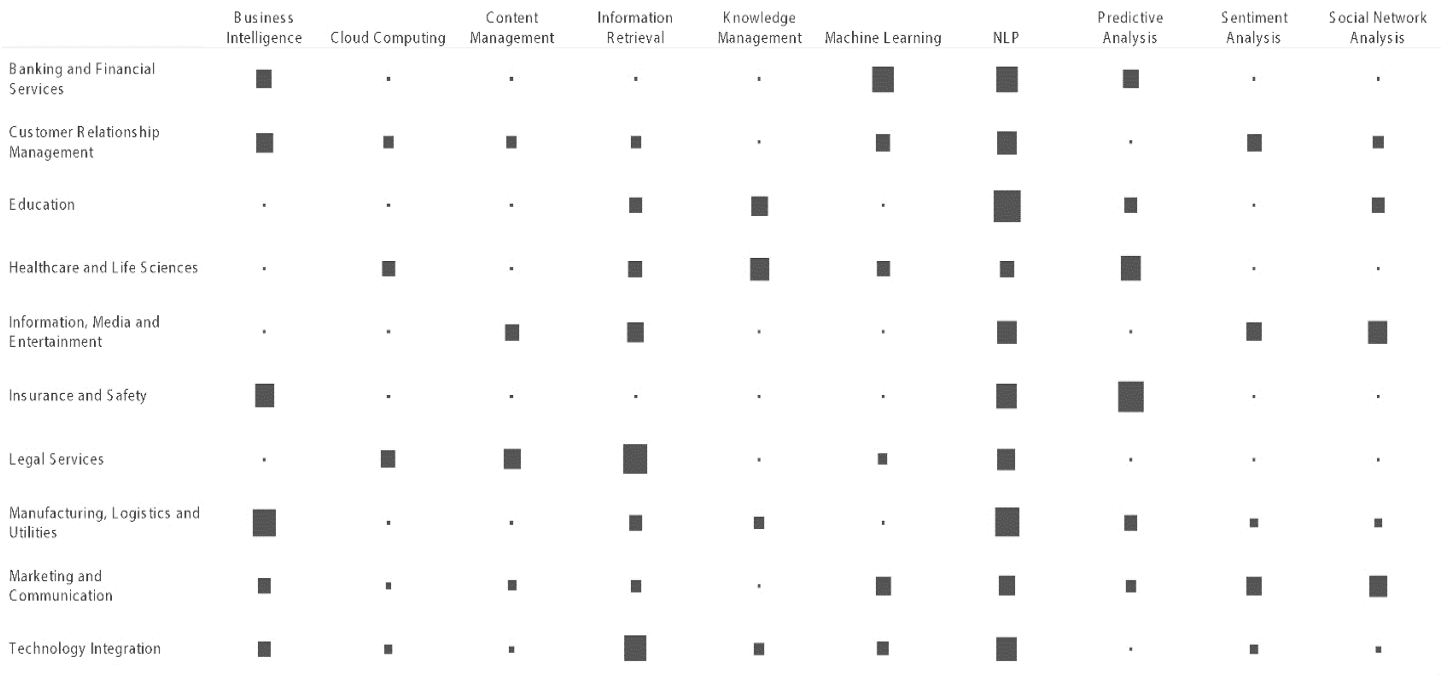

Table 8. Applications area in STEs sectors

3.5.3 Development phase

Every STE identified operates in a B2B market, and as we can notice in Table 9, 74 STEs (35%) are startups, enterprises in the early stage of their life, while 136 STEs (65%) are well-established companies. As already noticed, many companies have reconverted their own systems, while the great interest for SE has generated many new businesses, that we call innovative startups. If we consider that startups are defined as innovation movers, we should not be surprised by their so impressive presence inside semantic technology market. These findings suggest that in the last decade, interest in semantic technology opportunities grew significantly, and many businesses have enriched the market with different applications. However, many companies have not yet embarked towards Web evolution. Only 17% of STEs collected was already present before 2001, when it was published the article launching the Semantic Web. These companies, belonging to the old generation of ICT companies, have evolved over time seizing the innovation challenges. In particular, from 2001 to 2014, the number of STEs increased exponentially: on average, 12 STEs were born every year.

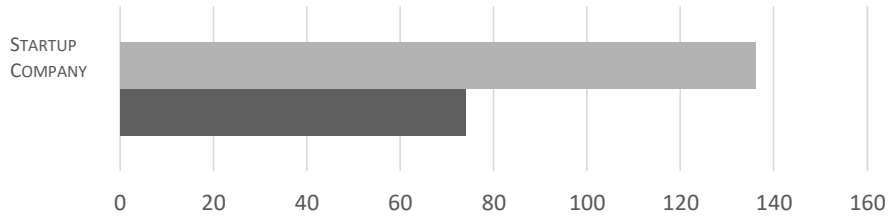

Table 9. Startup and company in STEs data set

3.5.4 Data Sources

We refer to data sources like a simply source of data. Such data might be located in a database of company computers, or another computer, or maybe they could be found in networks like data streams. In this study, specifically, we distinguish between internal data (inside company) and external data (not contained in a company database) and we considered combined data as a mix of both. In order to understand the behaviour of each sector, we built Table 4, in which we examined which type of data STEs use in percentage. Among those analysed, we did not find specific enterprises that work only with internal data. Therefore, in the following table, we present only the relation between external data and combined data. The single use of external data is more spread in almost each sector, except for Customer Relationship Management and Insurance and Safety, while in Banking and Financial Services and Education there is equity (50%). Another point of discussion regards Information, Media and Entertainment: STEs in this field are intensive as for information, and they are mostly hunting for continuous new data: they probably analyse data stream from the Web, with a critically relation between those using only external (90%) and combined data (10%).

Table 10. STEs data sources in different sectors

3.5.5 Survey results

The use of SE to analyse large amount of textual data produced on the Web is rapidly increasing. Organizations are recognizing the role of new

information to integrate them as processes inside enterprise single units. Although integrating semantic technology in enterprises requires big efforts in terms of economic resources, there is a growing tendency to outsource data analysis activities, thus allowing semantic technology market to define itself through the explosion of STEs. In order to understand better the characteristics and trends of the development of companies working with ST, we provided a state of the art of STEs, analysing textual data through TM and KDT tools, classifying them according to their degree of visibility on the Web. Starting from company websites, we identified major features of these companies. First of all, geographical distribution presents a higher concentration of STEs in the USA, to which follows Europe and sporadic companies around the world. Despite market's growth, with many emerging startups (65% of recognised companies), SE slows to establish a relation with management. Yet STEs focus their activities on data retrieval phases, so demonstrating that in terms of embedding, mining and analysis, applications are more advanced than data management: organizations are mostly interesting in researching and obtaining data, utilized to improve business units' performance, but they do not make efforts to manage them. Data show that most of STEs tend to analyse streams of data rather than single events, and that these data come from the outside of client enterprises. Accordingly with our definition, data that are not categorized into corporate databases are found mainly by means of social networks, as in the case of Customer Relationship Management or Marketing, from medical records and scientific studies regarding the Healthcare sector, the legal acts; by patents, as far as legal industry is concerned, and so on.

Another issue comes from the fact for which current TM products and applications there are still tools designed for trained knowledge specialists, and as already mentioned, they require a big work in terms of economic and human resources. As a part of knowledge management systems, future TM tools should be readily usable by technical users as well as management executives (Tan, 1999). The need not only to acquire data, but also to process and store them, perform modest operational tasks, and primarily analyse and interpret such data suitably, is increasingly becoming a shared need, which provides for the intervention of a specific professional, who many call Date Scientist, and who has, among all, many different experiences and skilled competencies.

In a context of continuous technological evolution, in which the boundaries have shifted from the national to the global, data management has become increasingly strategic to identify new customers, for the management of the current ones and the area of procurement companies. Within this, it is essential now to identify and correctly interpret data on the markets and the people who work there. Thus, business management at all times requires knowledge, and getting in touch with useful data reduces the time needed to learn about this market scenario, sometimes even allowing predictions on it. Our analysis shows that STEs are focusing on and working with data retrieval applications, be they dynamic or static, but companies are not yet ready to integrate inside them data management project that runs through the phase of assimilation of firm's knowledge. Probably, this aspect still stagnates in enterprise information phases, and it is not possible to have a long-term vision by integrating internal data with external ones. We have already noticed that companies prefer to retrieve data outside themselves, but in our opinion, the

most important result in this paper is that only 10% companies use combined data: this means that there is a large market space in this direction. Companies that learn to use combined data will have a strong competitive advantage if they are able to take these opportunities as soon as possible.

4

A NLP APPLICATION FOR BUSINESS DECISION SUPPORT

4.1 Text Mining as an exercise in Business Communication

John J. Clancy in 1999 wrote "The invisible power: the language of business" in which he described business as "cultural artefacts", not only as a matter of economics, marketing, and management. Clancy believed in invisible forces (metaphor and other figures of speech) used by leaders to deal more successfully with the economic, cultural, and environmental crises of our times. Communication, and more specifically the use of language, are critical issues for every aspect of business, particularly when you have to communicate a new process and a new product to market.

Austin (1962) affirmed that communication is an act, which can be associated to the identification of the errors that are committed in the use of certain words in everyday reality, and an innovative point of access on reality starting from the expressions we use to describe it. To be more precise, the utterances to which we are referring are not examples of nonsense, but according to the author, they "masquerade" their real meaning. An utterance has always a meaning, therefore it is linked to its context of use. By the context

in which we use words, we could obtain goals or effects, in addition to the conditions of success. As we show in Table 11, in order to recognize the performative utterances, starting from constatives, two criteria are available: the first based on grammatical frames, the second on conditions of happiness and unhappiness.

Grammatical	▪ Performative verbs (apologize, bet, sort, promise, baptize) ▪ First person present	Not all performatives follow this grammatical criterion: *"I'll come to your party"* *"Can you tell me where is Cathedral Square?"*
Conditions of unhappiness/ happiness	▪ Happiness or unhappiness ▪ True or false	▪ Even constatives have conditions of happiness or unhappiness (are also acts) ▪ If performative do not have truth conditions, may be more or less adequate to the facts

Table 11 .Criteria to recognize constative and performative utterance

The performative act is a statement that does not describe a certain state of things, does not expose a few facts, but rather allows the speaker to make a real action. Austin would express the sense bound right to an action. Through a performative act, it is fulfilled what you say to do; consequently, this act produces a real fact. From this discussion, we observed that:

- The clear distinction between constatives and performatives does not exist, because every utterance has a performative and constative dimension;
- Every utterance has a level of adequacy to reality or circumstances;
- Every utterance is an act. As such, it must comply with legitimacy and adequacy rules.

If we consider that every speech act has some effects/goals, thus when we formulate a sentence we could evaluate the effectiveness precisely, because of the effects it produces and the goals reached. In particular, this efficacy could be managed considering the hidden meaning contained in every word of the statements: they not only describe the action that is taking place, but also they achieve it. In this way, linguistics can help business to improve communication performances. Companies show significant need to communicate, and do it right. Particularly, written business communication, shows a syntactic structure not only more complex than oral, but also characterized by a reading "off-line", which is always subject to reader's interpretation and therefore probably far from the intention of who wrote it. This often causes misunderstandings: the use of automatic linguistic analysis

constitutes a solution that supports even the amount of documents and information that companies produces nowadays.

In the last decades, the interest for the use of language in business has grown: it is recognized that the hidden persuasive linguistic potential improves the positioning of company in the public consciousness. STs contribute to evaluate enterprise capability to communicate: applying TM is an exercise in communication. All documents are objects that exist between a sender and a receiver. We need to be able to understand the purpose of communication, in order to optimise our processes and lead our strategic directions. Daniushina (2010) observed how language treatment builds and maintains a good relationship with existing and potential customers or shareholders. Specifically, introducing this type of technologies will contribute to:

1) Improve the governance of public intervention in the evaluation phase of the financing of agricultural enterprises requests;
2) Increase the efficiency / effectiveness of financial public/private aid through greater integration of procedures, evaluation systems of public administrations, banking institutions and guarantee bodies;
3) Promote the simplification of the procedures for submitting applications for funding: it is expected a guided imputation of business project and its management by computer;
4) Draw up a functional business plan to define the company's objectives and to offer the same a self-assessment tool;

5) Ensure greater reliability of the business plan presented, including integrated control systems: connection with databases prices, yields, other territorial databases;
6) Ensure the continuous improvement of the system: management databases, information support to businesses and evaluators, calibration of the system over time.

4.2 The language of Business

Introducing the matter relating to business language, we must consider two fundamental aspects that make the analysis rather complex. To express the business activities in their complexity, as well as in their diversity, we have to consider on the one hand the sublanguages that characterize this world, and on the other hand the terminology used. For instance, sublanguages are employed to describe professional activities belonging to different business sectors: banking, trading, accounting, communication, logistic, administration etc. Another issue is referred to terminology: no one could say that business has a specific and limited vocabulary. The study of language in business contexts is highly interdisciplinary (Studer, 2013). Business activities are so complex that they require the application of several disciplines at the same time, therefore the use of specific languages. Although, it is always necessary that the circumstances in which terms are uttered, should be in some way, or ways, appropriate. The combination of business functions and processes is impacted by improved communication: from company to company, we have seen

language skills consistently deliver tangible business value and virtuous results for organizations that invest in language training.

Ford and Wang (2014) observed how the use of language in the field of strategic management has been the subject of many studies (Leontiades, 1982; Hoskisson et al., 1999; Nicolai & Dautwiz, 2010; Ronda-Pup & Guerras-Martin, 2012) just because there is no unique classification of words as it exists for other disciplines such as Economy. Every strategic document is a stream of decisions (Mintzberg, 1978) and actions, whereby it does not just describe reality but performs it in the same moment in which such decisions are representing it.

The language of business world is very multifarious: we tried to identify its features and behaviour, considering the evolution that it has suffered primarily with globalization of markets. As we saw in the previous paragraphs, in the last thirty years it has increased significantly the interest of researchers for the variety of specialist languages. However, in relation to different specialized varieties of the language, it has not been yet developed a unified terminology, and tags used in this field of research by various researchers are different. Nevertheless, we must consider that the use of certain terms has entered in the common language through mass media, as we know, and this passage often become the point of contact between specialists and people. Thus, we will have a kind of coded language typical of the economy, and another kind of language that instead has developed among the experts, a kind of jargon, which then became part of everyday life through media. For instance, in some previous studies (Elia et al., 2014) we dealt with the specific lexicon used by media to describe the phenomenon of startup companies. We

studied how the Italian terminology and this specialty language can be used in automatic text analysis routines. Using NooJ, an environment for the automatic processing of a natural language, through the application of electronic dictionaries of terminology and specialty, we analysed more precisely a corpus of 2000 journal texts centred on the startups topic. After the analysis, we detected about 400 entries, a great part of which belonging to the semantic fields of Economics and Informatics, and only a small part to Professionals, Revenue and Law. Moreover, it emerges from the analysis that the terminology of the world of startups is rich of foreign words, coming mainly from the United States. Through the study of the presence, frequency and origin of these lexical entries, it is possible to grasp certain phenomena implicitly expressed in the texts analysed, with the objective of a fuller study on the evolution of the ecosystem of startups. The specialty language that has been determined requires continuous and on-line monitoring of the dynamic and innovative vision, which is inside the specialized terminology field. On the other hand, from this it derives the fact that there is a very strong presence of borrowings from the English language in the lexicon of startups. This data could be taken as an invitation to extend the research by adapting these terms to the Italian language system, thus satisfying the need to find effective correspondents to describe certain concepts. This case shows us even how mostly technical words enter in our common language through mass media, and became our opportunity to comment some socio-economic events.

About this, nowadays, in the language of Business, we could identify two levels, i.e. *specialist* and *popular*. The specialist language includes all features of a sector language, while the popular one is spread through mass

media. The popular level resorted to some mitigations making the language less complex, also recurring to metaphors. Predominantly, economic dictionaries characterize the language of Business, but the enterprise system is so complex that it naturally requires the intervention of more specialized languages in the interaction processes, based on the nature of the enterprise and of the marked in which it operates. We consider the recognition of economic terminology only as the basis for a larger study that may involve other types of specialized language processing, within the analysis of textual documents that provide information to support strategic decision. Thus, the language of business is partially the language of Economics, as it uses many words that have a dramatic nuance ("crisi economica" as "economic *crisis"*) or military origin ("manovra finanziaria" as "budgetary manoeuvre") as shown by Piparantainen (2001). The most striking feature of the business language in Italian is the presence of foreign words and expressions, especially of English origin, so abounding of technicalities and terms that are often incomprehensible to experts. To obtain an efficient TM system and applying it to business document analysis, we have to consider a typical economic language, opening our analysis field to other knowledge domains. As for business language, following LG framework of natural language analysis, we can achieve what follows:

a) The studying of the dynamics underlying the introduction of foreign words in Italian, from the lexical point of view but also from the syntactic one, because very often a consideration in

Italian is not adequate and above all makes the communication between the parties less immediate and effective.

b) Business language is constantly evolving, because it is strongly linked to technological innovation, for which one defines new management models geared to the continuous improvement of organization, processes and products. This generates new terms with which to update TM systems, to produce a more effective analysis.

In the following paragraphs, we will provide an example of business document automatic analysis based on LG framework: we will analyse a corpus of Business Plans of the Agricultural sector to extract new linguistic resources, create *ad hoc* local grammars and other tools. In brief, we will develop a complex system to understand the features of the language used by experts in this field.

4.3 A NLP model of analysis

We present our model of DSS specifically aimed to recreate a *research environment*. A support system to business decision must be based on the complex conditions in which companies operate nowadays: decisions are taken in the shortest time and at the lowest costs possible. This scenario implies today a close cooperation between high-technologies tools and human skills, to

develop mutually sophisticated solutions. Human intervention cannot be ruled out; machines can only apply standard procedures, established by researchers, thus saving our time in the evaluation of the documentation. Specifically, human intervention comes before the linguistic analysis and the post analysis. In the phase of pre-analysis, it necessary to recognize the resources that we need, and make them available in a readable-machine. In the phase subsequent to linguistic analysis, researchers can intervene to interpret the linguistic data output by the machine releases, and only at that point, they can decide what type of application is most suitable to their case, i.e. one among those we evaluated in the previous chapters.

The component of a DSS overlap the Linguistic Model of document analysis:

1) **Input**: collecting and selecting Business documents to analyse.
2) **Linguistic Knowledge and Expertise**: inputs requiring manual intervention by researchers (term extraction, normalization and preparing text phase).
3) **Outputs:** text analysis provides linguistic data results, which are interpreted by researchers; at this time, data could be processed with several applications (statistical, CAT, strategic NLP applications).
4) **Decisions:** Business documents are deeply analysed and results generated by the NLP application can supply decision-making process.

Business documents are files that provide details related to a company. In fact, they are used to communicate, transact business and analyse productivity. In the meantime, business documents provide the profile of an

organization and may be referred to for years to come: it is very important that they are well prepared, in order not to convey a negative impression about the person who wrote it or the company for which it is written. Thus, writing excellent business documents is imperative for any working professional: they can be digital, occurring as electronic files, or in physical form, written or printed on paper. Business documents range from brief email messages to complex legal agreements. Some documents are prepared by employees and business owners, while others are drafted by professionals from outside of the company, such as accountants and lawyers. Shown below, there are the most important and internal business documents, and among those, we can choose to analyse some types.

a) *Business Plans*. The first representative document for a company is the Business Plan. This document outlines common goals and objectives of the business, along with a management plan, marketing strategies and a financial budget. The Business Plan is a document that is often presented in public, especially when it is directed to possible investors, so it is good to remember that this document can perform several functions in addition to the internal strategic planning. Besides, it is important to stress that a business plan, if prepared rigorously and continuously updated, can be also a useful tool for post-evaluations of the results achieved. During the first few years of business, comparing actual achievements with those

envisaged reported in the Business Plan, it could certainly help evaluating whether or not you are going in the right direction.

b) *Business Reports*. Business documents may also refer to business reports, including annual sales report, annual budget reports for production and monthly update reports from marketing departments and production lines. Business reports convey formal written information: include statistics, charts, graphs, images, case studies and survey results. Also, they contain many topics such as safety compliance, sales figures, financial data, feasibility studies and marketing plans. Business reports are often published for the benefit of investors. If a report is periodic, such as a monthly sales report, the format is closely severe and it is used to compare results with previous reports.

c) *Letters, mail and memorandum*. Business letters are used to communicate with individuals outside of the office as customers, colleagues in other businesses, service providers, professionals who advise the business, government officials and job applicants. A business letter is usually formatted in block style; it can be emailed or delivered by mail. If a letter is sent in the text of an email, the sender includes his name, job title and contact information at the bottom of the email. Instead, co-workers typically use email to convey information to each other. Memoranda, before mails, are still used in situations

where a message is meant to accompany a specific file, and in cases that require more privacy than an email. Typically, the text is formatted in one or more paragraphs, and both a memo and an email contain the sender, the recipient, and a subject line.

d) *Financial and Accounting Documents*. They are used to prepare budget proposals or evaluate returns on investments. These documents include receipt records, payroll reports, paid bills, bank statements, income statements, balance sheets and tax reporting forms. A business owner uses these documents to determine the financial success of the company, and to identify unproductive areas. A department head might use financial documents to prepare a budget proposal. Accounting activities of business must track all incoming and outgoing funds, both on a daily and annual basis. Accountants must issue invoices for clients, maintain monthly budgets for the company's departments, and write annual financial reports for investors and shareholders; while invoices are typically between one and two pages long, an annual report may be upwards of fifty pages.

e) *Operational Documents*. Operational documents are pages or reports with information that may be written on a daily, weekly or monthly basis. These documents relate specifically to operations and services provided by a particular department or division, and are distinct from the general administrative (housekeeping) records. Operational documents include

schedules, notes written during operational meetings and project proposals for internal and external tasks that the business must complete. These tasks may include interior restructuring and an external marketing campaign.

f) *Customer Documents.* Companies selling services and products directly to end customers often have a customer service department that sells these products and services, as well as provide support and answers for customers. Common business documents in the customer service department include order forms, customer complaint forms and brochures with descriptions of products and services. A company uses documents to transact business with its clients. To save time, these documents may be formatted as a form, such as an order form, transmittal page, invoice or receipt. The types of transactional documents used vary somewhat by the nature of a business. An insurance agent, for example, generates insurance applications and policies, while a lender uses loan applications and mortgage documents. In some fields, businesses enter into agreements and contracts with others, through the mediation of a lawyer.

After choosing the documents types to process, we proceed with pre-processing of unstructured linguistic data. This phase passes through the application of LG theory and language formalization methodologies (syntactic tables, electronic dictionaries and local grammars). Subsequently, we can

process the texts in NooJ NLP software environment. After this linguistic pre-processing phase, we obtain sets of results which can be integrated into applications of Computer Aided Translation (CAT), statistical tools or specifically on the NLP applications described in the previous chapter (KM, IR, CM, ML, BI, PA, SA, SNA, CC), as also shown in Figure 8. Having examined key business documents that can be subject to analysis, we must consider that there are several external documents to the assets but that support strategic decisions. For example, we may think of the national reports on the major issues of social responsibility and business ethics, on the progress of a country, or simply of information from other companies. Other documents may be found inside the large amount of data that the Web produces and that cannot be ignored by companies, especially when we refer to consumers' perception compared to products, services or general themes. Support for business decisions is a difficult task that requires several steps, and an enormous amount of material to be analysed to evaluate all assumptions. Discovering the meaning with NLP applications requires a complex analysis, considering also the presence of several linguistic forms in a large corpus. To provide a first approach to the aid that ST and, in particular, TM can provide, we will see in the following a practical example analysis of some business documents, to establish a useful logical development of an efficient system.

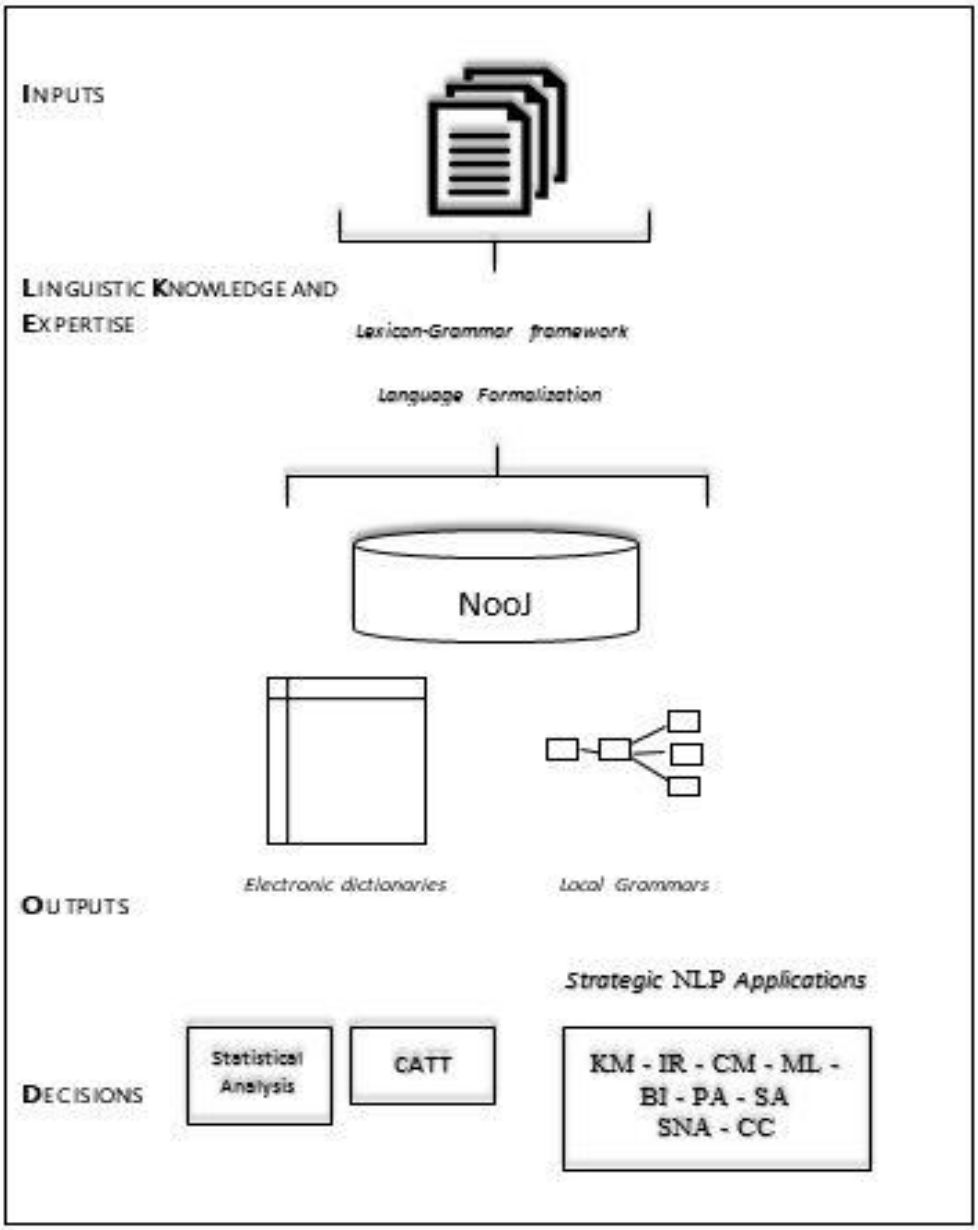

Figure 8. A NLP model of analysis

4.4 Pre-processing of linguistic data

In this example of NLP environment research applied to business documents, we take a linguistic knowledge approach to identify innovative language used to describe new entrepreneurial activities, more specifically in texts on innovative entrepreneurial activities. This is an interesting case showing how language changes continuously. An innovative startup needs to express itself through an upgraded innovative language, to describe processes and products. In order to prove our model of analysis and recognize new linguistic resources, we used several functions of NooJ to process the type of knowledge presented in business documents. This approach, based on LG framework, is extendible to every knowledge domain, although we processed only business documents of the Agri-food sector, which more than others presents critical issues in knowledge management and representation. To achieve this project's primary goal with NooJ, i.e. analysing a business document corpus, it was first necessary to identify and collect the documents. As already stated, among business records, Business Plan plays a strategic role in the description of a company: it is subject to evaluation by investors, institutions, providers and offers a complete view of the enterprise. Moreover, it has other important functions: to demonstrate clear, concise and precise communication skills; recommend necessary actions in future; examine available and possible solutions to a problem, event, situation, or issue; describe business activities; cover the company's situation; analyse business trend and financial activities. Specifically, we chose to apply our linguistic analysis to 5 Agri-food Business Plans, because this sector presents a lack in

knowledge management: a fragmentation problem that affects not only businesses, but also the organizations responsible to produce and disseminate knowledge, such as universities and research centres. This problem is also caused by a poor distribution and use of computer applications by Agri-food management. Furthermore, business documents in Agri-food are likely to be rich in technical and terminological words, also with reference to other knowledge domains: Biology, Chemistry, Medicine, Gastronomy, Economics, and Agriculture. We proceeded with the automatic text analysis using the Italian Module of NooJ. As already stated, NooJ is a complex NLP environment in which it is possible to read automatically digitized texts, locating inside them specific linguistic patterns in the form of concordances. In the task just shown, in order to tag all MWALUs, NooJ matched our corpus with the compound word electronic dictionary, and in different knowledge domains. The Italian Linguistic Resources and the electronic dictionaries we used, developed in accordance with LG framework, are mainly of two types, separable by the formal and semantic aspect of their content.

- Dictionary of simple words (called DELAS-DELAF) that include all the simple words of Italian, simple word atomic linguistic units (SALUs) and multiword atomic linguistic units (MWALUs).

- Dictionary of compound words (called DELAC-DELACF), including MWALUs and compound words, i.e. sequences of two or more words and which together form single meaning units.

Since the DELAC-DELACF is essentially a terminological dictionary, each entry has one or more terminology labels, matching the knowledge areas in which a specific compound word is attested. Nowadays, there are 180 knowledge areas inserted in the DELAC-DELACF, the most important of which are Medicine (tagged with MED, with some 63,000 inflected forms); Economics (tagged with ECON, with some 58,000 inflected forms); Computer Science (tagged with INF, with some 38,000 inflected entries); Law (tagged with DIR, with 14,000 inflected forms); and Engineering (tagged with ING, with some 5,000 inflected forms).

We added to these two main dictionaries a small dictionary created by a previous study (Elia et al., 2014) and reporting the startup ecosystem. This dictionary consists of about 400 single and compound words, many of which have no equivalent in Italian. In the development and management of these electronic dictionaries three main steps are crucial:

a. *Lexical acquisition*. During this on-going phase, MWALUs are extracted from corpora and/or certified glossaries.

b. *Morpho-grammatical and syntactic tagging*. Each lexical entry is given an inflectional paradigm, in order to be inflected. The following string gives a sample of this morpho-grammatical formalization procedure:

startup a vocazione sociale, N + Genere =f + Numero =s + Class = NPNA + Term = ECON + Eng = startup with social vocation, Class = NPAN[35]

c. *Testing on corpora*. The dictionary is used to automatically analyse and process large corpora.

These three criteria identify a greater number of compound words normally assigned to a given language. As we will see later, especially as regards terminology compound words, the analytical formulation put here in evidence allows a wide lexical coverage and is of great importance to all the lexical analysis activities, including those based on information retrieval and on natural language processing. In all the languages of the world, a close relationship exists between terminology and compound words: in fact, to express precise meanings terminology needs compound words, especially MWALUs, as evidenced by the presence in specialized lexicons of many words compounds, which in some cases reach 90% of the listed lexical items. However, we must not forget that the use of compound words is also well documented in records that are not marked terminologically. To lemmatise and identify correctly compound words, but also to differentiate them from free word groups, Silberztein (2014) adopts the following award criteria:

[35] In Italian, words maintain their singular structure also when they are referred to plural. The tag "N" (noun) indicates the grammatical function of the whole compound. Other elements indicate the morpho-grammatical patterns of each compound structure; NAPN (noun + adjective + preposition + noun) for the internal structure; "f" and "s" (feminine singular) give inflection indications; "ECON" (terminological tag) refers to the knowledge domain of Economics.

a) Semantic atomicity: if the precise meaning of a phrase cannot be deducted from the meaning of its components, then this word group is a compound word, and it will be lemmatise; which compared to their core adds elements that do not change the meaning but participating in the construction of complete sense, not literal in the set.

b) Distributional restriction: include this feature when the components of the word group can be substituted as belonging to certain specific distribution classes.

c) Shared and institutionalized use: some groups of words, including those from semantically and distributionally free, used with almost mandatory forms and theoretical opposition to other potential syntactic constructions that are equally valid but are almost never used. This happens when a foreign word enter in common use. In similar cases it will be necessary lemmatise such compound words.

Terminology compound words, unlike the generic use of simple words, are categorized unambiguously - or rather, belonging to different semantic fields, each of them has a single meaning. This feature is valuable for language terminology in the association of signifiers and signified, and needs to be as specific as possible. The main goals of terminology actually concern the unambiguous classification of objects and concepts and then, in a second step, the achievement of a non-dysfunctional technical-scientific communication.

Terminological language, by definition, cannot be ambiguous, therefore it finds in MWALUs the most appropriate communication and adapted linguistic formulation forms.

4.5 Automatic analysis and Linguistic Resources

In order to achieve the automatic processing, we converted every document in the MS-Word format; subsequently we unified our 5 Business Plans the in a single Text unit (TU) with a dimension of 71, 4 KB. As predicted, after the Linguistic Analysis with NooJ showed in Figure 9, our corpus presented many new linguistic resources, due to the innovative character of the business documents analyzed.

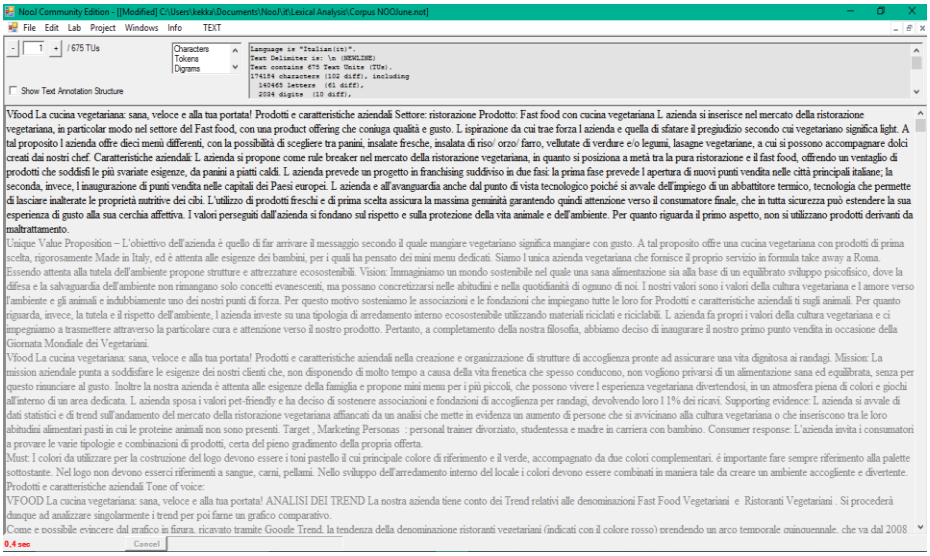

Figure 9. Text cleaning and Linguistic Analysis

As shown in the previous figure, among the 473 unknown words, we added about 80 new entries to our Italian electronic dictionaries; we associated to each entry a morph-syntactic code, a semantic code, and an inflectional paradigm. Based on their features and properties, we collected Innovative Linguistic Resources dividing them into 4 classes, as follows:

- *Prefixes* (16%). The morphemes posted at the beginning of the lexemes are often equipped with an autonomous meaning (**agri**campeggio, **bio**edilizia, **ciclo**turistici, **agri**festival, **eno**turismo).

- *Acronyms* (10%). New forms that express the meaning of MWALUs recently entered the common usage (www, html, ddl and http).

- *News* (19%). Completely new words, however, are derived from words already present in our linguistic resources (circolareggiante, acrilamide, flavonoidi, motivatore, vegetariano).

- *Foreign words* (55%). Most words used in business come from English language, or rather from the American world, and they fit very familiar in our entrepreneurial culture (startup, competitor, asset, follower, smart).

Using different terminology is inherent not only to various groups of users, but also occurs within a given group, e.g. of business analysts applying different labels to describe the same artefacts. This may lead to serious

problems in sharing, discovering, and reusing the already modelled processes as well as it hampers the effective collaboration in the process modelling phase.

According to the results of our analysis, we built a special dataset with MWALUs and compound words to evaluate our tools. The relation between these linguistic forms and terminology is rather strict. Terminology concerns the unambiguous classification of objects and concepts, with the aim to characterize technical-scientific communication. AS already stated, terminology cannot be ambiguous; therefore, it is in the compounds words, and MWALUs, that it finds the most appropriate forms.

In the following Table 12, we collected the most common compound words and MWALUs that NooJ recognized in our corpus. For each entry, we indicated: its category; its POS (internal structure); the Domain regarding existent terminological tags in our Italian electronic dictionaries; and the result with reference to presence in our Linguistic Resources (known word, unknown word, known but not as compound). Regarding totally unknown words, we noticed the presence of foreign words that mostly described business processes.

Entry	**Category**	**POS**	**Domain**	**Result**
Istituti di ricerca	N	NPN	DIGE	Known

Analisi della concorrenza	N	NPN	ECON	Known
Costi di gestione	N	NPN	ECON	Known
Capacità produttiva	N	NA	ECON	Known
Dati statistici	N	NA	STAT	Known
Economie di scala	N	NPN	ECON	Known
Posizionamento strategico	N	NA	ECON	Known
Agricoltura biologica	N	NA	AGR	Known
Settore agricolo	N	NA	ECON	Known
Marketing comparativo	-	-	-	Not considered as compounds

Impatto ambientale	-	-	-	Not considered as compounds
Parco fotovoltaico	-	-	-	Not considered as compounds
Business Plan	-	-	-	Unknown
Take away	-	-	-	Unknown
Swot analysis	-	-	-	Unknown
Unique Value Proposition	-	-	-	Unknown

Table 12. Example of MWALUs and compound words detection

Moreover, we detected with NooJ the most recurrent words among all. The first one is offerta (offer); followed by prodotto (product), azienda (company), attività (activities), produzioni (productions), mercato (market), servizi (services), qualità (quality), struttura (building), turismo (tourism). The prevalence of the word "offer" in the text shows us how is used the Business Plan, that is as a business document, not only to describe company activities

but also as a strategic tool of communication and promotion with stakeholders. Corpus exploration leads us to recognize a substantial number of operators that present these arguments (N) and that provide indications of circumstantial nature. Those observations allow us to establish a dozen verbal semantic groupings:

a) *Offrire, proporre, presentare*

b) *Assicurare, garantire, attestare*

c) *Soddisfare, accogliere, rispondere*

d) *Posizionare, immettere, inserire, introdurre*

e) *Mirare, ambire, puntare*

f) *Sviluppare, accrescere, espandere, incrementare, potenziare*

We can observe that those mentioned are all transitive verbs, where the action passes directly from enterprise (startup) to the object (person, animal or thing) that receives or suffers such as in those cases:

1) *L'azienda assicura la massima genuinità dei prodotti*
2) *L azienda propone strutture e attrezzature ecosostenibili*
3) *La mission aziendale punta a soddisfare le esigenze dei nostri clienti*

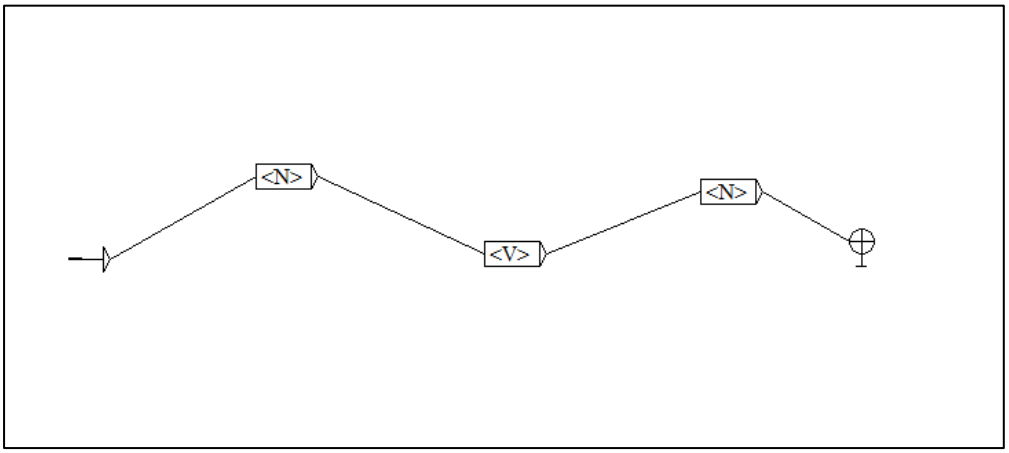

Figure 10. Local grammar simple graph

We created a local grammar on the base of the most frequent simple sentence forms that we found in the corpus. In the followed paragraph, we collected new linguistic resources, not present in our resources, due to their innovative nature representing startup ecosystem.

4.6 NooJ Local Grammars and Knowledge domains

Regarding Innovative Linguistic Resources, we identified about 80 new entries and inserted the inside Italian electronic dictionaries of compound words. In order to build local grammars for our new linguistic resources, we created six new inflexion codes, even for new formal structures, as for instance AAN. These new inflexion codes are required by the fact that more than half of the new linguistic resources are in foreign languages (almost entirely in

English); considering that we chose to avoid the use of long paraphrases, Italian translations are not completely adequate to represent their meanings. In Figure 11, we report an example of a compound word inflection local grammar for “early stage investment”.

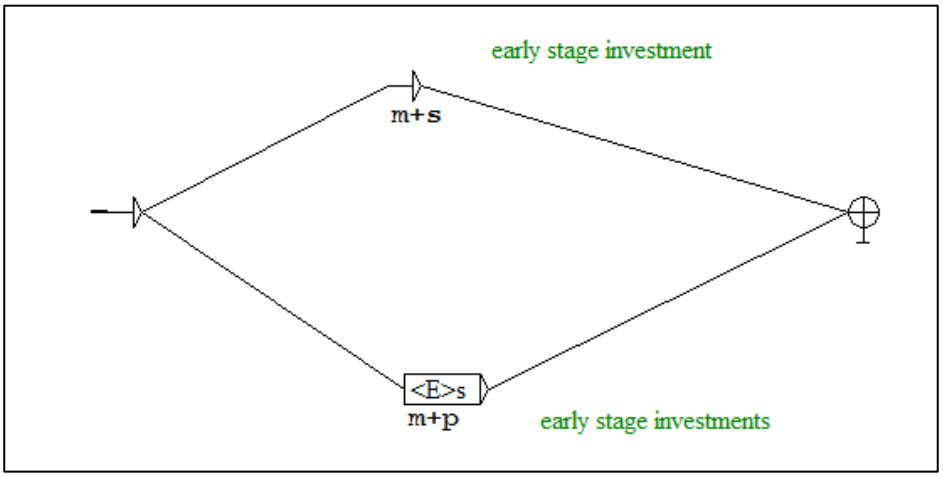

Figure 11. The graph of inflectional grammar for “early stage investment”

A further text classification task was performed on the already mentioned monitor corpus, to highlight the relationship existing between domain terminology and MWALUs, which are massively present in terminological texts. As regards the knowledge domains of our new entries, we observed that in the terminological tags recognized by NooJ there is a prevalence of Economy (63.4%); followed by Professions (6.8%); Medicine (5.9%); Generic

Dictionary (5.6%); Engineering (4.1%); Tourism (3.5%); Law (3.4%); Computer Science (3.1%); Gastronomy (2.5%); Mathematics (1.3%).

Knowledge domain	**MWALUs (%)**
Economics	63,4
Professionals	6,8
Medicine	5,9
Generic dictionary	5,6
Engineering	4,1
Tourism	3,5
Law	3,4

Computer Science	3,1
Gastronomy	2,5
Mathematics	1,3

Table 13. Knowledge domain Classification

The aim of our research was to process the type of knowledge presented in the Business Plans of five innovative startups operating in the Agri-food sector. Thanks to our analysis, we observed that companies in a primary phase of their life express themselves with an innovative language, which is not always understandable for stakeholders, and this determinates a communication gap. This make Agri-food present a sort of weakness in management and knowledge representations, due to a poor distribution and use of computer applications in the management processes.

We used NooJ to create semantic expansion networks, extracting concepts and representing them by means of clustering schemata: about 80 entries have been added to our Italian electronic dictionaries, and six inflectional codes have been created. We showed how the automatic processing of textual data reduces the amount of time spent for the measurement and analysis of a project, allowing a possible massive control of the documentation, critically lowering the dispersion of information related to

the company, identifying benchmarks and monitoring the quality and adequacy of the language used. In our future work, we will try to identify other Innovative Linguistic Resources: the growth and continue update is the main feature of startups and this determines an expansion of the vocabulary typically used to describe their traditional business functions.

Another issue will be to verify the compliance between the language used in business documents and the real entrepreneurial projects, and also the effectiveness of the language used with reference to the communication goals prefixed. This last point will be realized applying formal description techniques based on conceptualizations and, therefore, on ontologized concepts.

DISCUSSION AND CONCLUSIONS

To be competitive in the market and face innovation challenges, companies need to acquire specific knowledge, growing and communicating outside their values. STs can help companies to realize these goals: specifically, TM applied to business documents could support business decisions and improve the way in which companies communicate. At any rate, we need some clarification to understand better the general purposes of this model of linguistic data analysis. Natural language formalization is not the only way to process business documents. An efficient analysis of a large amount of different business documents could be achieved also by means of statistical text-analysis techniques: only with the cooperation among several tools of analysis, problems would receive more adequate and complete solutions. Despite the detailed level of the methodological and theoretical framework provided, which gives us great hope as for the analysis results, we acknowledge that the formalization of all linguistic phenomena is extremely complex. Also, it is prone to some limitations coming from the non-transferability of certain human speech acts, being the language a living and constantly evolving organism, and which requires a careful monitoring and maintenance. For this reason, we need to detect supplementary technologies to

support our research, sharing knowledge and hire professionals dedicated to the development of these tools. Adopting a new technology within an enterprise means to adapt an already existing structure to a new technological environment, creating integration mechanisms that facilitate innovation, despite a strong resistance to change. In fact, during our dissertation we have seen how we are still in the phase of early adoption of STs by businesses, especially mature, and what difficulties there are in the process. This naturally cerates the recourse to outsourcing, which is an easier and smart solution for companies more willingly to receive guidelines, as for instance a report with some analysis results. This choice has its advantages, as cost reduction; it does not require continuous maintenance, nor requires to convert all adapting technology innovation. Nevertheless, some actions cannot be overlooked: for instance, the employing of staff able to read and understand the results of data analysis, such as data scientists; yet it is necessary that the data collected are contextualize inside the knowledge and cultural fields in which a business organization acts.

We admit that this study is a first approach to the development of a linguistic support to embed inside decision-making procedures, or better in DSSs, with a particular reference to the document-driven analysis. However, we noted some limits, as for instance, the fact that the Business Plans of innovative startups does not have a standardized structure. They are very free, from a linguistic and structural point of view: so, if on one side we may exactly capture the most diverse linguistic phenomena, on the other one we may find it difficult to apply a systematic comparison between two or more documents, even of the same industry. Business Plans are interesting from the point of view

of terminology, as they use a specialized technical and highly innovative language, which enriches the Linguistic Resources built according to LG, through the creation of specialized electronic dictionaries, in a way to more and more detailed and efficient analysis. Moreover, business plans are useful to stress stresses the phenomenon of foreign words used in Italian companies language, as they represent a unique code to define processes, actors and tasks. It is for this reason that in our future work we hope to follow the line of research described in this dissertation, with the aim to explore its main theoretical topics, applications, and tools, together with their evolution. Also, we hope we will have the opportunities to revise our methodological and procedural errors, in order to manage the full-of-data world we have been describing.

REFERENCES

Alter, S. (1980). *Decision Support Systems: Current Practice and Continuing Challenges*. Reading, Mass.: Addison-Wesley, Inc.

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Barbera, M. (2013). *Linguistica dei corpora e linguistica dei corpora italiana. Un'introduzione*. Milano: Qu.A.S.A.R s.r.l.

Benjamins, V. R., Radoff, M., Davis, M., Greaves, M., Lockwood, R., & Contreras, J. (2011). Semantic Technology Adoption: A Business Perspective. *Handbook of Semantic Web Technologies* (pp. 619-657). Springer Berlin Heidelberg.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284 (5): 28-37.

Blomqvist, E. (2014). The Use of Semantic Web Technologies for Decision Support – A Survey. *Journal of Semantic Web*. IOS Press, 5(3): 177-201. Available online at http://www.semantic-web-journal.net/sites/default/files/swj299_1.pdf.

Bolasco, S. (2012). Appunti sull'analisi statistica dei dati testuali e cenni sull'analisi automatica dei testi. Retrieved June 24, 2016, from

http://www.memotef.uniroma1.it/sites/dipartimento/files/file%20lezioni/2_%20dispensa%20ADT%20x%20MEAD%202012.pdf

Bryant, R., Katz, R. H., & Lazowska, E. D. (2008). Big-data computing: creating revolutionary breakthroughs in commerce, science and society. Retrieved September 12, 2015, from http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf

Bū̄rger, T. (2008). The Need for Formalizing Media Semantics in the Games and Entertainment Industry. *Journal of Universal Computer Science*, 14 (10): 1775-1791.

Bussler, C. (2003). The role of Semantic Web technology in enterprise application integration. *IEEE Data Eng. Bull.*, 26(4), 62-68.

Carbone, F., Contreras, J., Hernández, J. Z., & Gomez-Perez, J. M. (2012). Open Innovation in an Enterprise 3.0 framework: Three case studies. *Expert Systems with Applications*, 39(10), 8929-8939.

Cardoso, J., Hepp, M. & Lytras, M. D. (Eds.). (2007). *The semantic web: real-world applications from industry*, (vol. 6). Springer Science & Business Media.

Casas-Bayona, A., & Ceballos, H. G. (2014, August 29-31). *Integrating semi-structured information using Semantic Technologies, An Evaluation of Tools and a Case Study on University Rankings Data*. Paper presented at DATA2014: 3rd International Conference on Data Management Technologies and Applications. Vienna University of Technology, Vienna, Austria.

Cataledo, M. (2009, November). *The Semantic Web's the Next Frontier*. American Banker. Retrieved May 5, 2016 from http://www.americanbanker.com/btn/22_11/the-semantic-webs-the-next-frontier-1003476-1.html?pg=1

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, 54(8), 88-98.

Chiari, I. (2007). Introduzione alla linguistica computazionale. GLF editori Laterza.

Chierchia, G., & McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics*. MIT press.

Clancy, J. J. (1999). *The invisible powers: The language of business*. Lexington Books.

Daniushina, Y.V. (2010). Business linguistics and business discourse. Linguística de negócios e discurso de negócios. *Calidoscópio*, 8(3) 241-247.

Davenport, T. H. & Harris, J. G., (2007). *Competing on Analytics: The New Science of Winning*, Harvard Business School Press.

Davenport, T.H. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities.* United States: Harvard Business School Publishing Corporation.

De Bueriis, G., Di Maio, F., Elia, A. & Monteleone, M. (2005). Le poliremathiche dell'italiano. In G. De Bueriis & A. Elia (Eds.) *Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche*. Progetto PRIN 2005, Atlanti Tematici Informatici. Salerno: Plectica.

della Volpe, M. (2013). *Imprese tra Web 2.0 e Big Data. Nuove frontiere per innovazione e competitività*, CEDAM.

della Volpe, M. & Esposito, F. (2016) Incrustation de datos: lo mas moderno de las empresas de tecnologia semantica. *Cultura Latino Americana*, 4 (2): 130-150.

Devoto, G. (1979). *Avviamento alla etimologia italiana* (2 ed.). Milano: Mondadori.

Domingue, J., Fensel, D., & Hendler, J. A. (Eds.). (2011). *Handbook of semantic web technologies*. Springer Science & Business Media.

Donovan, J. J., & Madnick, S. E. (1977). Institutional and ad hoc DSS and their effective use. *ACM SIGMIS Database*, 8(3), 79-88.

Echeverry, C. E. M., Trujillo, M. L., & Giraldo, M. M. (2013). Análisis de la gestión del conocimiento en pymes de Colombia. *REVISTA GTI*, 12(33).

Elia, A. Martinelli, M. & D'Agostino, E. (1981). *Lessico e strutture sintattiche. Introduzione alla sintassi del verbo italiano*. Liguori Editore, Napoli.

Elia, A. & Vietri, S. (2010). Lexis-grammar and Semantic web. *INFOTEKA*, 11(1): 15-38.

Elia, A., Vietri, S., Monteleone, M. & Marano, F. (2010). *Data Mining Modular Software System.* In SWWS2010 – Proceedings of the 2010 International Conference on Semantic Web & Web Services. LAS VEGAS, NEVADA, USA, 12-15 July 2010, p. 127-133, CSREA Press.

Elia, A., Postiglione, A., Monteleone, M. & Monti, J. (2011). *CATALOGA®: a Software for Semantic and Terminological Information Retrieval*. In: Rajendra Akerkar. WIMS '11 Proceedings of the International Conference on Web Intelligence, Mining and Semantics. Sogndal, Norway, 25-27 May, p. n.s., New York, NY, USA: ACM.

Elia, A. (2013). On Lexical, Semantic and Syntactic Granularity of Italian Verbs. In Kakoyianni Doa, F. (ed.), *Penser le lexique-grammaire: perspectives actuelles*, Editions Honoré Champion, Paris, France, pp. 277-288.

Elia, A. (2014). Operatori, argomenti e il sistema "LEG-Semantic Role Labelling" dell'italiano. In Mirto, I., *Relazioni irresistibili*, (p. 105-118), PISA: ETS.

Elia, A., Monteleone, M. & Esposito, F. (2014). Dictionnaires électroniques et lexique des startups. Un exemple d'analyse textuelle automatique. Dictionnaires électroniques et dictionnaires en ligne, *Les Cahiers du dictionnaire*, 6: 43-62.

Esposito, F. & Elia, A. (2016). *NooJ Local Grammars for Innovative Startup Language*, NooJ Conference 2016, [9-11 june 2016] (going to press).

Esposito, F. & della Volpe, M. (2016). *Using Text Mining and Natural Language Processing to support Business Decision: towards a NooJ application*, NooJ Conference 2016, [9-11 june 2016] (going to press).

Feldman, R. & Dagan, I. (1995, August). *KDT knowledge discovery in texts.* Paper presented at the 1st International Conference on Knowledge Discovery (KDD), 95: 112-117.

Feldman, R., & Hirsh, H. (1996, August). *Mining Associations in Text in the Presence of Background Knowledge*. Paper presented at the 2nd International Conference on Knowledge Discovery KDD (pp. 343-346).

Feldman, R., & Hirsh, H. (1997). Exploiting background information in knowledge discovery from text. *Journal of Intelligent Information Systems*, 9(1), 83-97.

Feldman, R., & Sanger, J., (2007). *The Text Mining Handbook*. Cambridge University Press,.

Fleisher, C. & Bensoussan, B. (2002). *Strategic and Competitive Analysis: Methods and Techniques for Analyzing Business Competition* (1st Edition). London, United Kingdom: Pearson.

Ford, E. W. & Wang, Z. (2014). Tackling the Confusing Words of Strategy: Effective Use of Key Words for Publication Impact, Business Management and Strategy, 5 (1).

Gachet, A. (2004). *Building Model-Driven Decision Support Systems with Dicodess*. Zurich, VDF.

Gartner (2015, December). *How to monetize your customer data?* Retrieved January 16, 2016 from http://www.gartner.com/smarterwithgartner/how-to-monetize-your-customer-data/

Gartner Research. (2013). Big Data. Available online at http://www.gartner.com/it-glossary/big-data/ (Accessed on March 20, 2016)

Global Entrepreneurship Monitor & WEF (2015, January). *Leveraging Entrepreneurial Ambition and Innovation: a Global perspective on Entrepreneurship*, Competitiveness and Development, World Economic Forum 2015.

Gobber G., & Morani M., (2014). *Linguistica generale*, McGraw-Hill Education, Milano, Italy

Graham, P. (2012, September). *Want to start a startup?* Retrieved September 26, 2015, from http://www.paulgraham.com/growth.html

Grimes, S. (Accessed on June 24, 2016). *A Brief History of Text Analytics*. B Eye Network. Available online at http://www.b-eye-network.com/view/6311

Gross, M. (1968). *L'emploi des modèles en linguistique*. Langages 9, Paris: Larousse, p.3-8.

Gross, M. (1972). *Mathematical Models of Language*. Englewood Cliffs, N.J.: Prentice-Hall.

Gross, M. (1975b). *Méthodes en syntaxe, régime des constructions complétives*, Paris, Hermann.

Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63, Larousse: Paris,

Gross, M. (1986). *Lexicon-Grammar*. *The Representation of Compound Words*, In AA. VV., COLING-1986. Proceedings, University of Bonn, Bonn, (pp. 1-6).

Gross, M. (1989). *La construction de dictionnaires électroniques*, dans AA. VV., Annales des Télécommunications, CNET: Issy-les-Moulineaux/Lannion, 44(1-2): 4-19

Gross, M. (1975a). *Méthodes en syntaxe*, Hermann: Paris.

Hackathorn, R. D., & Keen, P. G. (1981). Organizational strategies for personal computing in decision support systems. *MIS quarterly*, 21-27.

Harris, Z.S. (1954). Distributional Structure. WORD 10:146–162. Reprinted in Fodo, J. and Katz, J., The structure of language: Readings in the philosophy of language, Prentice-hall, 1964.

Harris, Zellig S. (1991). *A Theory of Language and Information: A Mathematical Approach*. Oxford & New York: Clarendon Press.

Hodgson, G. M. (1996). Corporate Culture and the Nature of the Firm. In *Transaction Cost Economics and Beyond* (pp. 249-269). Springer Netherlands.

Hoskisson, R. E., Hitt, M. A., Wan, W. P. & Yiu, D. (1999). Theory and research in strategic management: Swings of a pendulum. *Journal of management*, 25(3), 417-456.

IBM Corporation (n.d.). *What is Big Data? Bringing Big Data to the Enterprise*. Retrieved December 20, 2015 from https://www-01.ibm.com/software/au/data/bigdata

Jackendoff, R., & Jackendoff, R. S. (1983). *Semantics and cognition* (Vol. 8). MIT press.

Kao, A., & Poteet, R.S. (Eds) (2007). *Natural Language Processing and Text Mining*, Springer-Verlag: London.

Kilgarriff, A. (2001, March). *Web as corpus*. In Proceedings of Corpus Linguistics 2001 (pp. 342-344). Corpus Linguistics. Readings in a Widening Discipline.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3), 333-347.

Kodratoff, Y. (1999, June). *Knowledge discovery in texts: a definition, and applications.* In International Symposium on Methodologies for Intelligent Systems (pp. 16-29). Springer Berlin Heidelberg.

Kopackova, H., Komarkova, J. & Sedlak, P. (2008). Decision making with textual and spatial information. *WSEAS Transactions on Information Science and Applications*, vol. 5(3), p.259.

Kowal, S. & O'connell, D. C. (2004). 5.9 The Transcription of Conversations. A Companion to, p. 248.

Laporte, E. (2005). *In memoriam Maurice Gross*. Archives of Control Sciences, 15(3), 257-278.

Leontiades, M. (1982). The confusing words of business policy. *Academy of Management Review*, 7(1): 45-48.

Loh, S., Wives, L. K., & de Oliveira, J. P. M. (2000). Concept-based knowledge discovery in texts extracted from the web. *ACM SIGKDD Explorations Newsletter*, 2(1), 29-39.

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet R. (January 2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, Elsevier, Available from http://amzn.to/textmine

Mintzberg, H. (1978). Patterns in strategy formation. *Management science*, 24(9): 934-948.

Moore, G. A., & McKenna, R. (1999). *Crossing the Chasm: Marketing and selling high-tech products to mainstream customers* (Преодоление разрыва: маркетинг и продажа высокотехнологичных товаров массовому потребителю).

Nesselhauf, N. (2005). *Corpus Linguistics: a practical introduction*. Retrived November 25, 2016, from http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf

Nicolai, A. T., Dautwiz, J. M. (2009). Fuzziness in Action: What Consequences Has the Linguistic Ambiguity of the Core Competence Concept for Organizational Usage? *British Journal of Management*, 21: 874–888. Available online at http://dx.doi.org/10.1111/j.1467-8551.2009.00662.x

Niu, L., Lu, J., & Zhang, G. (2009). Cognition-driven decision support for business intelligence. Models, Techniques, Systems and Applications. *Studies in Computational Intelligence*, Springer, Berlin.

Nyce, C. (2007). *Predictive Analytics White Paper* (PDF), American Institute for Chartered Property Casualty Underwriters, Insurance Institute of

America. Retrieved April 3, 2016 from http://www.hedgechatter.com/wp-content/uploads/2014/09/predictivemodelingwhitepaper.pdf

Oettinger, A. G. (1965). *Computational Linguistics.* The American Mathematical Monthly, part 2: Computers and Computing,72 (2): 147-150.

Pardelli, G. & Biagioni S. (2013). Quando la linguistica incontra l'informatica: una riflessione terminologica, SCIRES-IT, SCIentific RESearch and Information Technology, *Ricerca Scientifica e Tecnologie dell'Informazione* 3(1), 67-78.

Piraquive, F. N. D., Aguilar, L. J., & García, V. H. M. (2009). Taxonomía, ontología y folksonomía,¿ qué son y qué beneficios u oportunidades presentan para los usuarios de la web?. *Universidad & Empresa*, 11(16), 242-261.

Polikoff, I. & Allemang, D. (2003, September). *Semantic technology*. TopQuadrant Technology Briefing, 1.1.

Power, D. J. (2000). Web-based and model-driven decision support systems: concepts and issues. *AMCIS 2000 Proceedings*, p. 387.

Power, D. J. (2001, June). *Supporting decision-makers: An expanded framework.* In e-Proceedings Informing Science Conference, Krakow, Poland (pp. 431-436).

Power, D. J. (2008). Decision support systems: a historical overview. In *Handbook on Decision Support Systems* 1, pp. 121-140. Springer Berlin Heidelberg.

Power, D.J. (2003). The above response is based upon Power, D., What are the characteristics of a Decision Support System? *DSS News*,. 4, No. 7, March 30.

Radimský, J. (2012). Actants, arguments et rôles sémantiques: combien de niveaux d'analyse?. In Tomaszkiewicz, T. & Vetulani, G., *L'apport linguistique et culturel français à l'Europe*, Volume: du passé aux défis de l'avenir, Lask, Leksem, p. 97-103.

Radovanovic, M. & Ivanovic, M. (2008). Text mining: approaches and applications, Novi Sad "Journal of Math", 38 (3): 227-234.

Raghupathi W. (2010). Data Mining in Health Care. In Kudyba S. (Ed) *Healthcare Informatics: Improving Efficiency and Productivity* (pp. 211-223). United Kingdom: Taylor & Francis Group LLC.

Riediger, H. (2012). *Che cos'è la terminologia e come si fa un glossario*. Retrived online November 22, 2016, from http://www.terminator.it/corso/doc/mod3_termino_glossa.pdf

Robehmed, N. (n.d.)*"What Is A Startup?". Forbes*. Retrieved 30, April 2016 from http://www.forbes.com/sites/natalierobehmed/2013/12/16/what-is-a-startup/#57a38b894c63 (2013, December 16).

Rogers, E. M. (1962). *Diffusion of innovations*. Simon and Schuster.

Ronda-Pupo, G. A. & Guerras-Martin, L. Á. (2012). Dynamics of the evolution of the strategy concept 1962–2008: a co-word analysis. *Strategic Management Journal*, 33: 162–188. Available online at http://dx.doi.org/10.1002/smj.948

Ryan, C.D. (2013). *Innovation in Agri-food Clusters: Theory and Case Studies*, CAB Books, CABI.

Searle, J. R. (1969). Speech acts: An essay in the philosophy of language, 626. Cambridge university press.

Segre, C. (1981). Testo, in *Enciclopedia Einaudi*, Einaudi, Torino 1981, vol. XIV, pp. 276-277.

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The Semantic Web Revisited, *IEEE Intelligent Systems*, 21 (3): 96-101.

Silberztein, M. (2003). *Nooj Manual.* Available online at http://www.nooj4nlp.net/NooJManual.pdf.

Silberztein, M. (2012). Corpus Linguistics and Semantic Desambiguation, in G. Maiello, R. Pellegrino Eds. *Database, Corpora, Insegnamenti Linguistici*. Linguistica , 63, Schena Editore/Alain Baudry et C.ie, pp. 397-410.

Silberztein, M. (2013). *NooJ Computational Devices*. Formalising Natural Languages with NooJ 2013: Selected Papers from the NooJ 2013 International Conference (Saarbrucken, Germany). Edited by Svetla Koeva, Slim Mesfar and Max Silberztein. Cambridge Scholars Publishing, Newcastle., UK: 01-14.

Silberztein, M. (2014). *NooJ V4.* Formalising Natural Languages with NooJ 2013: Selected Papers from the NooJ 2013 International Conference (Saarbrucken, Germany). Edited by Svetla Koeva, Slim Mesfar and Max Silberztein. Cambridge Scholars Publishing, Newcastle., UK: 01-12.

Silberztein, M. (2015). *Analyse et génération transformationnelle avec NooJ*. Elia A., Iacobini C.; Voghera M. (eds.), 2015, Proceedings of the 47th annual meeting of the Italian linguistic Society “Livelli di Analisi e Fenomeni di Interfaccia”, Rome, Bulzoni,.

Silberztein, M. (2015). *La formalisation des langues: l'approche de NooJ*. ISTE Ed: Londres.

Sprague, R. H. & Carlson, E. D. (1982). Building Effective Decision Support Systems. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Stephens, S. (2007). The Enterprise Semantic Web. *The Semantic Web* (pp. 17-37). US: Springer.

Studer, P. (2013). Linguistics applied to business contexts: an interview with Patrick Studer. *ReVEL*, 11 (21): 187-202.

Tan, A. H. (1999, April). *Text mining: The state of the art and the challenges*. In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases 8: 65-70.

Tesnière, L. (1953*). Esquisse d'une syntaxe structurale*. Klincksieck, Paris.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck, Paris.

Turban, E., Aronson, J.E., & Liang, T.P. (2005). *Decision support systems and intelligent systems*, 7th edn. Pearson prentice Hall, New Jersey.

Turney P.D. (2002, July) *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews* (pp.417-424). Paper presented at the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, USA.

Vaccà, S. (1986). L'economia delle relazioni tra imprese: dall'espansione dimensionale allo sviluppo per reti esterne. *Economia e politica industriale*, 51, 3-41.

Vietri, S. (2008). *Dizionari elettronici e grammatiche a stati finiti. Metodi di analisi formale della lingua italiana*, Salerno, Plectica.

Vietri, S. & Monteleone, M. (2014). *The NooJ English Dictionary*. In Formalising Natural Languages with NooJ 2013: Selected papers from the NooJ 2013 International Conference Pag.69-86 12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, Cambridge Scholars Publishing.

Wallis, S. & Nelson G. (2001). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery*, 5: 307–340.

Witten, I. H. (2005). Text mining. *Practical handbook of Internet computing*, 14-1.

Wittgenstein, L. (1953). *Philosophical investigations*. Philosophische Untersuchungen.