Complexity Analysis in the Language of Information Technology Companies

Mary Luz Mouronte López
Departamento de Ingeniería Telemática,
Universidad Carlos III de Madrid, Escuela Politécnica Superior
Av. Universidad 30, Edif. Torres Quevedo, Madrid 28911, Spain
mmouront@it.uc3m.es

***Abstract*—This paper uses a large amount of data, which, in turn, will interpret a broad spectrum of information. It applies the graph theory to map relationships among words using *network science* in order to analyze the reports of important telecommunication companies and to study the syntactic structure of their language. The following properties are analyzed in the word network of these companies, such as the words' co-occurrence, density, betweenness, distance between words (length, average diameter) and structures (presence of clusters, existence of motifs). We conclude that the company language shows characteristics of complex systems and that some features are common among organizations.**

Keywords-words network; mean distance; communities; motif; robustness.

I. INTRODUCTION

A language is a set of signs that allow human communication by means of their meaning and their relationships. It is different in creativity, form and content for several cultural groups because they do not use the same sounds, or grammatical structure [1]. Many tools for natural language processing have been developed due to language analysis, which is useful in different areas;

- The Internet has created many electronic documents, which are widely available. In order to recover the most relevant data, it is necessary to check these documents by means of language processing, a keyword-match based on data recovery, which supplies quick access to documents containing important information.
- In social marketing, understanding what the community is saying is a key issue. A study about abbreviations or variants of expression usage (small or capital letters, letter repetition or missing letters, orthographical variations, or non-alpha symbols) would be interesting.
- In medicine, the linguistic analysis may be useful in the investigation of mental illnesses. For instance, some studies have confirmed that schizophrenics have a syntactically less complex speech, which means that linguistic variable usage in a discriminant function analysis may help to predict diagnoses in many cases.

The objective of this paper is to study the syntactical structure of the language in information technology companies by means of *network science*. In [5], network science is defined as "the study of networks which contrasts, compares, and integrates techniques and algorithms developed in disciplines as diverse as mathematics, statistics, physics, social network analysis, information science and computer science".

This investigation studies more than 30 company reports with more than 6,000,000 words to interpret a broad spectrum of information. Specifically, the annual reports (in Spanish) of five companies are analyzed in: AMPER [23], JAZZTEL [24], VODAFONE [25], INDRA [26] and TELEFÓNICA [27].

We transform each company's annual report into a words network where some metrics such as: shortest distance between nodes, betweenness, detection of clusters and motifs are studied. These measures have also been used to analyze the structural properties of others natural and artificial complex systems [2] [3] [4]. Typical language properties are identified and the obtained results are compared among the analyzed companies.

We apply the method described in [6] to detect motifs and the Walktrap Algorithm [7] [22] to calculate clusters in the networks.

We develop several tools in C and phyton languages to analyze the language structure syntactically in the annual reports.

The rest of the paper is organized as follows: Section 2 shows the related work, Section 3 describes the method of analysis and results, and Section 4 establishes the main conclusions and the future work.

II. RELATED WORK

There are no previous works on applications about communication pattern analysis of information technology companies. Nevertheless, there are several works about the language characteristics.

Ferrer i Cancho and Sole [8] show that language organization, described in terms of a graph, displays two important features found in a disparate number of complex systems: (i) The so-called small-world effect. In particular, the average distance between two words, $< l >$, is shown

 ISBN: 978-1-61208-242-4

as $< l > \simeq 2^3$. (ii) A scale-free distribution of degrees and the fact that disconnecting the most connected nodes in such networks can be identified in some language disorders. The authors claim that these observations indicate some unexpected features of language organization that might reflect the evolution and social history of lexicons and the origins of their flexibility and combinatorial nature.

Steyversa and Tenenbaumb [9] present statistical analysis of the large-scale structure of 3 types of semantic networks: word associations, wordnet, and rogets thesaurus. This research shows that they have a small-world structure, characterized by sparse connectivity, short average path lengths between words, and strong local clustering. . The authors describe a simple model for semantic growth, in which each new word or concept is connected to an existing network by differentiating the connectivity pattern of an existing node. This model generates appropriate small-world statistics and power-law connectivity distributions, and also suggests one possible fundamental basis for the effects of learning history variables on behavioral performance in semantic processing tasks.

Markosova[10] revises recent studies of syntactical word web. He presents a model of growing network in which processes such as node addition, link rewiring and new link creation are taken into account. The author argues that this model is a satisfactory minimal model explaining measured data.

Freeman and Barnett, [11] try to identify, which characteristics of the language are used in the written messages sent to the employees in a manufacturing company of medical equipment. Authors of the framework focus their research on organizational culture.

Coronges [12] uses a network analysis approach to provide information that helps to compare structural indexes of associative organization of two populations varying in age and city location; associative connections between words reveal the organization of concepts in these populations.

Brasethvik and Atle [13] present an approach to semantic document classification and retrieval based on natural language analysis and conceptual modeling. A conceptual domain model is used in combination with linguistic tools to define a controlled vocabulary for document collection.

Our research studies the structure of the company language by means of the network science and gets interesting conclusions. The investigation shows that some features are common among organizations.

III. Analysis Method

We analyze chairman's letters of annual reports for the following companies and periods: TELEFÓNICA (2009, 2008, 2007, 2006), AMPER (2009, 2008, 2007, 2006), INDRA (2009, 2008, 2007, 2006), VODAFONE (2009,2008, 2007, 2006) and JAZZTEL (2009, 2008, 2005, 2004). Chairman's letters are written in Spanish.

A. Design of word network

A text can be plotted as a graph $G = (U, L)$; where $U = \{i\}_{(i=1,\ldots,N)}$ is the set of N words and $L = \{i, j\}$ is the set of connections between them. Adjacent words are defined as a couple i, j, which pertains to G , and where a binary relation or link exists.

There are several assumptions made, including: repeated words correspond to the same node i, word network is not case sensitive, and the words before and after a score are not neighbours in the network.

B. Words' co-ocurrence

Table 1 shows the words with a percentage of co-occurrence bigger than 1 % for the five companies studied. It should be noted that words in English are written in italics, different words in Spanish can be translated to the same word in English, and those words in bold are repeated. Prepositions, conjunctions, articles and adjectives have the higher percentage. One can notice that some words appear in more than one company's reports, for instance *"millones"* (*"millions"*) in JAZZTEL and TELEFÓNICA; *"compañía"* (*"company"*) in AMPER and TELEFÓNICA , *"desarrollo"* (*"development"*); in INDRA and VODAFONE. The company's own name appears in letters from AMPER, INDRA, JAZZTEL and TELEFÓNICA.

A matrix B of $5x5$ dimension can be built as a representation of the coincidence of words between companies. The element B_{tq} is the coincidence percentage (%) in the texts between t and q companies. $t, q = 1$ for TELEFÓNICA; $t, q = 2$ for AMPER; $t, q = 3$ for INDRA; $t, q = 4$ for JAZZTEL; and $t, q = 5$ for VODAFONE. We show below that the main similarities are: TELEFÓNICA and AMPER in 60.71%, AMPER and INDRA in 86.36 %, and JAZZTEL and VODAFONE in 77.78 %.

$$B = \begin{pmatrix} 100 & 60.71 & 57.14 & 39.29 & 50 \\ 60.71 & 100 & 86.36 & 40.91 & 63.64 \\ 57.14 & 86.36 & 100 & 32 & 64 \\ 39.29 & 40.91 & 32 & 100 & 77.78 \\ 50 & 63.64 & 64 & 77.78 & 100 \end{pmatrix}$$

C. Structural parameters calculation

1) General characteristics: We have calculated the average distances between nodes ($< l >$), average degree ($< k >$) and the most distant vertices (d) in the word networks. These parameters are similar to those of other complex technological networks ([11], [14], [15]) and close to the value obtained in random networks, where the small world property has been likewise reported [15]. We show these characteristics in Table 2 and Table 3.

 ISBN: 978-1-61208-242-4

Table I
CO-OCCURRENCE OF WORDS IN AMPER, INDRA, JAZZTEL, TELEFÓNICA AND VODAFONE COMPANYS

	Words (%)
AMPER	**de,** ***of*** (14.07 %); **la,** ***the*** (4.55 %); **que,** ***that*** (4.22 %); **el,** ***the*** (2.93 %); **las,** ***the*** (1.79 %); **a,** ***to***(4.22 %); **ha,** ***has*** (2.05 %); **en,** ***in*** (4.09 %); **y,** ***and*** (3.58 %); **un,** ***a*** (2.44 %); **con,** ***with*** (2.81 %) ; **del,** ***of*** (2.60 %); **los,** ***the*** (2.28 %);**nos,** ***us*** (1.30 %); **por,** ***for*** (2.28 %); **para,** ***to*** (1.02 %) ; **una,** ***a*** (1.95 %); **nuestra,** ***our*** (1.14 %); **nuestro,** ***our*** (1.14 %); AMPER (1.28); **euros,** ***euros*** (1.14 %); **compañia,** ***company*** (1,46 %);
INDRA	**de,** ***of*** (11.34 %); **y,** ***and*** (8.96 %); **en,** ***in*** (7.31 %); que, *that* (5.60 %); **la,** ***the*** (4.78 %); **los,** ***the*** (3.80 %); **a** ***to*** (3.36 %); **los,** ***the*** (3.02 %); **el,** ***the*** (2.92 %); **un,** ***a*** (2.05 %); **nuestro,** ***our*** (1.79 %); INDRA (1.79 %); **con,** ***with*** (1.75 %); **las,** ***the*** (1.34 %); **nuestra,** ***our*** (1.34 %) ; futuro, *future* (1.01 %); **del,** ***of*** (1.46 %); más,*more* (1.46 %); **una,** ***a*** (1.34 %) ; **nuestros,** ***our***(1.19 %); ha, *has* (1.19 %);**desarrollo,** ***development*** **%** (1.17 %); **para,** ***to*** (1.17 %); **por,** ***for*** (1.12 %)
JAZZTEL	**de,** ***of*** (15.17 %); a, *to* (3.48 %) ; **y,** ***and*** (3.62 %); **el** (3.62 %); JAZZTEL (2.84 %); **la,** ***the*** (2.84 %); **con,** ***with*** (2.28 %); **un,** ***a*** (1.74 %); **para,** ***to*** (1.74 %); **euros,** ***euros*** (1.74 %); **se,** ***is*** (1.52 %); **por,** ***for*** (1.45 %); **una,** ***a*** (1.42 %); **clientes,** ***clients*** (1.14 %); mercado, *market* (1.45 %); **millones,** ***millions*** (1.14 %); (1.14 %); equipo, *team* (1.14 %); **nuestros,** ***our*** (1.05 %); **nos,** ***us*** (1.05 %)
TELEFÓNICA	**de,** ***of*** (22.16 %); **en,** ***in*** (9.06 %); **y,** ***and*** (7.48 %); un, *a* (3.17 %); **del,** ***of*** (2.88 %); **con,** ***with*** (2.87 %); **los,** ***the*** (2.50 %); **una,** ***a*** (2.37 %); **se,** ***is*** (2.30 %); **por,** ***for*** (2.19 %); TELEFÓNICA (2.15 %); **ha,** ***has*** (1.73 %); como, *like* (1.69 %); es, *is* (1.58 %); al, *to the* (1.58 %); **compañia,** ***company*** (1.43 %); crecimiento, *growth* (1.29 %); **millones,** ***millions*** (1.29 %); resultados, *results* (1.19 %); sector, *area* (1.18 %); no, *not* (1.01 %); **clientes,** ***clients*** (1.01 %)
VODAFONE	**de,** ***of*** (13.57 %); **y,** ***and*** (6.58 %); **en,** ***in*** (5.91 %) ; **el,** ***the*** (4.80 %); la, *the* (3.60 %); **que,** ***that*** (3.29 %);**los,** ***the*** (2.80 %); a, *to* (2.06 %); actuaciones, *actions* (2.06 %); **un,** ***a*** (1.87 %); más, *more* (1.82 %); **por,** ***for*** (1.36 %); **para,** ***to*** (1.36 %); **desarrollo,** ***development*** (1.23 %) ; **del,** ***of*** (1.20 %); **las,** ***the*** (1.20 %) ;este, *this* (1.20 %); **una,** ***a*** (1.12 %); servicios, *services* (1.12 %);

Table II
GENERAL CHARACTERISTICS IN TELEFÓNICA, AMPER, INDRA, JAZZTEL AND VODAFONE COMPANIES

		2009	2008	2007	2006	2005	2004
TELEFÓNICA	Density	0.0070947	0.0076476	0.0094565	0.0080372		
	$< k >$	4.91	4.55	4.79	4.48		
	$< l >$	3.36	3.35	3.28	3.43		
	d	10	9	8	9		
AMPER	Density	0.0112201	0.0159254	0.0115064	0.0092859		
	$< k >$	3.98	3.25	3.53	3.62		
	$< l >$	3.61	3.78	3.63	3.498		
	d	10	11	11	8		
INDRA	Density	0.0112644	0.0111099	0.0122924	0.0146433		
	$< k >$	3.83	3.73	3.69	3.90		
	$< l >$	3.52	3.64	3.55	3.51		
	d	10	11	9	8		
JAZZTEL	Density	0.0166325	0.0202877			0.0132935	0.0128408
	$< k >$	3.48	2.76			3.48	3.67
	$< l >$	3.56	4.23			3.63	3.82
	d	8	10			7	9
VODAFONE	Density	0.0145563	0.0159402	0.0139675	0.0124659		
	$< k >$	3.52	3.49	3.72	3.10		
	$< l >$	3.60	3.57	3.41	4.07		
	d	9	9	8	10		

2) Betweenness: The betweenness b_i of a node i in a network is related to the number of times that such node is a member of the set of shortest paths, which connect all the pairs of nodes in the network. If g_{nl} is the total number of possible paths from *n* to *l* nodes, and g_{nil} is the number of paths from *n* to *l* passing through *i*, then g_{nil}/g_{nl} is the proportion of paths from *n* to *l* that pass through *i*. The betweenness for a node *i* is defined as: $b_i = g_{nil}/g_{nl}$.

The functional relevance of the betweenness centrality b_i of a node is based on the observation that a node located on the shortest path between two other nodes has a greater influence over the information transfer between them. The highly connected nodes (hubs) must have high-betweenness values because there are many nodes directly and exclusively connected to these hubs and the shortest path between these nodes goes through these hubs.

Table III
GENERAL CHARACTERISTICS IN RANDOM NETWORKS

		2009	2008	2007	2006	2005	2004
TELEFÓNICA	$< k >$	4.91	4.55	4.79	4.48		
	$< l_{rand} >$	4.21	4.24	4.07	4.40		
	d_{rand}	8	10	9	9		
AMPER	$< k >$	3.98	3.25	3.53	3.62		
	$< l_{rand} >$	4.18	4.21	4.85	4.71		
	d_{rand}	9	9	12	10		
INDRA	$< k >$	3.83	3.73	3.69	3.90		
	$< l_{rand} >$	4.35	4.21	4.65	4.31		
	d_{rand}	11	9	11	10		
JAZZTEL	$< k >$	3.48	2.76			3.48	3.67
	$< l_{rand} >$	4.60	4.65			4.54	4.44
	d_{rand}	10	10			10	9
VODAFONE	$< k >$	3.52	3.49	3.72	3.10		
	$< l_{rand} >$	4.45	4.37	4.29	4.73		
	d_{rand}	9	9	8	11		

In all studied networks, values for betweenness b are ranged over several orders of magnitude and there are low-connectivity and high-connectivity nodes, which exhibited a wide range of betweenness values. It is shown in Figure 1 for TELEFONICA and AMPER company in 2009, where betweenness, b, is plotted as a function of connectivity (degree), k. This result indicates the existence of a large number of nodes with high betweenness but low connectivity. Although the low connectivity of these words would imply that they are unimportant, their high betweenness suggests that these words may have a global impact. From a topological point of view, these words are positioned to connect regions of high clustering (containing hubs), even though they have low local connectivity. The existence of such words points to the presence of modularity in the network, and therefore suggests that these words may represent important connectors that link these modules. This behaviour was also found in the yeast protein interaction network [16].

Articles, pronouns, prepositions, conjunctions, adverbs and adjectives appear in the highest positions in the ranking and these words are similar in all texts. The first appearance of verbs and sustantives is lower in the ranking for all reports.

Depending on the year, in TELEFÓNICA and AMPER companies, the first ranked verb is *"ser"* (*"to be"*) or *"haber"* (*"to have"*), in JAZZTEL company the verb *"dudar"*, (*"to hesitate"*) also appears in that position; the first verb is *"haber"* (*"to have"*) in INDRA company and in VODAFONE company is *"suponer"* (*"to suppose"*), *"contribuir"* (*"to contribute"*), *"haber"* (*"to have"*) and *"descargar"* (*"to download"*). The first substantive in TELEFÓNICA company is *TELEFÓNICA*; in AMPER company the first nouns are *"mercado"* (*"market"*), *AMPER* and *"año"* (*"year"*); whereas in JAZZTEL company is

Table IV
DECREASING RANKING REGARDING TO BETWEENNESS IN TELEFÓNICA'S REPORT

Rank	2009	2008	2007	2006
1	de *of*	de *of*	de *of*	de *of*
2	la *the*	en *in*	y *and*	en *in*
3	que *that*	y *and*	el *the*	y *and*
4	en *in*	la *the*	en *in*	el *the*
5	y *and*	que *that*	la *the*	la *the*
6	el *the*	un *a*	del *of*	que *that*
7	a *to*	los *the*	que *that*	un *a*
8	los *the*	el *the*	un *a*	a *to*
9	las *the*	TELEFÒNICA	a *to*	las *the*
10	con *with*	del *of*	para *to*	por *for*
11	una *a*	a *to*	los *the*	una *a*
12	un *a*	con *with*	las *the*	nuestra *our*
13	para *to*	las *the*	con *with*	los *the*
14	TELEFÓNICA	para *to*	TELEFÓNICA	TELEFÓNICA
15	del *of*	por *for*	clientes *clients*	del *of*
16	su *its*	más *more*	crecimiento *growth*	más *more*
17	se *is*	su *its*	una *a*	con *with*
18	ha *has*	es *is*	al *to the*	es *is*
19	es *is*	...	2007	se *is*
20	más *more*	...	por *for*	crecimiento *growth*
21	al *to the*	...	como *like*	para *to*
...	...	...	...	...
27	...	...	ha *has*	
...	...	...	...	...
35	...	...	...	ha *has*
...	...	...	...	...

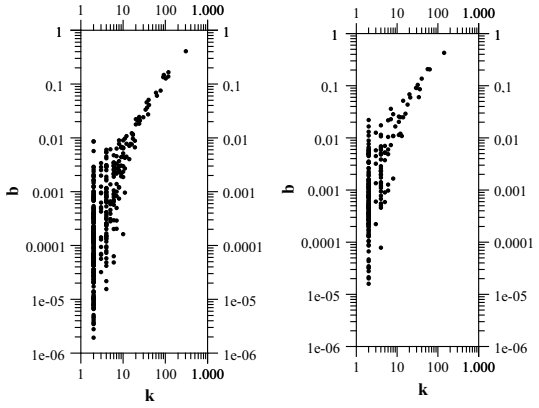

Figure 1. Degree (k) versus betweenness (b) plotted in logarithmic scale for 2009 reports in TELEFÓNICA company (left) and AMPER company (right)

"premios" (*"awards"*), *JAZZTEL* and *"mercado"* (*"market"*); in INDRA company the first sustantive is *"futuro"* (*"future"*), *"fruto"* (*"to bear fruit"*), *"confianza"* (*"confidence"*), *"soluziona"* (*without translation, project name*); and in VODAFONE company the first noun is *"actuaciones"* (*"actions"*), *"desarrollo"* (*"development"*)and *"servicios"* (*"services"*). We show these results in Table 4. It should be emphasized that words in English are written in italics and different words in Spanish can be translated to the same word in English.

3) Communities: Different methods have been developed in order to find communities in networks. Basically, these methods can be grouped as spectral methods (e.g., [18]), divisive meth-ods (e.g., [19]), agglomerative methods (e.g., [20]), and local methods (e.g., [21]). The choice of the best method depends of the specific application, including the network size and number of connections.

We have carried out a study of the community structure in the word networks by measuring the similarities between nodes by means of Walktrap Algorithm [7] [22]. Walktrap algorithm is an agglomerative approach to detect communities, which starts with a community for each node such that the number of partitions $|\rho| = n$ and build communities by amalgamation.

Walktrap method uses random walks on G to identify communities. At each step in the random walk, the walker is at a node and moves to another node chosen at random yet uniformly from its neighbors. The sequence of visited nodes is a Markov chain where the states are the nodes of G. An N x N dimension adjacency matrix $A(G)$ can be built as a bidimensional representation of the relationships between words, where $A_{ij} = 1$ when a connection between the nodes i and j exists and $A_{ij} = 0$ otherwise. At each step the transition probability from node i to node j is $P_{ij} = \frac{Aij}{ki}$, which is an element of the transition matrix P for the random walk. We can also write $D^{-1}A$, where D is the diagonal matrix of the degrees ($\forall_i, D_{ii} = k_i$ and $D_{ij} = 0$ where $i \neq j$. The random walk process is driven by powers of P: the probability of going from *i* to *j* in a random walk of length t is $(P^t)_{ij}$, which we will denote simply as P^t_{ij} . All of the transition probabilities related to node *i* are contained in the i^{th} row of P_t denoted as $P^t_{i\bullet}$. Then we define an inter-node distance measure as:

$$s_{ij} = \sqrt{\sum_{q=1}^{n} \frac{(P^t_{iq} - P^t_{jq})^2}{k_q}} = \| D^{1/2}P^t_{i\bullet} - D^{1/2}P^t_{j\bullet} \| \quad (1)$$

where $\| \bullet \|$ is the Euclidean norm of R^n . This distance can also be generalized as a distance between communities: s_{CiCj} or as a distance between a community and a node: s_{Cij}

We then use this distance measure in our algorithm. The algorithm uses an agglomerative approach, beginning with one partition for each node ($|\rho| = n$) . We first compute the distances for all adjacent communities (or nodes in the first step). In each step α , two communities are chosen based on the minimization of the mean σ_α of the squared distances between each node and its community.

$$\sigma_\alpha = \frac{1}{n}\Sigma_{C_i \in \rho_\alpha}\Sigma_{i \in C_i} s^2_{iC_i} \quad (2)$$

Instead of directly calculating this quantity, we first calculate the variations $\Delta\sigma_\alpha$. Due to the fact that the algorithm uses a Euclidean distance, we can efficiently calculate these variations as

$$\Delta\sigma(C_1, C_2) = \frac{1}{n}\frac{|C_1||C_2|}{|C_1| + |C_2|} s^2_{C_1C_2} \quad (3)$$

The community merges when the lowest $\triangle\sigma$ is performed. The transition probability matrix is then updated accordingly.

$$P^t_{(C_1 \cup C_2)\bullet} = \frac{|C_1|P^t_{C_1\bullet} + |C_2|P^t_{C_2\bullet}}{|C_1| + |C_2|} \quad (4)$$

and the process is repeated again, updating the values of s and $\Delta\sigma$ and then performing the next merge. After $n-1$ steps, we get one partition that includes all the nodes of the network $\rho_n = \{N\}$. The algorithm creates a sequence of partitions $(\rho_\alpha)_{1\leq\alpha\leq n}$. Finally, we use modularity to select the best partition of the network, calculating Q_{ρ_α} for each partition and selecting the partition that maximizes modularity.

We define modularity Q as the fraction of links within communities minus the expected value of the same quantity for a random network. Let A_{ij} be an element of the networks adjacency matrix and suppose the nodes are divided into communities such that node i belongs to community C^i. Then Q can be calculated as follows:

$$Q = \frac{1}{2m}\sigma_{ij}\{A_{ij} - \frac{k_i k_j}{2m}\}\delta_{C_i C_j} \quad (5)$$

where the $\delta_{C_i C_j}$ function is 1 if $C_i = C_j$ and 0 otherwise. m is the number of links in the graph, and k_i is the degree of a node i. The sum of the term $\frac{k_i k_j}{2m}$ over all pair nodes in a community represents the expected fraction of links within that community in an equivalent random network where node degree values are preserved.

All word networks have several main communities, which have high connected nodes but with few connections to the rest of the network. These so clearly defined communities suggest a network structure. The community rank by percentage of words is shown in Table 6. In Figure 2, we plot the community rank for TELEFÓNICA's reports in 2006 and 2007. We detected 88 communities in 2006 and 42 communities in 2007 within this company. In Figure 3, we display the community rank for INDRA's reports in 2008 and 2009; 33 communities in 2008 and 39 communities in 2009 were found.

In each community, we also detected that the higher probability of appearance according to the type of word occurs among nouns, verbs and adjectives. In Figure 4, we show the percentage by the type of word in rank 1 community and in rank 2 community for INDRA's report in 2007.

4) Motifs: Network motifs are connectivity-patterns (subgraphs) that occur frequently in the networks. Most networks studied in biology, ecology, communication and others fields have been found to show a small set of network motifs; in most cases these motifs are repeated. In [22] a network motif is defined as "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks"

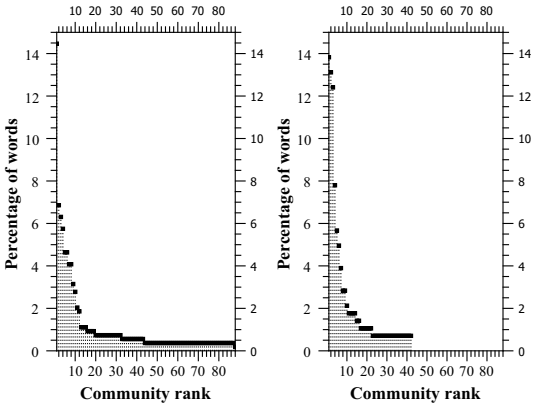

Figure 2. Left, community rank by percentage of words over 88 communities detected for TELEFÓNICA's report in 2006. Right, community rank by percentage of words over 42 communities found for TELEFÓNICA's report in 2007.

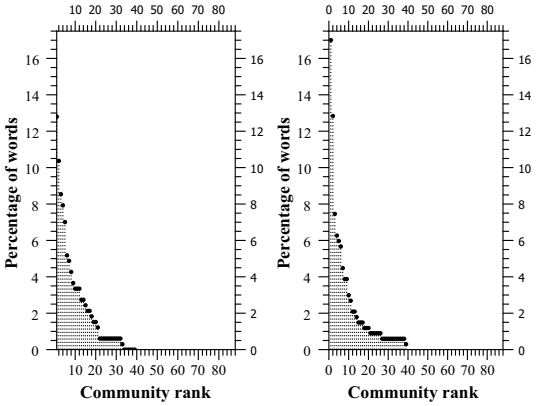

Figure 3. Left, community rank by percentage of words over 33 communities detected for INDRA's report in 2008. Right, community rank by percentage of words over 39 communities found for INDRA's report in 2009.

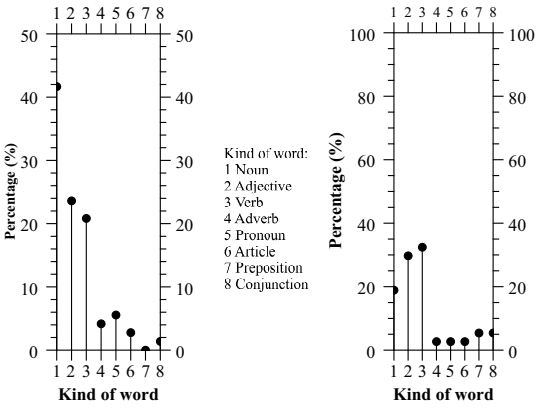

Figure 4. Percentage by kind of word in rank 1 community (left) and in rank 2 community (right) for INDRA's report in 2007.

Table V
COMMUNITY RANK BY PERCENTAGE OF WORDS IN TELEFÓNICA'S, AMPER'S, INDRA'S, JAZZTEL'S AND VODAFONE'S REPORTS

		2009	2008	2007	2006	2005	2004
TELEFÓNICA	Number of communities	83	79	42	88		
	Rank						
	1	11.35 %	10.25 %	13.83 %	14.47 %		
	2	8.59 %	9.74 %	13.12 %	6.86 %		
	3	7.71 %	8.89 %	12.41 %	6.31 %		
	4	6.55 %	6.67 %	7.80 %	...		
AMPER	Number of communities	40	21	31	22		
	Rank						
	1	18.29	16.49	13.49	16.62		
	2	14.29	15.46	8.22	15.83		
	3	11.14	13.92	7.57	12.66		
	4	9.14	7.22	6.25	11.35		
	5	6.57	6.70	...	8.44		
	6	...	...	...	6.33		
INDRA	Number of communities	39	33	30	31		
	Rank						
	1	17.01	12.80	26.03	17.62		
	2	12.84	10.37	12.33	16.86		
	3	7.46	8.54	10.96	8.43		
	4	6.27	7.93	...	6.90		
	5	...	7.01	...	...		
JAZZTEL	Number of communities	28	28			28	26
	Rank						
	1	14.57	8.40			15.32	20.14
	2	13.07	7.63			9.68	11.51
	3	12.56	6.87			7.26	9.35
	4	10.05	...			...	8.99
	5	8.54	...			...	7.91 %
	6	...	...			...	6.47
	7	...	...			...	6.12
VODAFONE	Number of communities	21	24	32	24		
	Rank						
	1	14.96	22.17	15.02	19.83		
	2	6.84	11.79	13.04	10.33		
	3	6.41	11.32	12.25	9.92		
	4	...	10.85	11.07	8.26		
	5	...	7.08	...	7.02		
	6	...	...	...	6.61		

The procedure used to detect network motifs in all word networks is described in [22]. This method samples the sub-graphs of the size n and estimates the appearances of them in the whole graph based on the frecuencies obtained in the samples. A sub-graph is sampled using a simple iterative procedure by selecting connected links until a set of n nodes is reached. The process is as follows:

L_S is the set of picked links.
N_S is the set of all nodes that are touched by the links in L_S.
L_S and N_S are inited to be empty sets.

1) Pick a random edge $l_1 = (h, i)$. Update $L_S = \{l_1\}$, $N_S = \{h, i\}$.
2) Make a list L of all neighboring links of L_S. Omit from L all links between two members of N_S. If L is empty return to 1.
3) Pick with a probability P edge $l = \{j, k\}$ from L.
4) Update $L_S = L_S \cup \{l\}, N_S = N_S \cup \{j, k\}$.
5) Repeat steps $2 - 3$ until completing n-node sub-graph S.
6) Calculate the probability P to sample S.

A sub-graph of size n-node can be represented as an adjacency matrix of n x n dimension M, where $M_{ij} = 1$ when a connection between nodes i and j exists and $M_{ij} = 0$ otherwise. For simplicity, in this study, we have symbolized this adjacency matrix as a long binary integer extracted by concatenation of its rows. This number is named Identity (Id).

We applied the method described above in all corporations and their different reports to look up 3-node and 4-node connected sub-graphs ($n = 3$ and $n = 4$). We detected two kinds of 3-node sub-graphs and six types of 4-node sub-graphs. These graphs are $Id = 78$, $Id = 238$, $Id = 4,382$, $Id = 4,698$, $Id = 4,958$, $Id = 13,260$, $Id = 13,278$ and $Id = 31,710$; in all company reports. In Table 6, we show the relationship between Id and M for them.

Table VI
RELATIONSHIP BETWEEN Id AND M FOR 3-NODE AND 4-NODE SUB-GRAPHS, WHICH HAVE BEEN DETECTED IN THE ANNUAL REPORTS FOR ALL COMPANIES

	Id	M		Id	M
3 node-sub-graph	78	$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$	4-node sub-graph	4,382	$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$
	238	$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$		4,698	$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$
				4,958	$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$
				13,260	$\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$
				13,278	$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$
				31,710	$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$

In the Figures 5 and 6, we show sub-graphs that have been found and their appearances for all analyzed companies in 2009. For instance, in TELEFÓNICA company the following appearances of 3-node graphs were detected: $40,393$ with $Id = 78$; 308 with $Id = 238$ and the following identifiers for 4-node graphs were found: $1,541,382$ with $Id = 4,382$; $255,232$ with $Id = 4,698$; $45,702$ with $Id = 4,958$; $2,079$ with $Id = 13,260$; $1,404$ with $Id = 13,278$; 23 with $Id = 31,710$

Uniqueness in a graph is the number of times it appears with completely disjoint groups of nodes. For all companies in their different reports, we have also calculated the graph's uniqueness percentage over its total appearances and this parameter is higher in the graphs with $Id = 238$, for AMPER, JAZZTEL and VODAFONE companies, and in the graph with $Id = 31,710$ for TELEFÓNICA and INDRA companies. In Figures 7 and 8, we show the percentage uniqueness in each sub-graph of the reports for all corporations in 2009. So, for $Id = 238$ in AMPER this uniqueness percentage is 29,17% , in JAZZTEL it's 34,78% and in VODAFONE it's 12,20%; for $Id = 31,710$ in TELEFÓNICA it's 8,70% and in INDRA it's 50,00%. We observe that the percentage uniqueness is greater in the sub-graphs with smaller appearances.

These results suggest common communication characteristics in business language.

We also estimate that the average appearances in random networks are 1,000 nodes. The method used to generate

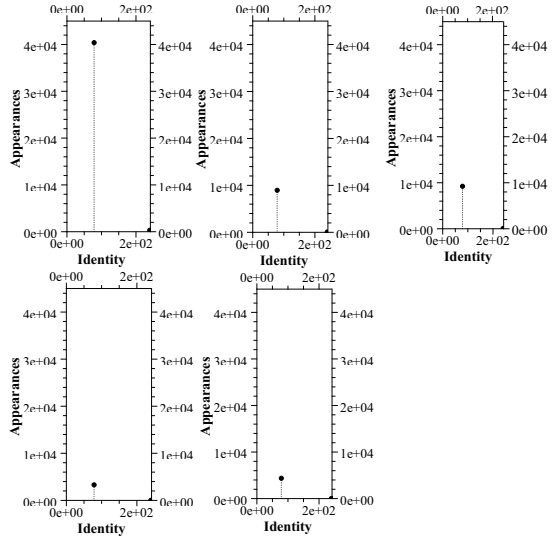

Figure 5. Appearances by Id for 3-node sub-graphs in 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-center).

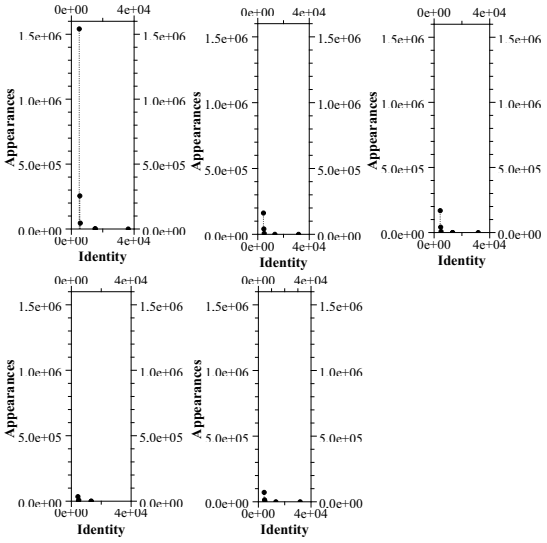

Figure 6. Appearances by Id for 4-node sub-graphs in 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-center).

random networks is the switching mechanism where we switch between links while keeping the number of incoming links of each node of the real network. The number of switches is a random number within the range of 100-200 times that the number of links appear in the network. In random networks the obtained results are also in good agreement with the real networks. If we consider the number of average appearances, the most frequent 3-node sub-graph is also: $Id = 78$ and the 4-node sub-graph with most appearances is also: $Id = 4,382$ and $Id = 4,698$, as it is shown in Figures 9 and 10 for reports in 2009.

IV. Conclusions and Future Work

In this paper, through the use of a large amount of data, we have been able to examine the syntactical language structure used by several information technology companies theoretically. We have checked their annual reports and measured various properties: average degree, main shortest path and betweenness. Additionally, we detected communities and motifs. There are common properties but other characteristics appear in some corporations only.

All company reports show a small world property which is a characteristic of complex systems. The parameters average degree ($< k >$) and the most distant vertices (d) are also similar.

Furthermore, TELEFONICA and AMPER, AMPER and INDRA show high coincidence in words. JAZZTEL and VODAFONE also have many common words. This result can suggest a common essence in the companies.

All word networks also have main communities which shows a high hierarchy among each network. In each community, the probability of appearance according to the type of word is higher for nouns, verbs and adjectives.

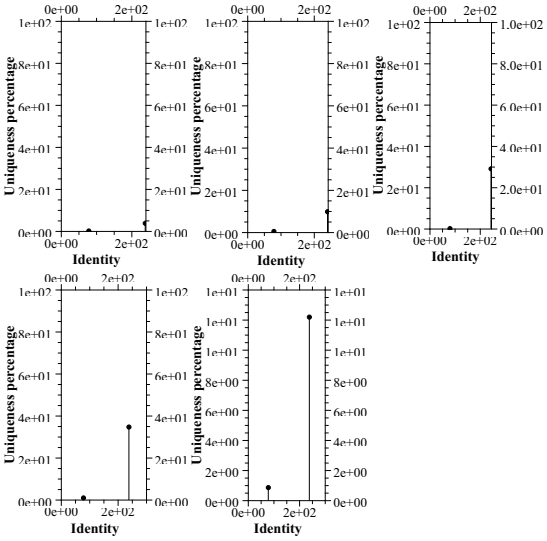

Figure 7. Uniqueness percentage for 3-node sub-graphs in 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-center).

In all corporation reports there are a large number of nodes with high betweenness but low connectivity; although the low connectivity of these words would imply that they are unimportant, their high betweenness suggests that these words have a global impact.

From a structural point of view, all companies have 3-node and 4-node common connected sub-graphs. The more frequent graph types are those with Id. 78, 4382, 4698 and 4958.

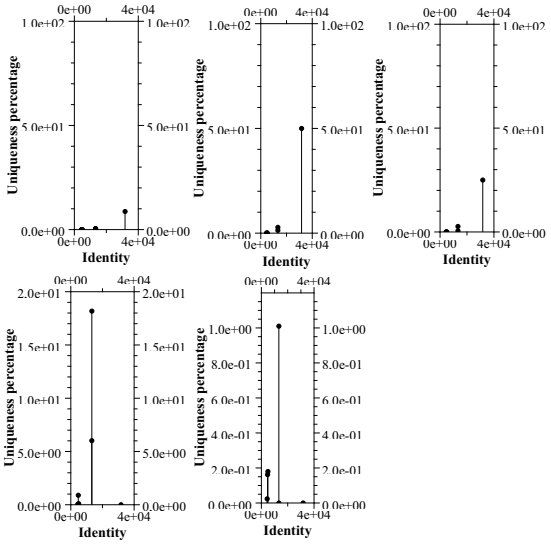

Figure 8. Uniqueness percentage for 4-node sub-graph in 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-center).

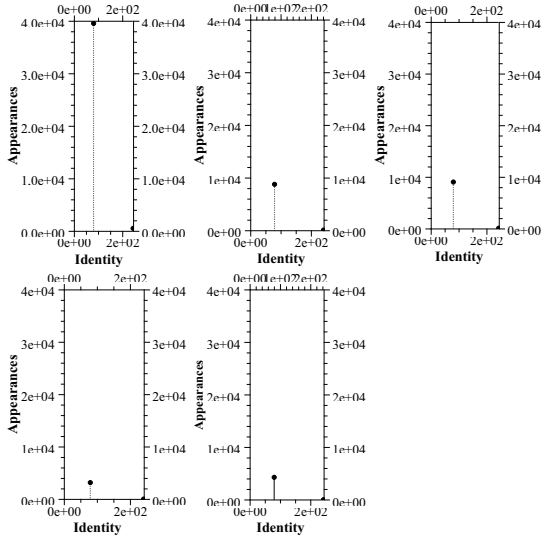

Figure 9. Appearances for 3-node sub-graphs in random networks for 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-right).

This study about the syntactic structure of the language in the organization reports, can help to identify specific and common properties in the companies. In future works, we improve this research by means of a semantic analysis.

REFERENCES

[1] E.R. Kandel, Principles of neural science, McGraw-Hill, Health Professions Division, 2000.

[2] M. Faloutsos , P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology", SIGCOMM: Conference on Communication IEEE,. 1999, pp. 251-262.

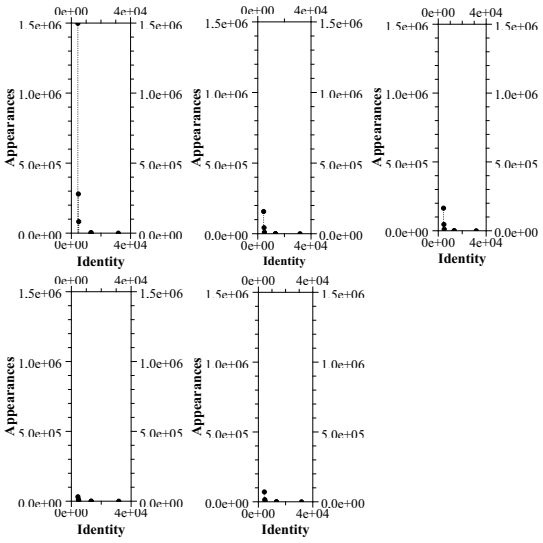

Figure 10. Appearances 4-node sub-graphs in random networks for 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-right).

[3] J. Spencer, D. Johnson , A. Hastie, and L. Sacks, "Emergent properties of the BT SDH network", BT Technology Journal, vol. 21, 2003, pp. 28-36.

[4] M.L. Mouronte et al., "Complexity in Spanish optical fiber and SDH transport networks", Computer Physics Communications, vol. 180, 2009, pp. 523-526.

[5] K. Brner, S. Sanyal, and A. Vespignani, Network science, Annual Review of Information Science & Technology, vol. 41, B. Cronin, ed., pp. 537607.

[6] R. Milo, et al., Network Motifs: Simple Building Blocks of Complex Networks, Science vol. 298, 2002, pp. 824-827.

[7] P. Pons, and M. Latapy, "Computing communities in large networks using random walks", ISCIS2005, 2005, pp. 284-293.

[8] R. Ferrer i Cancho, R. and R. V. Sole, "The small world of human language", Proc. R. Soc. Lond. B., vol. 268, 2001, pp. 2261-2265.

[9] M. Steyversa and J.B. Tenenbaumb, "The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth", Cognitive Science, vol. 29, No. 4178, 2005, pp. 4178.

[10] M. Markosova, "Network model of human language," Physica A , vol. 387, 2008, pp. 661-666.

[11] C.A. Freeman and G.A. Barnett, "An alternative approach to using interpretative theory to examine corporate messages and organizational culture", L. Thayer and Barnett G.A. (ed.): Organization Communication. Emerging Perspectives, Norwoord, New Jersey: Ablex, vol. 4, 1994.

[12] K.A. Coronges, "Structural Comparison of Cognitive Associative Networks in Two Populations",Journal of Applied Social Psychology, vol. 37, No. 9, 2005, pp. 2097–2129.

[13] T. Brasethvik and J. Atle, ”Natural Language Analysis for Semantic Document Modeling,” http://www.idi.ntnu.no/ brase/pub/nldb-brase-gullaV40.pdf, 2011.

[14] M.E.J. Newman, ”The Structure anf Functions of Complex Networks”, SIAM Review, vol 45, No. 2, 2003, pp. 167-256.

[15] S.N. Dorogovtsev and J.F.F. Mendes, ”Evolution of Networks. From Biological Nets to the Internet and WWW”, Oxford University Press, 2003.

[16] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, ”High-Betweenness Proteins in the Yeast Protein Interaction Network”, Journal of Biomedicine and Biotechnology, vol. 2, 2005, pp. 96-103.

[17] S. Shen-Orr1, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, ”Network Motifs: Simple Building Blocks of Complex Networks”, vol. 298, No. 2002, pp. 824-827.

[18] M. Newman, ”Modularity and community structure in networks”, Proceedings of the National Academy of Sciences of the United States of America , vol. 103, No. 23, 2006, pp. 8577-8582.

[19] M. Girvan and M. Newman, ”Community structure in social and biological networks”, Proceedings of the National Academy of Sciences of the United States of America, vol. 99, No. 12, 2002, pp. 7821-7826.

[20] A. Clauset, M. E. J. Newman, and C. Moore, ”Finding community structure in very large networks”, Physical Review E, 70, No. 6, 2004.

[21] A. Clauset, ”Finding local community structure in networks”, Physical Review E, Vol. 72, No. 4, 2005.

[22] B. Fields, et al., Analysis and Exploitation of Musician Social Networks for Recommendation and Discovery, IEEE Transactions on Multimedia, vol. 13, 2011, pp. 674-686.

[23] ”AMPER”, http://www.amper.es/index.cfm?lang=sp, July 2012.

[24] ”JAZZTEL”, http://www.jazztel.com/home, July 2012.

[25] ”VODAFONE”, http://www.vodafone.es/particulares/es/?cid=12072012-000000148, July 2012.

[26] ”INDRA”, http://www.indracompany.com/, July 2012.

[27] ”TELEFÓNICA”, http://www.telefonica.com/es/home/jsp/home.jsp, July 2012.

 ISBN: 978-1-61208-242-4