OPAALS PROJECT

Contract n° IST-034824

WP6: Socio-Constructivism and Language

Del6.13 – Complexities and frequencies of linguistic patterns (natural and formal) in online environments

Project funded by the European Community under the "Information Society Technology" Programme

Contract Number: IST-034824

Project Acronym: OPAALS

Deliverable N°: 6.13

Due date: April 2010

Delivery Date: July 2010

Short Description:

This deliverable investigates the complexity and evolutionary mechanisms of change in the context of system design. The integration of semiotic and frequency-based approaches is the central theme of this research.

Author: Oxana Lapteva (UniKassel)

Partners contributed: UniKassel, TUT, SUAS

Made available to: Consortium

Versioning

Version	Date	Name, organization
V1	01.03.10	Oxana Lapteva
V2	15.05.10	Oxana Lapteva
V3	05.06.10	Oxana Lapteva
V4	25.06.10	Oxana Lapteva, Josef Wallmannsberger
V5	30.06.10	Oxana Lapteva

Quality check

Internal Reviewers: Raimund Eder, Jaakko Salonen

Dependences:

Achievements*	Accomplished work: – theoretical investigations in the field of semiotics – applying the semiotic theory to the Digital Ecosystem – identifying the underlying structures of systems through the lens of semiotic – Continuation of research on underlying mechanisms of evolution – Statistical analysis of OPAALS language
Work Packages	The universal principles of a system structure and evolution provide an important input for any system development in the context of its dynamic character, self-organisational aspects and evolution. Looking at the OKS interface as a sign system helps to connect the developers with the users considering their socio-cultural background. This aspect illustrate a close connectivity to the research in different Work Packages: WP10, WP 5, WP2. Furthermore, the integration of semiotic and frequency-based analysis provides a different perspective on the problem of design and evolution in DE.
Partners	ALL
Domains	Linguistics: aspects of natural language evolution, semiotic theory and DE Computer Science: evolution in "formal" environments (knowledge spaces); design and development of sign systems (e.g. ontologies, user interfaces) Human-computer interface: possible applications of the frequency-based theory; design and implementation of the OKS.
Targets	OPAALS researchers, Scientific communities
Publications*	
PhD Students*	Oxana Lapteva – semiotic theory – sign systems in Digital Ecosystems – frequency in natural language – language networks
Outstanding features*	Integration of the linguistic perspective on the complexities and frequencies in Digital Ecosystems Discovery of the "universal" features of systems and the common mechanisms and laws of evolution in "natural" and digital/formal systems
Disciplinary domains of	Oxana Lapteva (Computational Linguistics)

authors*	

The information marked with an asterisk (*) is provided in order to address Recommendation n. 4 from the Year 2 review report

Contents

List of Figures

1 Executive Summary

In the Digital Ecosystem (DE), different elements (e.g. interfaces, language, knowledge representation) interact with each other. The goal of this research is to provide a common background of analysing such systems through introducing two ideas:

1. The complex components of DEs are sign systems having common structure

2. Frequency forces systems to change

This deliverable aims to continue the investigations of a frequency-based approach on evolution proposed in the Deliverable D6.8 [1]. Additional dimension of semiotics (study of sign systems) allows us to generalise the study of system structures and functionalities as well as to provide useful insights into development (and design) of such systems. The right choice of signs and sign systems is important in the representation of structural, communicative and functional components of any user interface (e.g. icon of a program or function). Independent of their origin (e.g. a program, natural language, visual system, and others), all systems existing in DEs are sign systems. Hence, the semiotic approach helps to understand their "universal" structure and to merge different poles (technical and human) through the interface design. In overall, the integration of these approaches endeavours the investigations towards universal principles of dynamic systems.

[1] Lapteva, O., Peukert, H., Nyknen, O. and Eder, R. (2009). Models of the Evolutionary Framework for Language, European Commission FP7 "OPAALS" project deliverable 6.8.

2 Scope of the deliverable

This deliverable identifies two important aspect of the design and development of DEs in general and single dynamic elements (interfaces, knowledge representations, and other) in particular. The semiotic view provides a common universal structure for all systems, i.e. the structure of signs. Looking through the lens of semiotic theory, it is possible to analyse the complexity of these structures. However, these systems are constantly changing followed certain mechanisms. Hence, the frequency-based approach is applied using the example of OPAALS language.

The proposed framework provides not only the theoretical platform for analysing complex dynamic systems and their interaction, but also introduces new dimensions into the design and implementation of such systems (e.g. OKS). Furthermore, the network analysis of natural language provides important insights into the problem and mechanisms of language change, which are crucial in the overall development of DEs. The output of this research can be applied to different areas of DEs, for example:

- socio-cultural norms and their reflection in digital environments (in form of sign systems)
- development of interactive dynamic user-interface environments (considering the integration of semiotic theory and the frequency-based approach on evolution)
- analysis of frequency patterns in natural language can be effectively integrated in the DEs and OKS (e.g. dynamic knowledge spaces and language networks)

3 Introduction

Language is one of the most important and at the same time complex systems in human societies. The same is true for Digital Ecosystems: not only the dimension of communication and social interaction is involved, but also the technical layer that underlies and supports these processes. Hence, the role of linguistic analysis is crucial for understanding the structure and character of DEs with humans as their primary actors. The linguistic lens opens a new perspective to the human-related aspects of communication and knowledge representation/sharing as well as to the technological development involved in the evolution of DEs.

This deliverable considers the linguistic view at different layers of organisation within the DE based on the self-reflection of the OPAALS as a community and the notion of the Open Knowledge Space (OKS) as the platform for collaboration, communication, knowledge representation and sharing. In all these processes, language is the underlying mechanism, which complexity and dynamic evolving character influence the overall development in societies.

Considering the fact that technological environment involves humans as well (e.g. software developers, engineers), we make a step further and apply the linguistic (semiotic) view on these parts of the DE development. Hence, our analytic framework connects different poles of digital communities: from technical creation through human-computer interaction to the evolving community.

Digital Ecosystems are highly complex environments that involve not only technological but also complex socio-cultural forms where language as a mechanism of communication provides an underlying platform. It is used in communication between different members of a community, forms up the knowledge representation and information sharing among users, reflects cultural/organisational norms and rules, and influences the evolutionary aspects of development. Language as a complex sign system with its rules, ambiguity and richness connects a variety of aspects in our life, from interactions with other members in the community or culture to the development of complex knowledge forms. Hence, language is inseparable part of the existence within the Digital Ecosystem.

In the following sections we are going to discuss the complexity of linguistic patterns not only in relation to the language structure and change, but also in connection with the DE and OKS involving different "actors" (developers, interfaces, and users). The interaction of language, culture, and technology causes the constant change and evolution of technological and socio-cultural environments at different levels of structural organisation (e.g. groups, organisations, communities, cultures).

The focus of this deliverable is twofold. On the one hand, we introduce the semiotic approach in order to analyse the complexity of different structures as sign systems existing in Digital Ecosystems. On the other hand, the frequency-based approach introduced in D6.8 adds the evolutionary aspect to the analysis of these structures.

4 Linguistic Perspective

In relation to the Digital Ecosystem, Open Knowledge Space (OKS) represents an environmental platform for communication, knowledge creation, representation and sharing. One of the important properties of this platform is its distributional character allowing dynamic growth and change at different levels of organisation including social network and information/knowledge distributed through this network. The evolution of the OKS is a complex multi-layered process, in which the language plays a central role. Considering a Digital Ecosystem as the underlying environment, in which the OKS grows and changes, it becomes obvious that the evolutionary aspects cannot be reduced to the practical "implementations", for example, version control of information, system, or network. In order to understand the dynamics, regulations and laws of the DE and its components, complex interplay of different layers, their structures and processes needs to be considered (Figure 1).

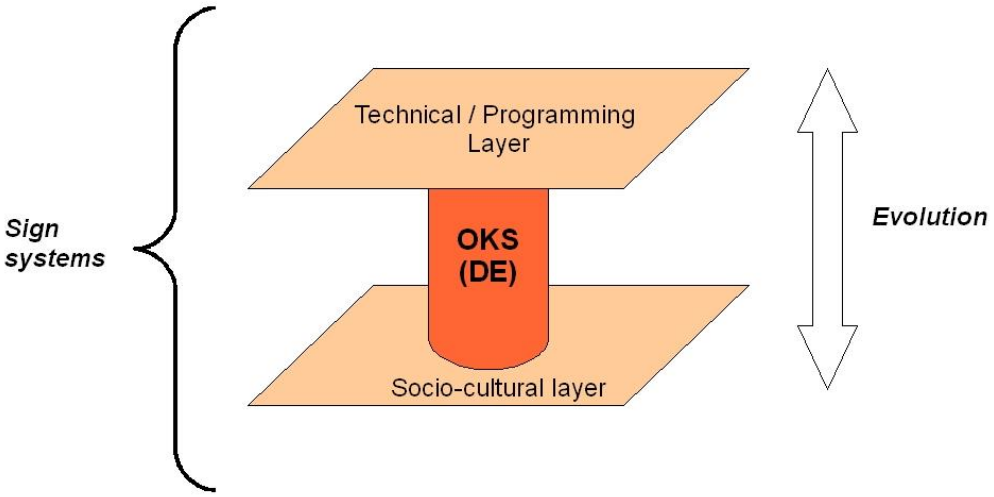

Figure 1: *Analytic Framework*

The technical, socio-cultural and interface (e.g. OKS) layers correspond to the different "systems" existing within Digital Ecosystems. Even though different types of actors work in each layer (e.g. software developers in the technical and interface systems; users are the part of interface and socio-cultural layers), we argue that the joined integrative view on the environment as a whole provides useful insights into the complexity of the structure,

patterns and evolution.

4.1 Technical/programming layer

The technical layer underlies the whole structure of the OKS and is crucial for the overall functioning of DE. Considering the dynamic character of DEs, the development of the corresponding infrastructure (e.g. Flypeer [2] within the OKS) influences the interaction between users and programs. Humans have the cognitive and physiological components that allow them to communicate. Same is true for digital environments. We need a platform for processing the information in human-understandable form. From the user perspective, there is no direct manipulation of the binary code: users "manipulate higher-level abstracted signs and representations of the data in order to perform tasks" (O'Neill, 2008, p. 15). At the intersection between technical and user-centred aspects of the OKS, the role of semiotic sign and theory will be discussed together with the language-based view on the evolution of the OKS.

4.2 Open Knowledge Space as a platform for interactions

"While binary code still remains at the center of the computational process, graphical and lexical sign systems that represent concepts and familiar activities have been introduced as the front end, or interface, that allow us to understand and operate them in our own terms"(O'Neill, 2008, p. 15).

OKS is the connecting component between technical and socio-cultural layers. Within the digital environments, the interaction between the technical, interface and user perspectives usually occurs in one of the following modes:

- system – user (e.g. a community member navigates through the OKS)
- user – system – user, i.e a system is a connecting element in communication and collaboration between users (e.g. a conference call using the OKS)
- system – system (e.g. adjustment of different knowledge representation systems such as ontologies without the human involvement)
- developer – system (e.g. development of the OKS)

[2]Dynamic P2P Infrastructure: http://kenai.com/projects/flypeer

These three modes co-exists in the proposed framework: the developer works primarily in the Technical/Programming Layer, a system exists within the second layer (OKS/DE) and the user is the primary actor within the socio-cultural layer. In relation to the DE and OKS development, we suggest to use the more advantageous mode of interactions, the triadic mode illustrated in Figure 2.

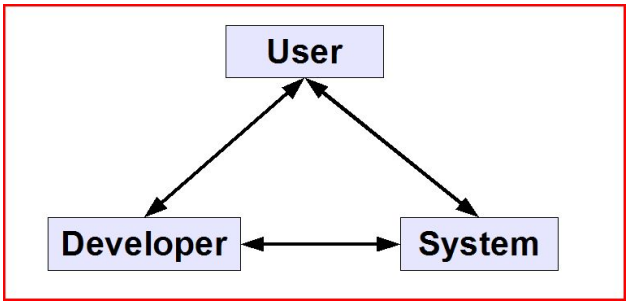

Figure 2: *Modes of interaction*

One of the important goals of community building, especially in context of the OPAALS project, is knowledge production, representation and sharing. The community plays a critical role in determining what we take to be true, and how we develop knowledge. Knowledge is always situated in a place, time, conditions, practices and understandings (Van House, 2004). Regarding collaborative knowledge production and management in the interdisciplinary communities, there is a need of the environment that can be used and managed intuitively from members of different domains and backgrounds. In order to achieve this goal, the "common language" by means of terms, their meanings and relations is necessary.

Furthermore, the users needs, their socio-cultural rules and traditions influence the use of digital platforms. This aspects must already be considered at the stage of design, for example, what signs and sign systems are acceptable and understandable in the certain community; what are the communicative traditions. Additionally, the aspects of adaptability, dynamics and evolution plays an important role in the context of the OKS development.

4.3 Socio-cultural layer

The complexity of natural language stems from the intricacy of language itself (Farghaly & Hedin, 2003). In the context of the community development, it is important to understand how linguistic items are selected and organized to serve the specific cognitive aims of a particular knowledge domain. Our choice of words and their usage may be conditioned by what we are talking

about and to whom we are talking (Farghaly & Hedin, 2003). Signs and their meaning are formed by social groups primarily as part of the social division of labor in society. These issues influence the linguistic work within a specific domain. A large number of groups may develop symbol systems (terms and concepts) and share knowledge, which they do not share with the rest of society. There may be a considerable degree of common knowledge and shared meaning (Hjorland, 2002). The structured terminology can be seen as an multidimensional interacting system that helps build and organize communities and their relations and interactions with others.

The socio-cultural layer provides the requirements used by a community in form of communicative styles, forms and rules, language and sign usage, types of tools, knowledge representation forms.

4.4 Sign Systems and Evolution

How the described layers, in respect to their complexity, can be conjointly analysed? We argue, that the integration of two approaches, semiotics (= sign systems) and evolution (= role of frequency) provides a platform of investigating the complexity and underlying mechanism of change in respect to the DE research.

OKS is a complex platform consisting "not only of simple computational forms but also of complex cultural forms (sign systems)" (O'Neill, 2008, p. 19). In relation to the Human-Computer-Interaction (HCI) research, the analysis of interfaces as systems of signs can be interesting and useful for two reasons:

1. Considering interfaces as signs we connect the objects to the meanings and possible interpretations by users.

2. Semiotic approach considers "the whole process of interaction as a semiotic process, where signs are transformed, exchanged and interpreted" (O'Neill, 2003, p. 1) at different levels.

The interaction of sign systems including natural language forces the evolution of their structure and elements.

In Digital Ecosystems, similar to the real-life environments, different natural pressures force change. Technical progress influences the way we use digital platforms: "Future computer systems promise to enhance the perceptual mapping through increasing use of graphical displays, including large screens, color, and three-dimensions, and the use of motion and sound" (Grudin & Norman, 1991, p. 614).

The second pressure of evolution is the contact with other groups (organisations, communities, cultures). The contact forces the change of language in order to make the communication between members of different groups successful and efficient. The change might involve the import of elements from one language into another one.

These pressures influences the change of the third type, the knowledge representation system. Through the contact of cultures and technological progress, the communities are forced to change their knowledge system accordingly. In other words, the underlying structure of the community's knowledge should react to the changes in language system and to be in line with the technological changes the DE faces. The evolution of knowledge systems can involve the changes of the high-level structure (e.g. meta-level of ontologies) and of the elements within the structure (e.g. new/old elements, relations between elements, the semantics of elements).

5 Semiotic approach

5.1 Semiotic Theory

The aim of semiotics is to understand and explain "the structure of sign systems in relation to the way they convey meaning" (O'Neill, 2008, p. 67). Starting with Saussure and Peirce, who proposed models of sign systems, the discipline of semiotics undergo essential changes tracing back to the technological progress in development of computers and digital environments.

Saussure (1996), focusing on natural language, introduced two distinct parts of a sign: "signifier" and "signified". A signifier corresponds to the expression (e.g. sound sequence of a word), whereas a signified represents a concepts at the level of meaning (Figure 3).

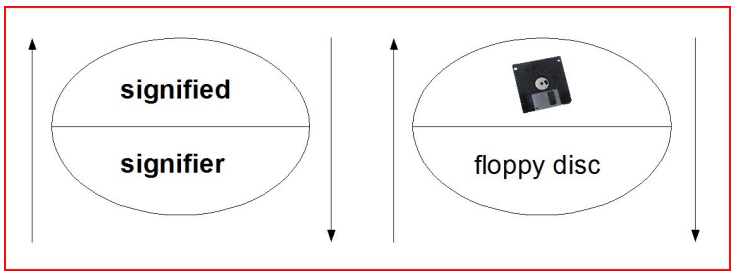

Figure 3: *Saussuean Model of Sign*

In contrast to Saussure, Pierce was concerned with more general conception of the sign, going beyond natural language and looking at different kinds of signs existing in human environments and their phenomenal qualities (O'Neill, 2008). The sign, according to Peirce, consists of three distinct parts: the object, the representamen, and the interpretant. Applying this model to the development of the OKS, it is necessary to consider how an object can be represented and the possible way(s) of its interpretation by users (Figure 4).

Furthermore, considering phenomenal qualities of different kinds of signs, Peirce proposed the following categorisation of signs:

1. Icon represents "their objects via direct likeness or similarity. Essentially, icons have features or qualities that resemble those of the objects they represent" (O'Neill, 2008, p. 70). There are three different types of iconic signs:

    - Images refer not only to the visual/pictorial signs, moreover, they are "any sensory qualities, or combinations thereof, that represent an object" (Johansen & Larsen, 2002, p. 37).

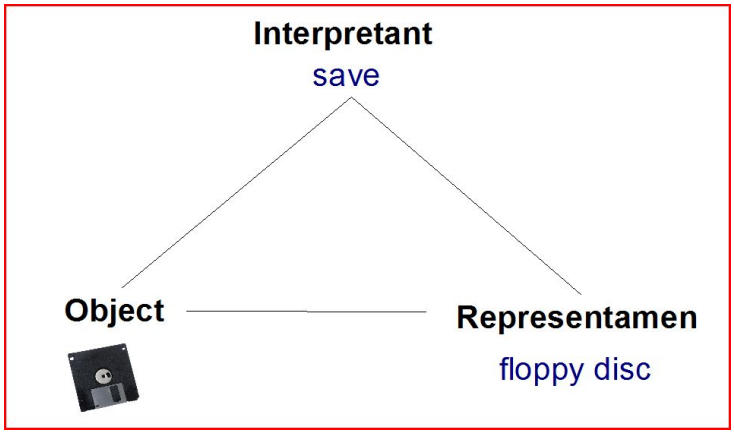

Figure 4: *Peirce: Triadic Model of Sign*

- Diagrams, functioning at the higher degree of abstraction, reflect structure and relations between elements.
- Metaphors connect at the first sight non-related elements[3].

2. Index indicates an object. As O'Neill (2008) points out, "There is a direct link between the object and the sign. Indices are signs or imprints often left in one physical entity, possibly a medium, by the passage of another physical entity that uses that medium. There is a clear connection here between the signifier and the signified, the form and the content" (p. 70).

3. Symbol refers "to their objects by virtue of a law or set of socially derived rules that cause the symbol to be interpreted as referring to that object" (O'Neill, 2008, p. 70). Symbols represent the conventions that are accepted within a certain culture.

These types of sings are necessary for the OKS in order to navigate, work and communicate through the platform. The choice of the signs has to be considered in respect to the social, cultural and contextual aspects (as proposed by Eco (1976)). In our life we use signs to transmit messages according to certain cultural norms and conventions. The same is true for Digital Ecosystems: certain norms exist that define the use of proper sign for communication with a system and through the system with other users. In the context of the OKS structure and evolution, "the actual sign usage, as expressed in the system processes, partly presupposes the users, partly

[3]see deliverables D6.3 "Metaphorological Tool Kit" and D6.7 "Metaphorological Tool Kit: implementation and results" for more details on metaphors

presupposes the system structure, which again presupposes the designer" (Andersen, 1997, p. 184). Changes in one element causes the chain reaction of changes in the whole environment.

5.2 Semiotic view on the OKS

OKS is meant to be a working/social platform for the knowledge communities and cultures within the complex environments making up the Digital Ecosystem. Hence, the role of human-computer interactions, communicative processes, language and knowledge representation become the focus of investigations. The development process of such interfaces is closely related to the study of signs, their creation and interpretation. As Kumiko Tanaka-Ishii (2010) pointed out:

> "Behind every computational system is a program. Programming languages are among the most widely applied artificial languages. Most programs are generated by human beings. An expression written in programming language is interpreted by both human and machines, these languages reflect the linguistic behaviour of both. The application of semiotics to programming languages therefore helps situate certain technological phenomena within a humanities framework. " (p.5)

The semiotic approach within the OKS development introduces the interaction between machine, developer, and user. The same element (sign) is differently represented at the level of hardware, software, and human.

Furthermore, it helps to understand the development of the system as interaction between hardware, software, and users. "As code structure that stand for design ideas, patterns can be treated directly as signs and this can lead to an effective categorisation scheme for patterns" (Khaled & Noble, 2004, p. 9).

Considering a sign as mechanism of coding and decoding messages according to the rules of a certain culture or community, we are facing the necessity of applying the semiotic approach to the analysis of OKS, especially in terms of user-interface design, communication and collaboration. From the design perspective, we differentiate static and interactive elements of the OKS. First group, including forms, layout, graphics, colours, and other elements, is important "in establishing a frame of reference from which to engage with its interactive elements" (O'Neill, 2008, p. 85). The static elements of the OKS have to provide the clear structure of the interactive and dynamic environments. The consistency of these elements helps the user to navigate through the digital space. As O'Neill (2008) pointed out:

> "Without the static elements to guide us on the screen, we would be lost in a maelstrom of interactive and dynamic elements. The balance between good static layout and dynamic interactive elements is clearly an important part of any interface design. ... The key aspect of the design of such graphics, then, has to be their ability to communicate their purpose to a user. However, it is not always as straightforward as assuming that a graphic might work in a singular, linear relationship with its function. Many other aspects of its makeup, its situation within a graphic structure as a whole, and the context in which it is viewed, can have an effect on its meaning. Moreover, the functionality of graphics often works in a metaphorical way that attempts to imbue the virtual world with a similarity to the real world (e.g. the desktop metaphor), which does not always display the same logic or affordandes that would be apparent in the real world" (p.85).

Applying the triadic model of signs (Peirce) to the OKS, we can identify that the representamen corresponds to the form of the sign in the environment, the object – to the underlying functionality of the sign and the interpretant – to the understanding by the user (interpretations generated in human mind).

Graphics are also involved in the interaction between human and digital environments as well as between different users/communities/cultures through these environments. Hence, these elements are often integrated in the dynamic, changeable systems together with other elements such as video, texts, images, and culturally-related symbols.

In the context of the OKS, we need to consider not only the social and cultural aspects of interaction. Open Knowledge Space is a working environment that has "a direct affect on the language that people use to talk about the tasks that they are performing in that environment" (O'Neill, 2008, p. 40)

The interpretation of different signs is based on knowledge and codes existing within a certain group, community and culture. This is true for both, virtual and "real-life" environments. The semiotic competence of the users includes, according to Johansen and Larsen (2002, p. 30):

1. the ability to *perceive* phenomena in our environment as signs, i.e. to understand the connection between present, (partially) hidden, and entirely absent phenomena
2. the ability to *produce* and transmit signs
3. the ability to *store* information and form interpretative habits on the basis of memory and learning processes.

The OKS as a type of user interface can be seen as a complex sign system containing single elements (e.g. images, buttons, icons, text, etc.). These elements allows the communication between users and interface environment. While interacting with the system, users interpret these signs. How important the triadic model of sign can be in the selection of the "correct" signs, shows the usability study described in the deliverable D.10.21 [4]. Considering the semiotic view on Guigoh [5], several aspects overlap with the user perspectives found in the socio-scientific study of the usability of the OKS. Confusion caused by the user experience was often based on the choice of signs to represent elements and their function in the environment (e.g. home-button and home-textlink; icon bar; "send file" function, and others). Furthermore, the underlying sign system, i.e. the consistency, integrity, connections of and between the signs, and the interpretation perspective, are not yet fully considered in the design of the OKS.

One of the major aspects in the development of user interfaces is the appropriate knowledge representation platform with the language as its underlying system. Independently of the type of language (natural, semi-formal, or formal), it is the system of signs that provides meaningful information according to the socio-cultural norms of a community. As the community constantly develops and changes, the knowledge and language changes influencing their representational aspects in form of complex signs. Within the OKS, introducing signs requires the socio-cultural knowledge of a community, e.g. what signs are accepted by this community, how are they interpreted, are there any ambiguity issues in the process of interpretation.

From the semiotic perspective of the OKS, following requirements need to be fulfilled (Johansen & Larsen, 2002, p. 31):

1. the development of the representative situation-independent systems
2. the development of production and reproduction techniques that ensure the reliable and preferably rapid recording and distribution of signs
3. the creation of conventions that ensure more or less congruent interpretations of the transmitted signs

In conclusion, semiotics "provides guidelines for organizing and using signs to represent something to someone for some purpose". Hence, it is

[4]Bruer, M., Steinicke, I., and Zeller, F. (2010). Report on recommendations for improving the usability of the OKS, European Commission FP7 "OPAALS" project deliverable 10.21

[5]http://www.opaals.org.br/

an important component in the design and implementation of interactive interfaces. Andersen (2000) highlights the following aspects in respect to the semiotic approach of interface design:

- System model and User model are "two interpretations of the same sign-complex produced by two groups that access different parts of it (designer and user)" (p. 420)

- Reinterpretation of technical issues

  > From a semiotic perspective, many processes which computer science sees as data storage and retrieval are really communicative processes. [...] This means that it must conform to the normal principles of communication by providing a context for interpretation (Andersen, 2000, p. 420).

  On the one hand, the formal systems have to be defined according to the certain forms and patterns. On the other hand, these systems have to be developed considering interpretational aspects within the certain socio-cultural and organisational contexts.

- The qualitative and quantitative social research methods are necessary to force the interactions between developers and users, between the sign system creation and its interpretation.

6 Frequency of Linguistic patterns

> "...evolution refers to the process by which organizations and their information systems change over time. Since the world changes, business changes, strategy changes and context change, the information systems need to adapt to such changes in order to deliver benefits. Further, changes in information systems also affect other systems. The change process is continuous, multi-dimensional and difficult to manage " (Lindgren, Hardless, Pessi, & Nulden, 2002, p. 1).

The major element influencing the change of a knowledge system in human societies is natural language. It is one of the core components of our life. People communicate through language, share their ideas, discuss problems, represent and produce their knowledge. As a complex system, language provides an intrigating platform for investigation of patterns and their dynamics on the time scale. In online communities, the complexity and changeability of language becomes crucial.

One of the interesting platforms for investigating linguistic patterns, their complexity and dynamics, is the OPAALS community. On the one hand, we have "different" scientific languages (e.g. social scientists, natural scientists, computer scientists and engineers) that interact with each other in order to provide a common global platform for communication and knowledge production. On the other hand, through interactions between community's members we are facing strong alterations of different language patterns (e.g., new words, concepts, expressions, and others).

In the context of language use in the OPAALS community, we can find the first evidence supporting the frequency-based approach on language evolution, which was proposed and described in the Deliverable 6.8. Two important aspects of analysing frequency patterns in the OPAALS' language, Zipfian distribution of words and the statistical patterns of language networks, will be presented in this section.

6.1 Linguistic Data: OPAALS community

For the purposes of analysing linguistic patterns existing within the OPAALS community, the corpus-based approach has been chosen [6]. The OPAALS corpus is a complex collection of data including:

- wiki corpus containing the web sites from the OPAALS wiki pages

[6] see Deliverable D6.4 for more detail on corpus-based approach.

In order to make this corpus as representative as possible, we applied a certain set of filtering criteria: the pages containing reports, discussions of research topics, descriptions and progress of the research tasks, etc. were included in the corpus. Contrariwise, web sites containing system-relevant information (for example, help pages), meeting information (place, time, directions, accommodation), and so on have been filtered out and were not included in the corpus.

- deliverable corpus

  Deliverables reflect the knowledge of OPAALS community as a result of interdisciplinary collaboration. Hence, all documents submitted since the beginning of the OPAALS project have been included in the corpus.

- email corpus

  In such communities like OPAALS (i.e. interdisciplinary, distributed, and multilingual), emails are one of the major sources for communication. OPAALS email corpus was build based on the messages communicated through all@opaals.org.

- newsletter corpus

  As a holder for the latest news and summary of the scientific work within the OPAALS community, newsletter provide important data for analysis.

- questionnaire data

  At the beginning of the OPAALS project, the questionnaire has been developed in order to analyse dynamics of the OPAALS community on the time scale. One part of the questionnaire addressed linguistic issues (e.g. understanding of concepts relevant for the OPAALS research). The resulting data covers four waves made within the life of the project.

The timeline for the OPAALS data collected during the OPAALS project is illustrated in Figure 5.

6.2 Zipfian Distributions in Language of the OPAALS community

One of the observations we consider in relation to natural language is called the Zipf's law proposed by the linguist George Kingsley Zipf in 1935. Basically, it describes the probability of occurrence of elements in a given corpus

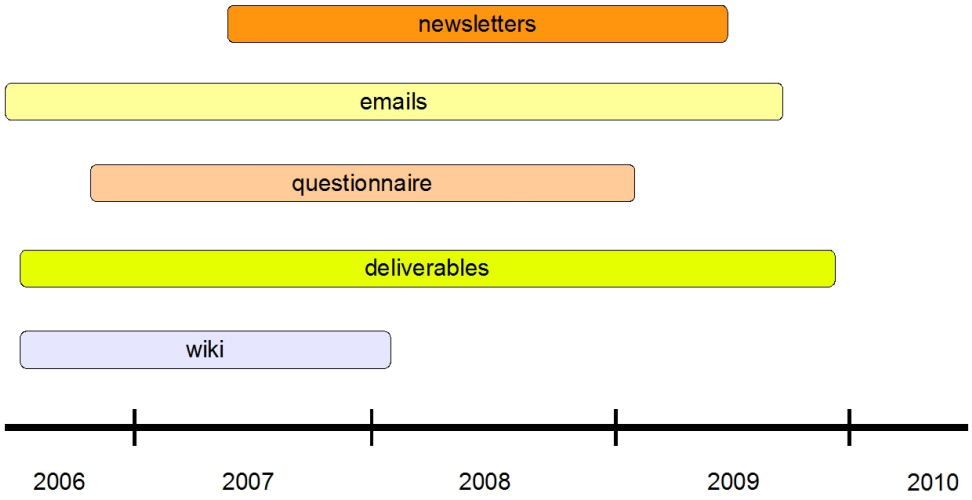

Figure 5: *OPAALS corpus: timeline*

of natural language and states that the rank of words in terms of their frequency is inversely proportional to their actual frequency (Tullo & Hurford, 2003).

The mathematical representation of this law is $f = \frac{k}{r^{\beta}}$, where f is the frequency, r is the rank and k is a constant.

In relation to natural language, we use the Zip's law in two relations:

1. the relation between the frequency of words and their rank
2. the relation between the length of words and their frequency of usage

Based on the collected data of OPAALS community, we analysed the frequency distribution of words. All corpus data shows the tendency described by Zipf's law. The results in the Figure 6 illustrate the Zipfian distribution with log axes. Ranks are represented by the X-axis, the frequencies – by the Y-axis.

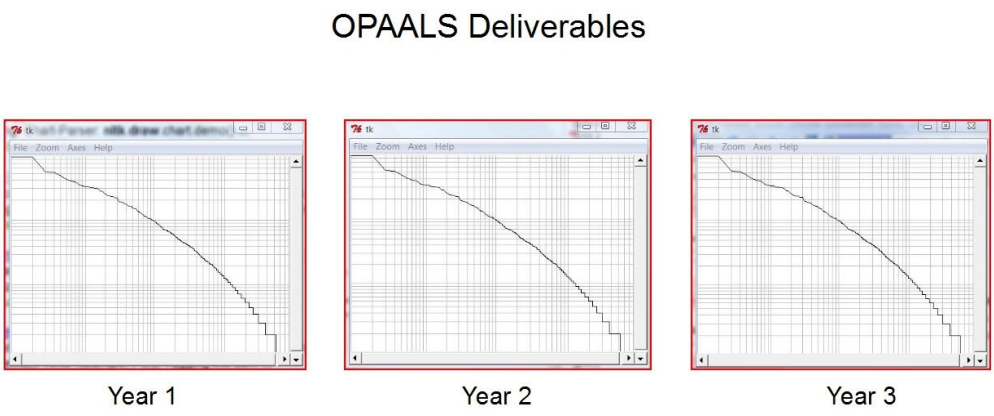

Figure 6: *Zipfian Distribution*

Our observations show that the words do not follow the Zip's law with the mathematical precision. This is due to the fact that the empirical data of OPAALS (and of any other natural language) consists of many words that share the same frequency. This is especially the case in the low-frequency areas. More importantly, however, is not the mathematical precision, rather the tendency we observe while analysing such complex systems. This aspect has been found in other studies focusing on the Zipfian distribution in languages (Li, 1992; Hatzigeorgiu & Carayannis, 2001; Nemeth & Zainko, 2002).

Another aspect, we can analyse using Zipf's law is the relationship between the frequency of occurrence and the length of a word. "it seems reasonably clear that shorter words are distinctly more favoured in language than longer words" (Zipf, 1993). This observation is also true for the language of the OPAALS community.

What do these observations mean for the natural language and its change? And can we find similar observations in other sign systems?

Using this observed tendency, we can try to explain the emergence of irregularities in language. Thus, beside the Zipfian correlations, linguists are familiar with the fact that there is also a correlation between the frequency of words and constructions and their morphological or syntactic irregularity. For instance, the most frequent verbs in a language are also those most likely to be irregular (Tullo & Hurford, 2003, p. 63).

The Zipfian distribution can be found in other forms as well. In 1986, Ellis and Hitchcock (1986) observed that computer users create command abbreviations. Moreover, these aliases follow the Zipf's law in terms of the relation between the length and the frequency of usage. This means, that the shorter terms were used for high-frequency commands (Grudin & Norman, 1991). The Zipf's law provides the platform for investigating certain properties of a system. In relation to natural language, these properties are:

- the efficiency of language (e.g. shorter words are more frequent)
- the combination of and connection between words in a language

As Cancho and Solé (2003) pointed out, "the emergence of a complex language is one of the fundamental events of human evolution" (p.788). Zipf's law as one of the fundamental principles of organisation in such systems provides a model for the analysis of structural properties, power and efficiency.

This approach considers the single elements, their form and position within a system. However, the evolutionary processes depend also on the connections between these elements. Hence, the research can be expanded through introducing the network analysis, which is the focus of the next section.

6.3 Language networks

Natural languages are networks that are robust and efficient in respect of their evolution. "If the network characteristics are properly identified, they may be applied in the construction of many kinds of artificial networks. These characteristics are expected to be culture independent" (Makaruk & Owczarek, 2008)

The statistical properties of language networks provide valuable information about processes of development and evolution and can be applied to the different levels of organisation (e.g. language of individual, group, organisation, community, culture). Furthermore, since the network structure underlies the representation of knowledge in digital environments, the investigations of its structure and dynamics are necessary.

In the analysis of the OPAALS language, we chose the collocational network that represents the connections between terms used in the OPAALS community. The exemplary fragment of the network presented in Figure 7 shows the collocations used in the deliverables of the year 1.

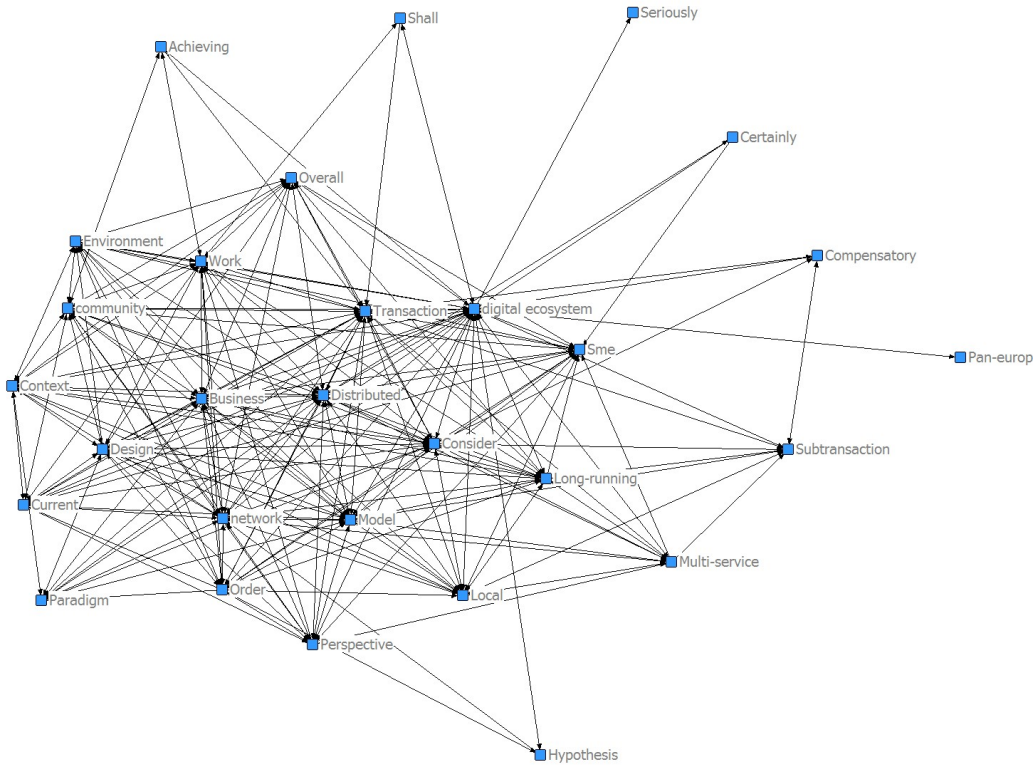

Figure 7: *Collocational Network*

In the analysis of statistical properties, we consider different measures:

1. Average path length, which is the number of relations in the shortest way from one node to another

2. Clustering Coefficient, which mirrors the closeness of nodes in the network.

3. Degree distribution

4. Degree centrality, which reflects the number of connections a node has. In the collocational network we differentiate between the Out-Degree Centrality (total number of ties sent) and In-Degree Centrality (total number of ties received). Centralisation characterises a network in terms of how unequal the distribution of centrality is and how much variance there is in the network.

Additionally, the random network of the same size has been created in order to provide a reference point in our analysis. The statistical measures of both networks are provided in Table 1.

Measure	Notation	Collocation	Random
Clustering Coefficient	CC	9.615	0.994
Weighted Clustering Coefficient		7.975	0.993
Average path length	P	1.565	1.374

Table 1: *Statistical Measures*

As illustrated in Table 1, the collocational network shows $P = 1.565$ and $CC = 9.615$. In comparison, the corresponding random network has $P_{random} = 1.374$ and $CC_{random} = 0.994$. These results are the hallmarks of the small-world phenomenon. Small world network can be defined as a graph with n vertices and average vertex degree k that exhibits $P \approx P_{random}(n, k)$ and $CC >> CC_{random}$ (Watts, 2003). P refers to the path-length of the network and CC is the clustering coefficient. P_{random} and CC_{random} are the values of the path-length and clustering coefficient for the random network with the same k and n. The degree distributions within this network are illustrated in Figure 8. Our results partially correspond to the work of Motter, Moura, Lai, and Dasgupta (2002) who investigated the properties of language network in English Thesaurus. The entries in the Thesaurus are vertices of the network and synonymic relationships build up the corresponding links between them. According to the authors, this network "presents a small-world structure, with an amazingly small average shortest path, and appears to exhibit an asymptotic scale-free feature" (p. 1).

In our analysis of the collocational network, we also use the centrality measures that refer "to the identification of the 'most important' actors in the network" (Parhi, 2008, p. 13). The degree centrality of the network is represented by two measures: In-Degree and Out-Degree. Out-Degree indicates the influence of nodes in the network, whereas In-Degree refers to the prestige and popularity of the nodes in the network. In other words, nodes

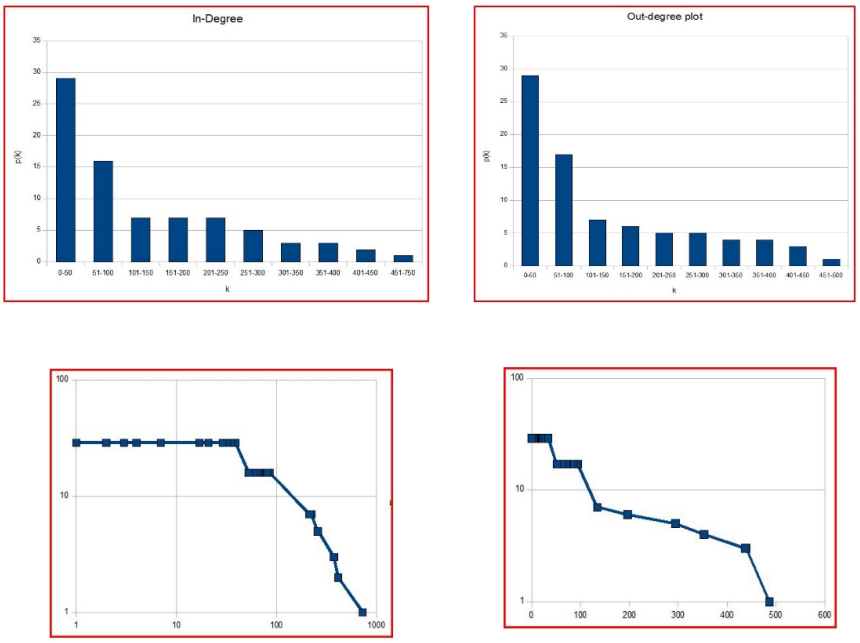

Figure 8: *Degree Distribution*

"receiving many ties are often said to be prominent or have high prestige" (Parhi, 2008, p. 14) in the network. The nodes with the high out-degree values are "those who are able to exchange with many others" (Parhi, 2008, p. 14). The out-degrees of the nodes in our network shows the sparseness with the relatively high number of nodes having the small value of degree range. Furthermore, both the Out-degree and In-degree vary in values among the nodes within the network (see Table 2). Analysing the centralisation

Measures	Out-Degree	In-Degree
Mean	113.310	113.310
Standard Deviation	141.990	160.976
Maximum	0.000	1.000
Maximum	487.000	729.000
Network Centralisation	8.277%	13.637%

Table 2: *Degree Centrality: statistical measures*

measures, we can conclude that there is a small amount of concentration or centralization in this network. The network centralisation measures are useful "for indicating levels of network coordination" and illustrate to what degree "the entire network is focused around a few central nodes" (Hagen, Killinger, & Streeter, 1997).

In our analysis, three measures (the average path length, clustering coefficient and degree distribution) mirror the intriguing properties of the collocational network and provide the important insights into the structure and growth of language networks in general. The first results of analysing the properties of the collocational network in the OPAALS community show the tendency to the small world structure, which "indicates that language evolution might have involved the selection of a particular arrangement of connections between words" (Ferrer-i-Cancho & Sole, 2001, p. 2264). This means, that the described type of networks has a high representation of cliques, and as the result, has several sub-networks in its structure. The high value of the clustering coefficient provides the evidence for this conclusion. Second concept, the average path length, illustrates the efficiency of a system in which the distance between two nodes in the network is small. The analysis of In- and Out-Degree distributions reveals that the network has a power-law form "allowing for a few nodes of very large degree to exist" (Wang & Chen, 2003, p. 14). In other words, there is a small number of well-connected hubs (= words) and many nodes with few connections. This aspect is crucial in terms of robustness and growth of a network.

The statistical analysis in language networks provides important proper-

ties responsible for robustness and efficiency in their structure and during their evolution. Comparing our results with other investigations of natural language networks, we can conclude that these characteristics tend to be culture independent. These findings are crucial in multilingual interdisciplinary communities such as OPAALS.

7 Conclusion

The complexity and frequency patterns have been studied in the context of Digital Ecosystems, the OPAALS community and the OKS. The theory of Digital Ecosystem in general as well as the development of dynamic interfaces for human-computer(-human) interaction in particular, require the analysis of these patterns. In this deliverable, we argue that the semiotic approach helps to understand the complexity through connecting technological and socio-cultural views and finding the common ground at the more general and abstract level. A variety of systems can be seen as networks with interconnected elements. Hence, the understanding of the structure and dynamics of change (identifying statistical properties) helps in the development of adaptive and dynamic systems. Language is a complex, dynamic and constantly evolving system. It is a network, which is constantly optimized for both robustness and efficiency during the development. If the network characteristics, which are responsible for the efficiency and robustness, are properly identified, we might be able to apply them in the modelling of different kinds of networks (e.g. ontologies, communicational networks, etc.). These characteristics seem to be universal, i.e. culture independent.

The proposed frequency-based approach was exemplarily applied to the "language of OPAALS". The gained results need further investigations in terms of increasing the amount of data, introducing the time dimension (e.g. how the language system develops and changes over time) and applying this study to other types of dynamic systems. In order to understand and implement the dynamics of change in digital environments, we suggest to apply the frequency-based analysis to the following areas:

1. Knowledge Representation system (e.g. ontology-based)
2. Interaction between different systems (e.g. assembly, integration and adaptation of ontologies)
3. Interactions between users through the interface (frequency-based approach within the sign systems)
4. Interactions between systems (e.g. different types of OKS or other interfaces)

More generally, we argue that the semiotic view allows us to generalise and analyse systems independently of their types. All of them are sign systems following certain laws and regulations. Similarly, the frequency-based approach on evolution seems to be universal. Analysing statistical properties of such systems provides information about their structure and dynamics,

which can be considered not only at the theoretical level of DEs, but also at the implementational level of development.

References

Andersen, P. (1997). *A theory of computer semiotics*. New York, USA: Cambridge University Press.

Andersen, P. (2000). What semiotics can and cannot do for hci. In *Knowledge-based systems* (pp. 419–424).

Cancho, R. F., & Solé, R. V. (2003, February 4). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 788–791.

Eco, U. (1976). *A theory of semiotics*. Bloomington: Indiana University Press.

Farghaly, A., & Hedin, B. (2003). Domain analysis and representation. In A. Farghaly (Ed.), *Handbook for language engineers* (pp. 21–58). Stanford, CA: SCLI Publications.

Ferrer-i-Cancho, R., & Sole, R. V. (2001, November). The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, *268*(1482), 2261-2265.

Grudin, J., & Norman, D. A. (1991). Language evolution and human-computer interaction. In *Proc. 13th annual conference of the cognitive science society* (pp. 611–616).

Hagen, G., Killinger, D. K., & Streeter, R. B. (1997). An analysis of communication networks among tampa bay economic development organizations. *Connections*, *20*(2), 13–22.

Hatzigeorgiu, G. M. N., & Carayannis, G. (2001). Word length, word frequencies and zipfs law in the greek language. *Journal of Quantitative Linguistics*, *8*(3), 175–185.

Hjorland, B. (2002). Domain analysis in information science: Eleven approaches - traditional as well as innovative. *Journal of Documentation*, *58*(4), 422–462.

Johansen, J., & Larsen, S. (2002). *Signs in use*. New York, USA: Routledge.

Khaled, R., & Noble, J. (2004). Extreme programming system metaphor: A semiotic approach. In *Proceedings of the 7th international workshop on organisational semiotics.*

Knudsen, T. (2001). Zipfs law for cities and beyond: The case of denmark. *American Journal of Economics and Sociology*, *60*(1), 123–146.

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 1842–1845.

Lindgren, R., Hardless, C., Pessi, K., & Nulden, U. (2002). The evolution of knowledge management system need to be managed. *Journal of Knowledge Management Practice*, *3*.

Makaruk, H. E., & Owczarek, R. (2008). Hubs in languages: Scale free networks of synonyms. *CoRR, abs/0802.4112*.

Motter, A. E., Moura, A. P. S. de, Lai, Y.-C., & Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E, 65*.

Nemeth, G., & Zainko, C. (2002). Multilingual statistical text analysis, zipfs law and hungarian speech generation. *Acta Linguistica Hungarica, 49*(3–4), 385–401.

O'Neill, S. (2003). Theory and data: The problems of using semiotic theory in hci research. In J. Rimmer & B. Zayas (Eds.), *Square pegs in round holes? the relationship between empirical research and theoretical frameworks. the proceedings of the 7th human centered technology workshop.* University of Sussex.

O'Neill, S. (2008). *Interactive media: The semiotics of embodied interaction.* London: Springer Publishing Company, Incorporated.

Parhi, M. (2008). *Impact of changing facets of inter-firm interactions on manufacturing excellence: A social network perspective of indian automotive industry* (Working Papers of BETA). Bureau d'Economie Thorique et Applique, UDS, Strasbourg.

Tanaka-Ishii, K. (2010). *Semiotics of programming.* New York, USA: Cambridge University Press.

Tullo, C., & Hurford, J. (2003). Modelling zipfian distributions in language. In S. Kirby (Ed.), *Proceedings of language evolution and computation workshop* (pp. 62–75). Vienna.

Van House, N. (2004). Epistemic machineries of environmental online communication. In A. Sharl (Ed.), *Environmental online communication* (pp. 199–208). London: Springer.

Wang, X. F., & Chen, G. (2003). Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE, 3*(1), 6–20.

Watts, D. (2003). *Small worlds: the dynamics of networks between order and randomness.* Princeton, NJ: Princeton University Press.