OPAALS PROJECT

Contract n° IST-034824

WP6: Socio-Constructivism and Language

Del6.8 – Models of the Evolutionary Framework for Language

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

Contract Number: IST-034824

Project Acronym: OPAALS

Deliverable N°: 6.8

Due date: May 2009

Delivery Date: May 2009

Short Description:

This deliverable covers the mechanisms and driving forces of natural language evolution. Within the theoretical framework we link the findings to the research on Digital Ecosystems. Special attention is paid to the evolutionary processes occurring in Formal Knowledge Spaces.

Author: Oxana Lapteva (UniKassel), Hagen Peukert (UniKassel), Ossi Nykänen (TUT), Raimund Eder (SUAS), Frauke Zeller (UniKassel)

Partners contributed: UniKassel, TUT, SUAS, LSE, IITK

Made available to: Consortium

Versioning

Version	Date	Name, organization
V1	29.01.09	Hagen Peukert (UniKassel), Oxana Lapteva (UniKassel), Ossi Nykänen (TUT)
V2	25.03.09	Hagen Peukert (UniKassel), Oxana Lapteva (UniKassel)
V3	08.05.09	Hagen Peukert (UniKassel), Oxana Lapteva (UniKassel), Raumund Eder (SUAS)
V4	20.05.09	Hagen Peukert (UniKassel), Oxana Lapteva (UniKassel), Ossi Nykänen (TUT)
V5	25.05.09	Hagen Peukert (UniKassel), Oxana Lapteva (UniKassel), Frauke Zeller (UniKassel)

Quality check

Internal Reviewers: Dr T.V.Prabhakar (IITK), Gerard Briscoe (LSE)

Dependences:

Achievements*	Accomplished work: – critical analysis of the biological and cultural views on natural language evolution – identifying the underlying laws and driving forces of evolution through the lens of natural language – Theoretical framework of language evolution in Digital Ecosystems – Evolution in Formal Knowledge Spaces
Work Packages	The mechanisms of language change and variation provide an important input for any system development in the context of its dynamic character, self-organisational aspects and evolution. The frequency-based approach of evolution is revealed to occur not only in natural language systems. It can be found in networks (e.g. P2P), knowledge spaces and dynamic interfaces. These aspects illustrate a close connectivity to the research in several WPs (WP2, WP3, WP5, WP10). Furthermore, this work provides a different view on the evolution (in comparison with biological explanations) and its underlying laws by looking at usage frequency. This opens a new platform for discussion, collaboration and integration with the natural science domain (WP1) in the OPAALS research community.
Partners	ALL
Domains	Linguistics: aspects of natural language evolution Computer Science: evolution in "formal" environments (knowledge spaces) Human-computer interface: possible applications of the frequency-based theory
Targets	OPAALS researchers, Scientific communities
Publications*	1. Nykänen, O. (2009). Semantic Web for Evolutionary Peer-to-Peer Knowledge Space. In Birkenbihl, K., Quesada-Ruiz, E., & Priesca-Balbin, P. (Eds.) *Monograph: Universal, Ubiquitous and Intelligent Web*, UPGRADE, The European Journal for the Informatics Professional, Vol. X, Issue No. 1, February 2009, ISSN 1684-5285, CEPIS & Novática. Available at http://www.upgrade-cepis.org/issues/2009/1/upgrade-vol-X-1.html
PhD Students*	Oxana Lapteva – biological and cultural factors of language evolution – language evolution in Digital Ecosystems – language networks

Outstanding features*	Integration of the linguistic perspecive on evolution into the research on Digital Ecosystems Discovery of the common mechanisms and laws of evolution in "natural" and digital/formal systems
Disciplinary domains of authors*	Hagen Peukert (Computational Linguistics), Oxana Lapteva (Computational Linguistics), Ossi Nykänen (Computer Science), Raimund Eder (Computer Science)

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

List of Tables

1 Executive Summary

The Evolutionary Framework for Language is an integrative and collaborative research arena focusing on the dynamics and driving forces of evolution. Starting with the investigation of natural language change and variation, this deliverable aims to find the core-mechanisms of evolution and self-organisation and apply them to the domain of Digital Ecosystems. Special attention is paid to the evolutionary processes in knowledge spaces.

2 Scope of this deliverable

Different views on the problem of language evolution in Digital Ecosystems create an intricate theoretical platform dealing with dynamic processes of change and self-organisation as well as their underlying laws and forces. This work aims to explore the mechanisms and driving forces of natural language evolution and link them to the research on Digital Ecosystems and Formal Knowledge Spaces. The mechanisms of language change and variation provide an important input for any system development in the context of its dynamic character, self-organisational aspects and evolution. This research establishes a different view on evolution (in comparison with the biological and cultural explanations) and its underlying laws by looking at the usage frequency. Our investigations reveal that the frequency-based approach explaining the natural language change and variation can be effectively applied to the evolving formal systems (networks, knowledge spaces, dynamic interfaces, etc.).

3 Introduction

In the context of Digital Ecosystem (DE) we explore different aspects of language evolution in order to understand the underlying laws and mechanisms driving a system (language, knowledge system, network, digital environment) to change. Natural language evolves over time due to a variety of factors (e.g. language contact, communication) affecting different levels of language organisation (phonology, lexicon, semantics, morphology, syntax). The linguistic view on language evolution provides interesting insights into the research on evolutionary processes and mechanisms occurring in digital environments, not only in respect to the formal languages and formal representations, but also in the context of a system's design.

The deliverable is structured as follows: First, a critical review of theories on natural language evolution is presented. We look at the cultural and biological views of it from a linguistic perspective. Considering the evolutionary aspects of natural language and formal systems in DEs, this research reveals the main drawbacks of these theories. Second, the frequency-based approach proposed by Haspelmath (2008b) is discussed with regard to natural language evolution. As opposed to the cultural and biological theories, the frequency-based approach proves to be more appropriate as an explanatory model. To illustrate, we survey the critical fields in the research of language evolution. Third, the discovered mechanisms and factors are applied to the different areas of research on DE. In this section, language networks, evolution of programming languages and Human-Computer Interaction are investigated. Language networks are frequency dependent and as such undergo the same mechanism of change. Finally, the evolutionary aspects in formal knowledge spaces are covered. It comprises the characterisation and analysis of evolution in the context of information and knowledge management, which is one of the major objectives of OPAALS.

We argue that all of the above is a necessary prerequisite for designing DEs and knowledge platforms as in OPAALS. Hence, the integration of "natural"(e.g. human language) and "formal" (e.g. formal languages, HCI, formal knowledge spaces) constituents existing within Digital Ecosystem through the lens of evolution is the leitmotif of this deliverable.

4 Natural Language Evolution: objectives and theoretical findings

4.1 Overview

In biology, evolution is the change in the inherited traits of a population from generation to generation. These traits are the expression of genes that are copied and passed on to offspring during reproduction. Mutations in these genes can produce new or alternative traits, resulting in heritable differences (genetic variation) between organisms. New traits can also come from transfer of genes between populations, as in migration, or between species, in horizontal gene transfer. Evolution occurs when these heritable differences become more common or rare in a population, either non-randomly through natural selection or randomly through genetic drift. The key point here is mutation, change (variation), initially random variation is directed by the evolutionary process through what is known as natural selection. So, the process of evolution requires this circular process of variation and selection (Figure 1).

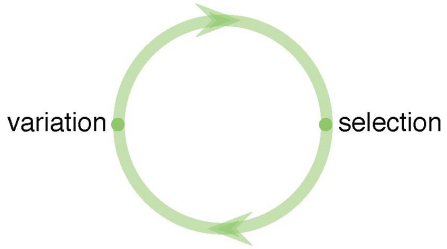


Figure 1: *Abstract representation of evolution*

In linguistics an evolutionary framework for language can be viewed from different perspectives. First, how did language evolve in the last 70 thousand or so years from simpler communication systems? Second, once language is there as a species-specific property, how does/did it change? Although taking account of the different approaches of investigation, in both cases a researcher focuses on finding the underlying laws and driving forces of an evolutionary process. While the first question centres around the emergence of language, the second one looks at how this system has changed. In both cases, evolutionary processes, that is, developmental laws, are under scrutiny.

In chapter 5, we would like to find out what underlying law could possibly explain these processes. We suspect it to be self-organizing in nature as opposed to the prevalent mainstream of linguistic research that tries to explain evolutionary phenomena by, for example, universalist (e.g. Jackendoff

(1977a)) or cognitive-functional (e.g. Croft (2008)) theories. Of course, none of the mentioned can be proven wrong; all are commendable findings. However, they are not compelling. What is missing is a theory that unites all of the varieties, exceptions and circular adjustments given in each of the approaches.

Provided that we can convincingly show that usage frequency as suggested by Haspelmath (March 2008)[1] can be seen as a simple but well-explaining theory of all of the above, it is reasonable to make predictions for other communication systems using natural language. On this basis, we can then make some suggestions for improving the design of Open Knowledge Spaces (OKS) and consider possible developments when representing content in natural language to the public as well as deal with software requirements from a user perspective. Against this background, we are going to explore different aspects of natural language evolution and link them to the research on Digital Ecosystem.

4.2 Objective: Evolutionary Linguistic Analysis as an indispensable Input to Open Knowledge Space Architectures

Why would we need a linguistic analysis of language change to construct knowledge spaces? The answer becomes obvious when thinking about the components of a knowledge space. In the first place, a knowledge space should not be a static collection of data. It should be as dynamic as knowledge changes. One should pay close attention to the user who is working with digital representations. Knowledge has to be represented in some form. Most of the time sign systems as natural language or pictographic visualisations are used. This is comprehensible because the human perceptory system is the connecting link between knowledge processing and its outer representation. Hence, these sign systems are as dynamic as other cognitive phenomena underlying the human mind: language, customs, and values should be treated as such. Thus a knowledge space should adapt constantly to the needs of the user and the knowledge it is representing. All of that is highly dynamic. To understand these dynamics, one has to look at the dynamics of its components, of which one is language.

[1] The findings on usage-based frequency presented here are primarily the work of Martin Haspelmath taken from presentations given in March 2008 in Leipzig (Haspelmath, March 2008) or recent publications (Haspelmath, 2008a, 2008c, 2008d). We merely intend to underline the benefit of the frequency-based theory of syntactic structures when compared to other sign systems such as vocabulary evolution in ontologies, networks, or knowledge spaces.

Second, the use of vocabulary in some of the recently developed models in OPAALS needs existential grounding. As knowledge itself, vocabulary is changing; it is dynamic. Thus, it should be of great interest how these changes take place and whether they are predictable at least to the extent that is useful in the design of an ontological representation of these vocabularies.

Systematic and scientific research on language and language change started some 200 years ago. Knowledge acquired during this time frame amounted to a substantial body of work that can also be exploited for other research areas. Most of the time, linguistic research on language change concentrated on detailed descriptions of how a sound, a word or a phrase etc. changed without making recourse to a general theory explaining the driving force behind the change. Recently, an attempt has been made to base all observations on the usage of frequency (Haspelmath, 2008a).

Taking the usage frequency as an explanatory model seems promising for language as we intend to show in the following sections. In addition, it should be easily aligned with other components of the Open Knowledge Space (OKS)[2]. Usage frequency claims that frequently used constructions decrease in complexity. Conversely, simplifying a well-defined system at one point might lead to a more complex pattern at the other end. Hereby, the connection between both changes is far from clear.

Intuitively, one may think of other formal languages in which simplifications take place. In formal programming language expressions like $i = i + 1$ become $i{+}{+}$ because they are used quite often, that is, their usage frequency is high. Structures such as for-loops become easier to handle by defining more specifics in the initials. Within Logic, all logical constructions can be reduced to combinations of $\wedge$ and $\neg$. Since the notational system looses tangibility when depicting logical constructions of larger sizes, three more operators are defined that shorten a complex arrangement of $\wedge$ and $\neg$. However, these five operators are not prolonged, one could at least theoretically consider other operators as done in computer algebra (XOR or $NAND$). Five seemed the optimal number of operators to make the logic system as efficient as possible. Mathematics is yet another example. Exponential expressions are just a shortening of a long line of adding up the same number. Multiplying can be seen as a similar shortcut for adding the respective numbers. In these cases, it seems that a middle of the road compromise is taken between the complexity that emerges out of huge quantities of equal entities and structural complexities that come up when different symbols are combined in a certain way.

[2]See section 6.4.4 as a possible application of the frequency-based model for the OKS

4.3 Language evolution: biological and cultural considerations

In the context of natural language evolution, one of the crucial questions discussed in the literature is "how is the human language system transmitted? Is it primarily in a genetic fashion (through the human genome)? Or is it primarily in a cultural fashion (through learning)?" (Steels, 2004, p. 72). There are mainly two types of research directions:

- Language evolution as a cultural phenomenon
- Language evolution as a biological phenomenon

In the following sections we are going to discuss these approaches and critically review their explanations of language evolution.

4.3.1 Language Evolution as a cultural phenomenon

The first view sees the reason and effects of evolution in cultural phenomena. This position relies primarily on learning and language contact as the ways in which language is transmitted and changed. In the cultural framework, language users are considered to be the active agents that "shape and reshape their language" (Steels, 2004, p. 76). In terms of coherence, the researchers focus on language that all the members of a community must share (e.g. the same language conventions and the same conceptualisations).

In the context of the cultural view on language evolution, one of the recent trends is to explore languages as self-organizing systems in order to investigate:

- Origins of natural languages (Steels, 1997)
- The evolution of natural languages (Croft, 2000)
- Dynamic processes of language evolution at a cultural level (Vogt, 2007)

One of the approaches proposed within this view on language evolution is Darwin's theory of evolution as applied to the cultural dynamics of change. The main idea is that natural languages evolve through variation, competition and selection of language material (see Croft (2000); Mufwene (2002); Vogt (2007), for more details), rather than of genetic material.

4.3.2 Language Evolution as a biological phenomenon

The biological view on language evolution considers language not as a cultural artefact, but as "a distinct piece of the biological makeup of the brain" (Pinker, 1994, p. 4). The observations supporting this approach refer to pattern similarities (e.g. structural features) among different languages and to the amazing ability of children to learn language(s) very fast independently of their cultural background. In order to explain the roots of these features, several researchers proposed the existence of a specific language faculty (Chomsky, 1972), a bioprogram (Bickerton, 1984), or a language instinct (Pinker, 1994). Different approaches within the biological view on language evolution try to find a criterion or function that could explain the emergence of human language. According to Pinker and Bloom (1990), the biological view "offers clear criteria for when a trait should be attributed to natural selection: complex design for some function, and the absence of alternative processes capable of explaining such complexity. Human language meets this criterion: grammar is a complex mechanism tailored to the transmission of propositional structures through a serial interface"(p. 707). This apparatus is the Universal Grammar (UG) that dominated linguistic theory for some decades in the last century and has exerted an enormous influence on the linguistic community up to the present day. Syntactic structures of natural languages is the last resort of purely human language abilities that have never been detected in any other species, which is why they are seen as part of the genetic endowment of humans. Complex syntax is the key argument for researching the paradigm of a restricted framework, that is, Universal Grammar. This innatist approach describes language phenomena that should be general enough to be applicable to all languages in the world. When syntax is genetically encoded, it must function independently of a specific language and a subset of common language principles and rules must exist upon which syntactic structures of all the languages rely.

Probably the most influential of these principles explained by X-bar Theory (Chomsky, 1970; Jackendoff, 1977b; Stowell, 1981). In short, X-bar claims that each word X is a head of its projected phrase XP whereas X stands for all possible grammatical categories[3]. Generally X-bar is specified as:

$$XP \rightarrow Y[x'XZP]. \tag{1}$$

This rule allows instantiations for English as in the following sentence:

- $NP \rightarrow D[n'NPP]$] the [horse on the meadow]

[3]For example, N = Noun, V = Verb, NP = Noun Phrase, VP = Verb Phrase, P = Preposition, PP = Prepositional Phrase, Adv = Adverb

- $VP \rightarrow Adv[v'VNP]$ often [eats a flower]

- $Adv \rightarrow Adv[p'PNP]$ right [under the tree]

Provided that the observation of X-bar-Theory is correct (and it seems to be true for the investigated languages so far), the question arises why forms like

$$NP \rightarrow [v'VPP]^4 \tag{2}$$

do not show up in the world's languages. The nativist[5] description is unsatisfactory since it only claims that the restrictive framework of the Universal Grammar accounts for a certain set of syntactic structures but not for others. In this respect, the argument is circular because the premise is taken as a prove. The Language Acquisition Device (LAD) could not acquire the rule in (2); it only allows the recognition of structures of the specified form in (1). Without the innatist claim, there is no other explanation (Haspelmath, March 2008). Of course, linguistic explanations in generative and functionalist domains are no explanations in the strong sense of the word. They happen to emerge over time. Researchers try to find general rules and laws that explain these historical accidents.

- The dog chased the cat

- *The dog the cat chased

While the first example emerged as the correct sentence in English syntax, the second one has an incorrect word order. Possibly, the second item could have developed as a correct sentence and it could have been explained by the same general rule. Despite of all this, artefacts may remain which are not explained uncontroversially. They do not fit the pattern of the general rule. Yet, once it is found that language-specific data support a general rule without exceptions, later they become a general universal fact. This universalist perspective is not to be confused with nativist approaches that also obey the procedure just stated but make further claims that are incompatible with universalist approaches.

On the basis of Chomsky's concept of UG, the biological view represents the idea that the brain must contain a program or procedure that enables the process of constructing an unlimited number of sentences from a finite collection of words. As a part of it, the "biolinguistic perspective" focuses on the traditional problem of determining the specific nature of the faculty

[4]An example of such structure would be *The spider crawled along.*

[5]In linguistics, nativists believe that certain language phenomena are genetically inherited, i.e. they are not learned from the environment.

of language and reinterprets it as the problem of discovering the genetic endowment that underlies the acquisition and use of language (Oller, 2004).

Some interesting investigations have been carried out towards determining analogies and resemblance between natural language and DNA (Steels, 2004; Sereno, 1991). It has been argued that the associative character and similarities between these two systems could help to explain the mechanisms of human cognition and language dynamics. The possible analogies are represented in Fig. 2.

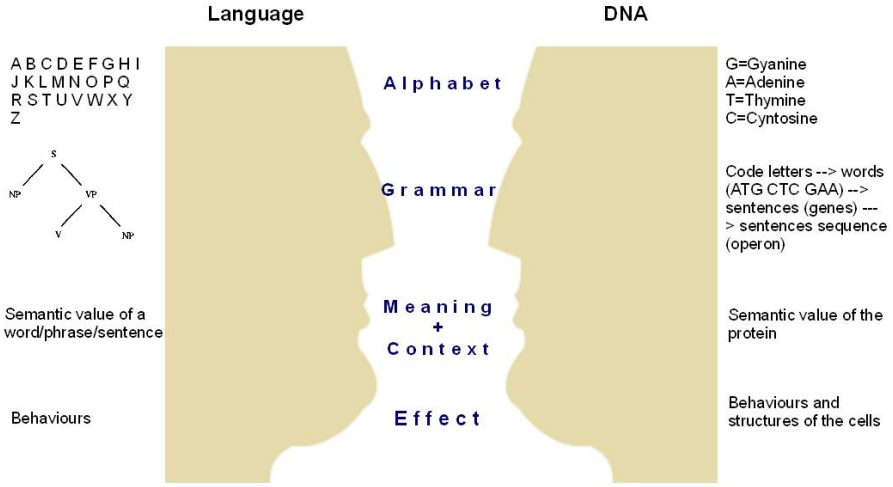


Figure 2: *Natural language and DNA*

One of the crucial questions related to this research is the ability to go beyond such analogies. Can we find the explanation of language evolution by looking at it through the lens of DNA? How can DNA research help us to model the evolutionary framework? The impact of finding "the key" to the question on language change and variation based on the DNA does not go beyond analogies between these two systems. The interaction between DNA and human behaviour is still not understood. Hence, the complex systems (language, social environment, and others) and their behaviour cannot be reduced to the DNA[6]. This approach does not explain the core-mechanism(s) of language transmission leading to variation and change. One of the reasons is the complexity of human language. At the current stage of research, this direction does not lead us towards identifying the underlying law of language evolution, and neither towards understanding the dynamics of language at the natural and formal levels within Digital Ecosystems.

[6]See Marcus (2005) for a discussion.

The biological and cultural impacts on language evolution provide interesting insights into its roots, mechanisms and driving forces. However, each theory has its own drawbacks in providing a "unique" answer to all the related aspects of natural language evolution. The cultural explanation which assumes learning as the underlying mechanism of language evolution still has a problem of explaining the mechanisms of sharing.

The biological explanation, by contrast, faces the problem of accounting for the rapid rise and spread of new language items (e.g. new concepts) in every human language. Consider, for example, the technological progress forcing the enormous expansion of our vocabulary and the rapid semantic change of words. Many concepts of the Internet are already a stable part of human vocabulary: *home page*, *server*, *browser*, *e-mail*, etc. Hence, from a historical point of view, the language change that has occurred during the last 40-50 years is enormous.

Another issue the biological view may face is the question of storage. According to Worden (1995), humans do not have enough storage in genetic form for the huge amount of data (e.g. all types of linguistic information) they need to process day by day.

In conclusion, the biological and cultural approaches have difficulties in explaining two crucial aspects of natural language evolution:

- strong variation and rapid evolution
- mechanisms of language transmission.

5 Frequency as a core mechanism in language evolution

5.1 Factors of language evolution

First, we should define more precisely what is understood by a frequency-based approach. In simple terms, frequency-based models explain changes in patterns – be they mere linguistic data or data in general – by a predictor that is a frequency distribution. It states that a more or less unilateral relation between the changed pattern and the frequency of the parameters exists. This allows to make predictions on the dependent variables, for example word length, given the independent variable, for example the usage frequency of the word. In this particular case we predict that the word length decreases while the usage frequency of the word increases. Likewise, we can observe that frequently used data structures become simplified. Based on that approach, we will now turn to a collection of linguistic data and see how this approach applies to it.

What are the potential triggers and motivations of language change? One important factor is linguistic economy, i.e. the tendency towards efficiency and simplification in the linguistic system. The motivation of the language change can be often an analogy, e.g. the development of new forms based on similarity to already existing forms within the same language (Table 1).

to lase	laser
to bake	baker
to research	researcher
to develop	developer

Table 1: *Analogy as one of the motives of language change*

There are two important characteristics that we might identify not only in the context of natural language change and variation, but also within the scope of human communication, Digital Ecosystems, Human-Computer Interfaces and many other systems. These are power and efficiency that compete in the evolutionary process of language change and variation. Power is related to the "relative ability of the system to transmit the information or manage the social relationships that might be relevant to survive", whereas efficiency is the "relative ability to communicate rapidly and at low cost in energy". (Oller, 2004, p. 51).

The principle of efficiency shows up in a variety of ways. One illustrative example is the fact that most common words are short and often (depending

on the language) monosyllabic. Table 2 illustrates the most frequent words in the OPAALS text corpus [7].

Words	Counts
the	8576
of	4617
and	4277
to	4091
a	3035
in	2610
is	2164
for	1562
be	1538
that	1515
as	1218
this	1179
are	1002
it	979
on	951

Table 2: *Frequent words in OPAALS corpus*

With regard to natural language we refer to the aspect of efficiency as a "Principle of Least Effort" (Zipf's law). According to this principle, "the rank of a word (in terms of its frequency) is approximately inversely proportional to its actual frequency" (Tullo & Hurford, 2003).

Efficiency in natural language is tied up with the process of simplification. Abbreviation and different kinds of shortening of complex structures are usually based on the usage frequency, in other words we tend to minimize the size of frequent words in our vocabulary. Table 3 below illustrates different mechanisms of simplification (abbreviation/acronyms, clipping, blending):

Language evolution provides another way of shortening in order to obtain efficiency. According to Grudin and Norman (1991), irregular verbs have shorter forms than their regular counterparts. They examined a list of 173 irregular English verbs and found out that in 166 cases the irregular form was shorter, in 6 cases of the same length, and in only one case the irregular form was longer than its regular counterpart. This case was the verb *bought* (in comparison to *buyed*). Bybee (1988) argues that an irregular verb reverts to the regular form when the frequency of its usage drops.

[7]The OPAALS text corpus has been taken from the Metaphorological Tool Kit.

Abbreviation/acronyms	
TV	television
Radar	RAdio Detection And Ranging
Laser	Light Amplification by the Stimulated Emission of Radiation
BASIC	Beginners' Allpurpose Symbolic Instruction Code
Clipping	
Phone	telephone
Zoo	zoological garden
Fridge	refrigerator
Fax	facsimile transmission
Blending	
Vegeburger	vegetarian + hamburger
Smog	smoke + fog
Motel	motor + hotel
Pulsar	pulse + quasar

Table 3: *Simplification mechanisms*

The question we would like to ask within the evolutionary framework for language is: *what is the mechanism underlying the power and efficiency of language, and which, at the same time, explains the origins of language change and variation.*

5.2 Diachronic Functional Explanations

Working on language change has an aftertaste of merely collecting descriptive data sets without providing real explanations for the observations. Such research results describe how language phenomena have altered, however, ignoring detailed expositions on the interrelatedness of their data collections. Frequency-based explanations are different in that sense. In fact, they are not restricted to synchronic observations. There is a strong tendency for frequent forms to be reduced diachronically (Haspelmath, 1999). Asymmetries in the morphological or syntactic structure (observations that are not in line with a general rule in a certain research paradigm) might go back to past frequency asymmetries. Greenberg (1966, p. 32) makes reference to the example of singulars being universally more frequent than plurals (see Table 4). Within the framework of usage frequency this is not a linguistic phenomenon, but only a reflection of the world as it is: single items are generally more in use than several items, which is then reflected in the grammar of a language.

	Singular	Plural	Dual	number of nouns
Sanskrit	70.3%	25.1%	4.6%	93,277
Latin	85.2%	14.8%		8,342
Russian	77.7%	22.3%		8,194
French	74.3%	25.7%		1,000

Table 4: *Singular and plural frequency for different languages*

most common x-syllable word	word	number per mil. words
1-syllable	the	61,847
2-syllable	into	1,634
3-syllable	government	622
4-syllable	information	386
5-syllable	international	221
6-syllable	responsibility	93

Table 5: *Frequency as a predictor for syllable length (Haspelmath, March 2008)*

Bearing this in mind, it is now easy to explain once incompatible artefacts. To give an example, the Latin plural word for *wolf* is shorter than the singular. This indicates that the plural *lupi* was more often referred to than the singular *lupus* which can be explained using real world knowledge: for Romans, wolves were seen as an acting group. It is then interesting to see that a particular language derived from Latin, i.e. Spanish, does not follow the attested pattern. As in English, the plural is indicated by adding an s-morpheme to the stem. Does this mean that the conception of wolf's life has changed? Possibly, but even if this is not the case, Spanish had to compensate for an impoverished morphological system. Thus, the language did no longer mark the accusative, which is not used as often as for example the nominative. These compensations finally led to the longer plural form in Spanish.

As we have seen, frequency effects have a great potential for functional explanations in language change. Provided that language is constantly changing, we can look at a simple frequency count of multi-syllabic words in English to see that the more frequent expressions tend to be shorter (see Table 5). Even though it seems intuitively right, this approach is hotly debated in the linguistic community (see Chapters 5.4.1 through 5.4.2). Here the methodology is quite easy. Look at (morpho-)syntactic asymmetries and figure out

	Singular	Plural	
Nominative	val	valjos	'horse(s)'
Accusative	valez	valjosty	'horse(s) direct object'
Ablative	valls	valjosles	'from the horse(s)'
Absessive	valtek	valjostek	'without the horse(s)'

Table 6: *Udmurt's case system (taken from Dayley (1985))*

whether their frequency patterns show such asymmetrical behaviour.

For single word counts the matter seems clear, but discussions tend to emerge on more complex syntactic phenomena:

(1) The dog chased the cat in the garden.
*The dog chased at the cat the garden.

Considering example (1), Haspelmath (March 2008) asks why the locative phrase works with a preposition, while the patient phrase does not? Why is there no other language in which this is the case? According to the frequency-based approach, this is because patient constructions are more frequent than locatives. Again, more frequent expressions tend to be shorter.

5.3 Forms of Frequency Effects

We would now like to illustrate three types of frequency effects applied to linguistic domains: relative frequency of paradigmatic alternatives, absolute frequency of word forms, and type frequency.

5.3.1 Relative frequency

Relative frequencies can be seen in linguistic categories ordered by frequency. This is also the case for other efficient sign systems, e.g. phone numbers, where one can predict that the more frequent forms (e.g. area codes) are shorter. The same is true for linguistic categories. Yet, as opposed to the unilateral dimensions of ordering phone numbers, we have to note an important addition since linguistic categories are interdependent on each other. Whenever two forms are related, the more frequent form tends to be zero-coded. Case-marking languages are a typical example. Udmurt, a Finno-Permic language, has a case system that shows these features in two dimensions (see Table 6).

First, one should notice, that one dimension is case and the other is number. Since singulars are more frequent than their plural counterparts,

Dimension	Categories (ordered by frequency)
number	singular > plural > dual
case	nominative > accusative > dative
person	3rd > non-3rd (1st/2nd)
degree	positive > comparative > superlative
voice	active > passive
mood	indicative > subjunctive
polarity	affirmative > negative
tense	present > future

Table 7: *Frequent and rare categories*

	gwel-d (see)	myn-d (go)	gwneu-d (do)	do-d (come)
1 SG	gwel-es i	es i	nes i	des i
2 SG	gwel -est ti	est ti	nest ti	dest ti
3 SG	gwel-oss e	aeth e	naeth e	daeth e

Table 8: *Suppletion and other irregular forms in Welsh*

they are shorter in length. By the same token, nominatives are more frequent than accusatives, nominatives are shorter. Now the nominative singular form is related to its plural counterpart as well as to accusative, ablative, absessive and it is zero-coded, that is, the root of the word which is never shortened anywhere else. These findings can be summarized in table 7 (Haspelmath, March 2008).

5.3.2 Absolute Frequency

While the differential for relative frequencies is predictability resulting in shortness of more frequent forms, for absolute frequencies it is memorizability producing irregular forms. Highly frequent forms show a tendency to suppletion and irregularity as the example found in King (1985, p. 183) reveals (see Table 8).

5.3.3 Type Frequency

Depending on the productivity of type-frequent patterns as a function of the lexical strength in memory, type frequencies impact on new word formation processes. To illustrate, the German plural of *-e* has a very high type frequency. There are some hundreds of nouns that take this suffix. It

is therefore productive. Consequently it is also applied to new nouns as *fax* that becomes *faxe*.[8] To make this point clear, another German plural form is *-er* (as in *Buch* and *Bücher*) and has got a low type frequency. So it is rather unproductive and not applied to new words.

5.4 Explaining Markedness and Iconicity from a Frequency – Based Approach

Markedness and iconicity are the fields of investigation that should be of highest advantage when natural language is used in ontologies and knowledge spaces. These fields are typically addressed in linguistics when general rules are scrutinized. Especially iconicity should be of particular interest since principles of analogy and similarity are often applied to isomorphic or iconic processes of visualisation. As such, they are not only language-specific but also cognitive in nature. We could say that the OKS meets all the properties of iconic signs. Presenting a general principle of how iconicity evolves, one can reasonably conclude that these underlying laws are true for all icons and therefore apply to OKS-platforms as well.

5.4.1 Iconicity of Complexity

Reviewing the literature of the last four decades, one can observe an astonishing agreement on the question of markedness matching. Lehmann (1974, p. 111) writes that the complexity of the semantic representation correlates with the complexity of its phonological representation. The same idea is expressed by Mayerthaler (1988, p. 25):

> What is more semantically should also be more constructionally.

Givon (2008, 2.2) says that a larger chunk of information will be given a larger chunk of code. Haiman (2000) explains this idea in terms of the abstract nature of concepts:

> The more abstract the concept, the more reduced its morphological expression will tend to be. Morphological bulk corresponds directly and iconically to conceptual intension (p. 283).

Finally, Langacker (2000, p. 77) notes that the iconicity of the preposition *of*, its phonological value and its semantics are apparent because *of* is the most

[8] fax is just another example of a shortening process due to frequencies. It was long used as facsimile by a small group of people. Once it became part of the vocabulary of everybody and used more frequently, it was shortened to *fax*.

tenuous of all English prepositions. In general terms, more complex meanings are expressed by more complex forms. Often it is described as an iconicity of markedness matching. Put differently, marked meanings are expressed by marked forms. Even Jakobson (1966) argues for this observation:

> language tends to avoid any chiasmus between pairs of unmarked/marked categories, on the one hand, and pairs of zero/nonzero affixes [...] on the other (p. 270).

Plank (1979, p. 139) elaborates on this phenomenon as follows:

> The formal markedness opposition iconically mirrors the conceptual-semantic markedness opposition.

Haiman (1980, p. 528) argues that morphologically and syntactically marked categories are also marked semantically. Givon (2008) delineates a *meta-iconic-markedness-principle* stating that all categories that are cognitively marked (and these are complex!) have a tendency to structural markedness. Aissen (2003, 3) believes that iconicity favours the morphological marking of syntactically marked configurations. In summary, formally marked means that something is expressed overtly.

Typical examples are given in Table 9. Greenberg (1966) noted these universal formal asymmetries without invoking iconicity to explain them. However, he believed indeed that frequency asymmetries explain formal asymmetries: on this account, less marked forms would be more frequent and more marked forms would be less frequent across languages.

In what follows, we adopt Haspelmath's conception of the frequency-based approach (Haspelmath, 2008c) and argue against iconicity of complexity (as argued by Croft, 2003, Haiman, 1980, Haiman, 1983, Haiman, 1994, Haiman, 2000).

Iconicity of complexities makes wrong predictions (see Table 10). Too many exceptions have to be added, all of those were unknown at the time Greenberg (1966) published his work. Frequency, by contrast, makes the right predictions in all of these cases. So it meets necessary and sufficient conditions for explaining the relevant phenomena. Langacker's (2000) argument of the preposition *of* can be answered by the same logic. *of* is not only the most semantically tenuous, but also the most frequent of all English prepositions.[9]

When Lehmann (1974) or Haiman (2000) assert that grammatical morphemes are universally shorter than lexical morphemes and that this iconically mirrors a more abstract or complex meaning, we can now account for

[9]As we have already presented in Table 2, this is also true for the OKS.

	less marked/unmarked	(more) marked
number	Singular (tree-∅)	Plural (tree-s)
case	Subject (Latin: homo-∅)	Object(homin-em)
tense	Present (play-∅)	Past (play-ed)
person	Third (Spanish: canta-∅)	Second (canta-s)
gender	Masculine (petit-∅)	Feminine (petit-e)
causation	Non-causative	Causative
	(Japanese: ik-u = go)	(ik-ase-ru = make go)
object	inanimate	animate
	(Spanish: Veo la case)	Veo a la mujer

Table 9: *Examples of markedness matching*

	less marked/unmarked	(more) marked
number	Plural	Singular (tree-s)
	Welsh: *plu* feathers	*plu-en* feather
case	Object case (Latin: homo-∅)	Subject case
	Godoberi: *mak'i* child	*mak'i-di* (ergative)
tense	Present (play-∅)	Past (play-ed)
person	Second P. Imperative	Third P. Imperative
	Latin: canta-∅ (sing)	canta-to (Let her sing)
gender	Female	Male
	English: widow-∅	widow-er
causation	Causative	Non-causative
	German: öffnen	sich öffnen

Table 10: *Against markedness matching*

X word Y	(function-word expression)
X Y	(juxtaposition)
X-Y	(bound expression)
Z	(portmanteau expression)

Table 11: *Cohesion scale after Haiman (1983)*

the same facts by using frequency as an underlying principle without risking any wrong predictions. Iconicity explanations wrongly claim that lexical items with highly abstract or simple meanings should be consistently shorter than items with more concrete or complex meanings. This would predict that the term *entity* should be shorter than *thing*; *animal* should be shorter than *cat*; *perceive* should be shorter than *see*. We can observe that this is not the case.

5.4.2 Iconicity of Cohesion

A second argument mentioned in the literature is that meanings that belong together more closely are expressed by more cohesive forms. In this context, Haiman (1983, p. 782) designed a cohesion scale (Table 11). He notes that the linguistic distance between expressions corresponds to the conceptual distance between them.

Since some authors (Newmeyer, 1992, p. 761) conflate cohesion and contiguity, we define explicitly cohesion preferable to distance and cohesion unequal to contiguity. As an illustration for the iconicity of cohesion, we will consider two examples often encountered in the respective literature: possessive constructions and causative constructions.

Haiman (1983, p. 793) reasons that expressions of indispensable possession (e.g. body parts, kinship) reveal at least the same degree of cohesion as terms for alienable possession (e.g. clothes, house). The reason is that semantically the *possessed* and its *possessor* belong together more closely in one case but are more distant in another. Abun, a West Papuan language, is cited as evidence for the argument (see 2, Berry and Berry (1985)). Note that for English the degree of cohesion is equal.

(2) a. *ji bi nggwe* my garden
I of garden

b. *ji syim* my arm
I arm

In Buru, an Austronesian language spoken in Indonesia, we find an example for causative constructions (Dixon, 1991, p. 69). Haiman (1983) ar-

	X-Y	Z
comparatives	dri-er	worse
past tense	play-ed	went
negation	has-n't	won't
gender	actr-ess	nun
diminutive	pig-let	puppy

Table 12: *Bound and portmanteau expressions*

gues that the degree of cohesion is dependent on the directness of causation. Whenever cause and result belong together more closely (e.g. 2(b)), the degree of cohesion increases. Indirect causation (as in 2(a)) indicates causative constructions with less cohesion. In English, *cause to die* and *kill* make the same point. *kill* implies a direct action, whereas *cause to die* does not necessarily have this denotation.

(3) a. *Da puna ringe gosa*
3Sg.A cause 3Sg.O be.good
He (did something which, indirectly,) made her well.

b. *Da pe-gosa ringe*
3Sg.A Caus-be.good 3Sg.O
He healed her (directly, with spiritual power)

The frequency-based approach argues that absolute frequencies explain the contrast between portmanteau expressions and separate expressions (see Table 12). The items that show greater formal cohesion are more frequent in an absolute sense. Relative frequencies can then explain the contrast between function-word expressions and zero expressions (first two lines in Table 11), but also between juxtaposition and bound expressions (2nd and 3rd line in Table 11).

Croft (2003) asserts that the frequency-based shortness is due to entrenchment.

> The theoretical explanation for economy (e.g. Bybee, 1985) requires absolute frequency. Economy effects are due to degree of entrenchment of linguistic forms (morphological forms or constructions such as the possessive) in the mental representation of linguistic knowledge. Entrenchment leads to routinisation of

> the production of the form by a speaker, which in turn brings about reduction of that form. But entrenchment is a result of exposure to the number of tokens of the linguistic form; that is, entrenchment is a function of the absolute frequencies of forms, not relative frequencies.

Haspelmath (2008d) replies that the view by Croft (2003) cannot be reconciled with some basic facts. Routinisation co-occurs with reduction in form because routinised forms are also predictable for the receiver of the message. This does not mean, however, that routinisation is not the cause of reduction. It is rather the speaker's ability "to save energy when part of the message is predictable" (Haspelmath, March 2008). Whenever predictability is not possible (e.g. when someone dictates a phone number), no reduction takes place. This idea can already be inferred from Zipf's work.

> In listening to spoken language, we notice that, among other things, the speaker invariably emphasizes these two: first, what is new or unexpected to the hearer; second, what the hearer desires to make especially clear [...] But that which is unexpected, unusual, or unfamiliar to the hearer is, by definition, the seldom (Zipf, 1929, p. 5).

Haspelmath (March 2008) rejects the idea of entrenchment and summarizes as follows:

> Thus, frequency-induced reduction is to a large extent a hearer-based phenomenon and is not due to routinisation, but to predictability. It should also be noted that predictability need not be due to linguistic frequency. Stereotypical situations allow massive reduction, simply because the context makes the utterance content easy to predict.

5.5 Specifications of Frequency-Based Contractions

How can we now put some structure into the findings of the frequency-based approach? It seems to be appropriate to choose four categories, in which the linguistic peculiarities fit in quite nicely. In keeping with the observations made, frequent expressions tend to be zero-coded while the rare expression is marked overtly. By analogy, three more categories can be suggested: frequent ones are shorter and the rare ones are longer; existent and non-existent, or lastly, straight-forward and round-about constructions. For each of these, some examples will justify the arrangement made (Table 13).

Frequent expression	Rare expression
(zero-coded)	(overt)
Sg: book-∅	plural: book-s
3rd P. Spanish: canta-∅	2nd P: canta-s
present: I ∅ sing	future: I will sing
(shorter)	(longer)
Tamil inanimate locative: -il	animate locative: -itam
Latin dative Sg.: -o/-ae/-i	dative plural: -is/-is/-ibus
Russian refl.: -sja	ordinary reflexive: sebja
(existent)	(nonexistent)
Tzutujil: w-ati7t	*ati7t
(my grandmother)	(grandmother)
juyu7 (montain)	*w-juyu7 (my montain)
Who do you think that I meet?	*Who do you wonder why I met?
(straight-forward)	(round-about)
Gabriel's friend	a friend of Gabriel's
I gave her it	I gave it to Aisha.
German: Ich will spielen	Ich will, dass Du spielst.

Table 13: *Categories of frequency-induced shortness*

5.6 Pathways to Well-Coded Patterns

In this section we present three mechanisms of how language patterns develop based on frequency functions: phonological reduction, preservation and analogical change – all create economical patterns in language evolution. They illustrate why it is so hard to acknowledge the frequency-based approach as an overall general theory. Still, it is undeniable that frequency is the major force for the changes taken place.

The first mechanism observed suggests that frequent phonological patterns get reduced. For example, linguists can observe a reduction in the marking of the 2nd and 3rd person singular in the development of Polish and Russian out of Proto-Slavic. While *you write* and *she writes* are three-syllable-words in Proto-Slavic, the last vowel was reduced and so they now consist of two syllables in Russian and Polish alike.

The plural and singular indicators were reduced in modern English. The English singular became zero. In Old English *daeg* (Sg.) and *dagas* (Pl.); in Gothic even *dags* and *dagos* were cut off in Modern English as we can see today (day/days). In Russian the reflexive pronoun *-sja* evolved out of the Proto-Slavic *sebja*, which is the same mechanism. In English, the complement clauses of *want* become *wanna* depending on subject agreement.

At first sight phonological reduction seems to contrast with the finding of a *conserving effect*, that is, more frequent constructions are preserved. Really, this is again a matter of predictability, which implies less need of overt coding. In Classical Arabic all nouns can take possessive affixes indicated by *ii*. This frequent pattern is not shortened but preserved. Indeed, the *ii*-form is morphemic, though represented by a phoneme. Since reduction is not possible, the frequent form is preserved, but then applied to all other nouns.

Economical patterns may be a result of analogy as well. In Modern German some markers from Old High German were preserved by analogy similar to the Arabic example just given. *Knoton* (knot) is the accusative singular in Old High German and became *Knoten* in Modern German. *Knoto* is the singular nominative form of knot in Old High German, which also changed to *Knoten* in Modern German. *Affo* (Nom.Sg) and *affon* (Acc.Sg.) took the same course. We can observe that *o* changed to *e* by analogy.

6 Natural Language Evolution in Digital Ecosystem

6.1 Overview

Can we apply the frequency-based approach and its findings to the digital environments? Do efficiency and power occur in digital environments? The answer is yes.

The proposed frequency-based approach helps to understand the driving forces of natural language change and variations. This is true not only for human populations, but also for digital environments. The linguistic perspective on language variation refers to the fact that languages vary according to region, social group, and function. Language changes over time due to the variety of factors such as language contact and new communicative needs. Without change languages die. Living languages are in a continuous process of change and adaptation that may affect all levels of language organisation: phonology, lexicon, semantics, morphology, and syntax.

Let us now turn to the domain of Digital Ecosystem and its functionality: when language is spoken in different areas, changes that occur in one area do not necessary spread to or influence other areas. The same dynamics have to be true for the distributed, self-organised digital worlds. Abstractly speaking, we can say that different peers (in our case languages) interact with each other at different levels of organisation. And the change (evolution) of a system depends on the strength of connectivity between these peers. At this stage of research, we argue that the frequency-based theory (Haspelmath, 2008) leads towards a theoretical framework of evolution (not only language evolution, but also evolutionary processes occurring in Digital Ecosystems) through explaining the mechanisms of change, variation and self-organisation. Consider, for example, the communication processes between users based on any knowledge system. When one user introduces a new expression into the knowledge system (for example, a new name for a business product), the survival of it depends on how often other language users use it while their communicative contacts. Non-usage means the death of a concept. In terms of the peer-to-peer networks, the strength of the "language"-peer depends on its connections, i.e. frequency of use. Through this mechanism, the self-organisational aspects of language and knowledge systems can be explained and traced.

Another important area of applicability of the theoretical framework can be collaborative tagging. Different factors existing in natural language complicate the processes of creating a dynamic tagging platform. One of the

major causes is the fact that many words are polysemic. In collaborative tagging systems, the polysemy occurs when the same word is used for tagging different information (= different meaning). Furthermore, people use a lot of synonyms for their tags, e.g. different terms are used for tags related to the same subject (= same meaning). Another difficulty one faces in collaborative tagging systems is the fact that "different users can use terms at different levels of abstraction to describe the same resource" (Gendarmi, Abbattista, & Lanubile, 2007). Since each user has his or her own individual space of tags, it is difficult to create a collective set that is representative for a specific community. With the frequency-based approach of language evolution we see a high potential of moving towards a dynamic, self-organised tagging system. The same is true for other knowledge representation systems.

As described in the previous chapters, the change itself strongly depends on the statistical properties of usage. We see the high potential of linking the findings from evolution of natural language to the research on Digital Ecosystems. One important consideration that needs to be taken into account is that the system will always change. To make this process dynamic and natural (e.g. self-organised), the statistical properties of usage need to be looked at. The basic law of *frequent use →simplification* has different kinds of appearance which mirror the power and economy of any system (language, network, digital environment).

In the following sections, we are going to discuss the evolution of language from different perspectives:

- linguistic mechanisms of semantic change
- human language as a network including its structure and properties
- evolution of programming and Human-Computer Interaction

6.2 Semantic change and variation in Digital Ecosystem

The evolution of natural language occurs constantly at different levels of linguistic organisation. In the context of a DE, one of the fundamental aspects that is affected through the processes of change is semantic change. From the linguistic point of view, this process involves the creation of new words and semantic change of existing words. This occurs in the form of several mechanisms. The creation of new words, for example, is done through compounding, derivation, back-formation, clipping, and blending (Trask, 1994).

1. Compounding refers to the process of combining existing words into a new one, for example: *blackboard, girlfriend, overhead, paperback, darkroom, smalltalk.*

2. Derivation is a process of adding suffixes and prefixes to existing words, for example: *culture - cultural, person - personal, economy - economic - economical, funding - underfunding, familiar - unfamiliar, magnet - non-magnetic.*

3. Back-formation is a process of removing affixes. This is one of the illustrations referring to the shortening. Some of the examples are: *editor - edit, sculptor - sculpt, aviation - aviate, biography - biograph, evaluation - evaluate, laser - lase, semantics - semantic (adjective).*

4. Clipping refers to the process of extracting part of a longer word and using it with the identical meaning: *exam - examination, fax - facsimile, gas - gasoline, memo - memorandum, pop - popular music, flu - influenza.*

5. Acronyms are one of the most typical examples of frequency-based approach in language change. The more frequently a complex term is used by the linguistic community, the shorter it tends to be. In the OPAALS community, we have a variety of them: *DE, OKS, P2P, SBVR*, etc.

6. Blending is a process of creating a new word from portions of the existing words, for instance: *breakfast + lunch = brunch; camera + recorder = camcorder.*

These processes are illustrative examples of the proposed *frequency →simplification* law in language evolution.

Expanding the meanings of words is the second powerful technique of semantic change. For example, due to the technological progress, the words *chip*, *disk*, *program*, *memory*, *mouse* got new meanings. Because of the statistical properties of vocabulary use, old meanings tend to become less frequent or even disappear in the lexicon. Change in meaning happens often through the following mechanisms:

1. Metaphor: the word *mouse* as a "rodent" got a new meaning, a"computer device"[10].

[10]Metaphor is a powerful technique that gives new meanings to words (see Deliverables 6.3 and 6.7 for more details on metaphors).

2. Metonymy: *The White House tried to avoid the scandal*, which refers to the representatives of the White House.

3. Specialisation of meaning, i.e. the meaning of a word becomes less general. In terms of taxonomic representation, the meaning of a words "shifts down" in a hierarchy. For example, the word *meat* originally meant "food, solid food". However, it is currently restricted to the "flesh of animals".

4. Generalisation of meaning, i.e. the meaning of a word becomes more general. In contrast to the process of specialisation, here a word moves upwards within the hierarchical representation of relations. The word *holiday*, for example, meant "a day for religious significance". Nowadays, it means "day of festivity or recreation".

One of the reasons for rapid semantic change in our life (as well as in DE) is the fact that words are usually polysemic (having several meanings).

The selection of new words as well as of new meanings due to the described mechanisms is again based on the statistical properties of use within a language environment. However, knowledge of these mechanisms might be relevant for the design of a dynamic knowledge system within digital environments. The semantic change and evolution in formal knowledge spaces is the focus of chapter 7. One important aspect in the context of tracing the semantic meaning is the role of syntax. The structure of the sentence reveals its meaning and the meaning of the elements.

6.3 Language Networks

Looking at language through the lens of a network can help to analyse and explain the evolutionary process through characterizing its statistical properties and development. Sole et al. (2005) identify different types of language networks and their dynamics: co-occurrence, syntax and semantic networks. In the following section, we present and discuss the role and impact of these networks in Digital Ecosystem through the lens of the evolutionary processes.

6.3.1 Co-occurrence networks

This is one of the simplest networks (from the linguistic point of view). It is based on the co-occurrence of words in a sentence (Figure 3). This means that when two words appear in at least one sentence (or in a pre-defined window), they are connected within the network. The degree of connections

is directly related to the frequency of their appearance in a text. The co-occurrence networks are able to partially represent the syntactic and semantic relations between words (= nodes). However, these relations do not underlie the structure of the network, but rather they occur by chance. The analysis of these networks in human populations leads to an interesting observation revealing the dependency between lexicon size and flexibility: "Children with smaller lexicon display a network with higher connectivity" (Sole et al., 2005, p. 4).

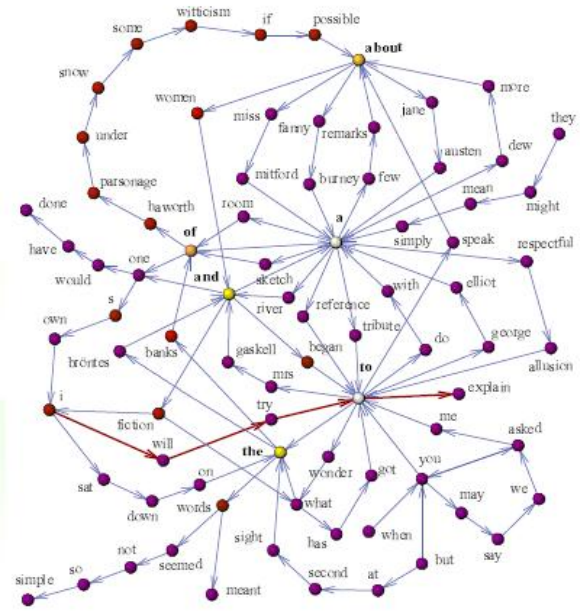


Figure 3: *Co-occurrence network (taken from Sole et al. (2005))*

6.3.2 Syntax networks

A language network is highly dependent on the syntax. Such networks (Figure 4) can be built up on either constituent or dependency approaches[11]. Analysing language as a network with regard to the dependency formalism is advantageous because "it retains the words themselves as fundamental nodes in a graph" (Sole et al., 2005). Consider, for example, the phrase *gas smells*. The dependency between these two items is implicit in the constituency network (Figure 5(a)), but explicit in the dependency approach (Figure 5(b)).

[11]A detailed description of syntactic analysis and its role in digital environments is provided in deliverables D6.4 and D6.7

Several studies have analysed language networks in order to explore their properties (see for example Brede & Newth, 2004, H. Liu & Hu, 2008).

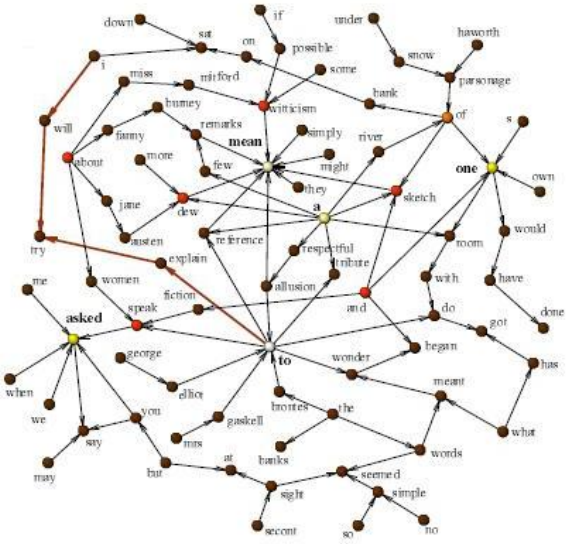
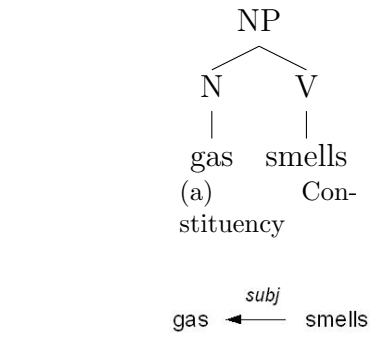


Figure 4: *Syntactic network (taken from Sole et al. (2005))*



(a) Constituency

(b) Dependency

Figure 5: *Syntax and language networks*

6.3.3 Semantic networks

Based on the semantic relations between concepts, this type of network is currently one of the most popular forms of knowledge representation (Figure 6). In the context of language evolution, the typology of such networks has gained a lot of interest during the last decades. Several studies show that

this type of network has "a highly efficient organization" (Sole et al., 2005, p. 3). Steyvers and Tenenbaum (2005) have analysed the structure of three semantic networks in statistical terms: word associations, WordNet, and Rogets Thesaurus. The results reveal that different networks have similar features in terms of sparsity (high), path length (very short in average), clustering (strong local), and power-low distribution. Ravasz and Barabási (2003) studied the language network that is generated on the basis of the synonymic relations between words in Merriam Webster dictionary. The resulting network contains 182,853 nodes and 317,658 links and it is scale-free with degree exponent a = 3.25. Further analysis of the network topology has led authors to the conclusion that it has a hierarchical structure.

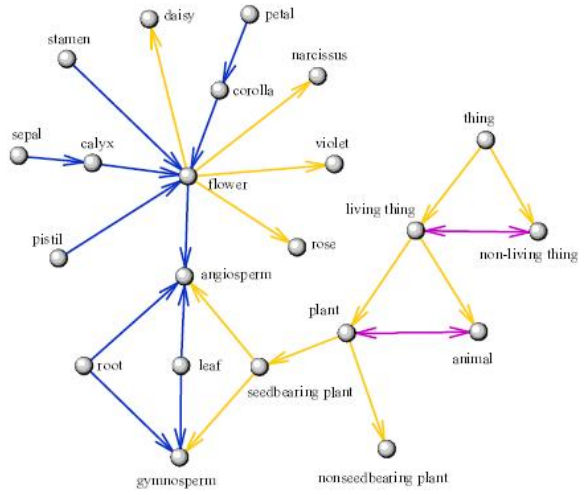


Figure 6: *Semantic network (taken from Sole et al. (2005))*

6.3.4 Language networks and their properties

In the context of our own research, we argue that the statistical properties of language networks (co-occurrence, syntactic, semantic, or others) provide valuable information about the processes of language development and evolution at different levels of organisation (e.g. individual and group levels).

In general, the language as a complex dynamical system can be analysed through the lens of network topology. As Sole et al. (2005) have pointed out:

> It exhibits highly intricate network structures at all levels (phonetic, lexical, syntactic, semantic) and this structure is to some extend shaped and reshaped by millions of language users over

> long periods of time, as they adapt and change them to their needs as part of ongoing local interactions (p.3).

These types of language networks provide a useful landscape for studying evolution with regard to the language elements at different linguistic levels (e.g. syntax, semantics). Moreover, the language network approach can be used for analysing language transmission and change at the intersection between individuals and/or communities, e.g. the emergence of language through individual interactions. Language networks can be analysed according to two basic features (J. Liu, Wang, & Wang, 2008):

- "small world" structure
  Human language as a complex, dynamic and adaptive system reveals the characteristics of the "small world". Cancho and Solé (2001) discovered that the "average distance between two words, d (i.e. the average minimum number of links to be crossed from an arbitrary word to another), is shown to be $d \approx 2\text{-}3$" (p. 2261).

- "scale-free" topology
  A variety of recent investigations of language networks provides evidence of the scale-free topology (Motter, Moura, Lai, & Dasgupta, 2002; H. Liu & Hu, 2008; Barabási, Dezso, Ravasz, Yook, & Oltvai, 2003; Ferreira, Corso, Piuvezam, & Alves, 2006).

The investigations of language dynamics in digital environments, based on the analysis of the network structure and typology, can help with the design of complex, adaptive, and self-organised systems. The statistical behaviour of language networks provides a valuable input to the questions of the evolution, i.e. how the system emerges, changes and varies.

6.4 Evolution of Programming and Human-Computer Interaction

The evolution of programming languages in the last decades was mainly driven by the lack of computing power on the one hand and by increasing complexity on the other hand. As computers became more powerful, software grew evermore complex and soon different programming paradigms became unsatisfactory. With the advent of the Internet, programming languages supporting the development of distributed software propagated further. In this process the concept of programming languages was migrated to the concept of development platforms, as more and more tools and libraries grew together

to support the developers. With the upcoming change to the *Cloud Computing* age these platforms will change to support programmers in building even more powerful and complex software constructs.

The following sections provide examples of how language evolution, as described in this deliverable, has already happened in the evolution of programming languages. This is followed by a section on evolution of Human-Computer Interaction (HCI) and finally brings together concepts by referring to the Code Structure Generation Prototype from WP 2.

6.4.1 Frequency-Based Contractions in Programming Language

In section 5.5 the concept of Frequency-Based Contractions that can be found very often in the evolutionary history of programming languages is explained. In programming languages this is a very important factor, as these developers do not need to write any surplus code. Therefore, languages are constantly changed to support simpler code constructs for frequently used functionalities. One of the prominent examples is the increment or decrement operators which are available in different varieties. In the following an example of the post-increment operator in C++ is given:

```
int i = 0;
i = i + 1;
```

which changed to

```
int i = 0;
i ++;
```

There are many more examples of this kind and programming languages steadily evolve in order to provide developers with the possibility to write the code faster. The easiest approach is to simplify the most common structures. Another example for evolution in software development is situated on a higher level and is a part of Object Oriented Programming: the concept of inheritance.

6.4.2 Object Orientation as basis for Mutation, Adaption and Selection

According to the TIOBE Index (TIOBE Software), 55,3 % of used programming languages are object-oriented as of May 2009. Such languages themselves provide a remarkable flexibility to model natural evolution due to the design principles: Encapsulation, Abstraction, Inheritance and Polymorphism. These principles enable programmers to define different child-classes

with different behaviours with defined interfaces. This makes it possible to change the used implementation at *compile-time* and during *run-time*. If algorithms prove to be flawed, they can easily be replaced without changing the original code, but by adding new code to evolve a new different *species*. If both *species* have both advantages and disadvantages, chances are high that both implementations will coexist - e.g. one may choose which algorithm to use. If one algorithm appears to be superior to another, the unused implementation will probably disappear at some point in time (Isernhagen & Helmke, 2004; Gamma, Helm, Johnson, & Vlissides, 1995).

On an even higher level, Plugin-Based software snaps in. As software products became more and more complex, it has become more important to be able to separate them into components. If these components have clearly defined interfaces and can be installed separately by the user due to the user's specific demands, then we are talking about Plugin-Architectures. When talking about Object Oriented Languages and Java, one of the most widespread Plugin-Platforms is Eclipse (The Eclipse Foundation).

Although when talking about Eclipse the Integrated Development Environment (IDE) package is meant. The Eclipse IDE really is a plugin platform with a huge number of different plugins for all kinds of purposes. Users can install needed plugins during runtime which lets the IDE *evolve* according to the users needs (Budinsky, Steinberg, Merks, Ellersick, & Grose, 2003; Daum, 2005).

6.4.3 Survival of the fittest Programming Languages

More mature programming languages do not change as fast as new programming languages because they need to keep their support for legacy code. This means that the bigger the adoption of languages is, the more stable they have to be and, consequently, the slower they can evolve. This gives every now and then the opportunity to other languages to win over developers using established languages.

Such an example is *D* which has tried to gain acceptance since 2007. D claims to be the successor to C and C++. The creators of D argue that the standards of C and C++ have become too large and complicated, which has led to the need for a simpler successor with the same capabilities. A detailed feature comparison between C++ and D can be found in the Overview page available at Digital Mars. After a fair amount of hype, D seems to be loosing the battle according to the TIOBE Index (TIOBE Software). The next months will show whether D still can survive among a large amount of very mature and established languages (Digital Mars; ISO).

Another example is Java which was introduced in June 1991 for an en-

tirely different purpose: to develop programs on television set-top boxes. The first stable release was published in 1995 and according to the TIOBE Index (TIOBE Software) it is now one of the leading programming platforms. Since the launch of Java the unique selling proposition of *write once, run everywhere* boosted Java to be now the leading platform for enterprise software (Gosling, Joy, Steele, & Bracha, 2000). This can be interpreted in the sense that unique features when combined with natural selection may drive forward certain concepts to areas where they were not planned to be.

6.4.4 Human-Computer Interaction (HCI)

More examples for evolution can be found in the research area of Human-Computer Interaction (HCI). Among other interests HCI tries to unite powerful and simple traits as regards user interfaces. Important and often used functionalities, so called key-features, need to be accessible within very few interaction steps. This also leads to an important aspect in web design, where every page within a web site should be easily accessible without navigating through a complex structure (Shneiderman, 2002; Hoekman, 2007).

As the functionality of applications increased, the need for *faster ways to do things* emerged. Consequently *Keyboard Shortcuts* are used in order to help users access commonly used functionality without navigating through the growing menu structures of applications. Nowadays there is hardly any used software that does not support Shortcuts of some kind.

Another example in HCI where the concept of evolution is built as a functionality into software is *smart menus* introduced in Microsoft Office, where icons that are not used for a longer time vanish, until they are explicitly used. This concept has been developed even further to support a dynamic reordering of the icons by the usage quantity. The latter approach has the disadvantage that a rearranging of the icons can collide with user habits and therefore lead to users not accepting this feature. Users not accepting a feature can consequently lead to switching off the functionality, which at a broader scale can be seen as a species dying out by natural selection (Coyne & Orr, 2004).

For that reason it seems appropriate to follow the evolution of HCI on the one hand and try to integrate the concept of evolution on the other hand into systems supporting Digital Ecosystems. The latter approach could be achieved by the integration of Generative Programming.

As part of Work Package 2 as described in D 2.1 and D 2.2, a prototype for code structure generation was presented. In one sentence, this prototype showed that it was feasible to generate code structure out of Structured Natural Language which then could be used by the Grails Platform to provide

a User Interface. With an appropriate feedback loop, the input data could be adapted during runtime to change or evolve according to the users' need, for example, rearranging functionality in order to provide faster workflows within an application.

6.5 Natural language and knowledge representation

In Digital Ecosystems, the aspect of knowledge representation involving natural language becomes crucial. Generally, our research on evolutionary structure, factors, processes, and mechanisms involves two major areas:

- evolution of human language systems
  This aspect was at the focus of the previous chapters.

- evolution of formal systems
  The connections between natural and formal systems through the lens of evolution is a leitmotif of this deliverable. The frequency-based approach of language evolution has also been found in formal systems (see section 6.4 for more details). However, the detailed critical focus on the formal knowledge systems and evolutionary processes within it is still missing. This aspect is the subject of the next section.

7 Semantic Change and Evolution in Formal Knowledge Spaces

All information systems undergo some changes during their lifetime. Similarly, all information systems are vulnerable to changes which were not anticipated in their design. Since this observation equally applies to Client-Server, Peer-to-Peer (P2P), etc. information systems alike, taking change explicitly into account in system design is needed. The abstract working hypothesis in many large-scale information or knowledge systems, including the OPAALS OKS, is that the natural goal-oriented usage yields into useful, gradual development in the system. Thus, the development of information systems does not take place in a single specify-design-implement step. Rather, once an adequate bootstrapping or enabling system is specified and designed, it is developed - and develops in itself - according to the actual use. This incremental development takes place in a process that can be described using the terms evolutionary, self-organizing, or self-sustaining (i.e. autopoietic, see Dini (2008)). While the driving force of physical systems might be captured in terms of the laws of physics (physical phenomena), or energy, the driving force of a knowledge system might be captured in terms of the laws of social systems, socio-economics, etc. (behaviour of people), or information.

Next, we shall characterise and analyse change and evolution in knowledge spaces and information and information evolution management through the lens of the Semantic Web (see the W3C Semantic Web Activity, `http://www.w3.org/2001/sw/`). This simplification is rational since at the end of the day, we expect to technically manage evolutionary information with respect to Semantic Web technologies in applications related knowledge transformation and information visualisation (see Nykänen (2007b); Nykänen, Salonen, Haapaniemi, and Huhtamäki (2008)).

This section (7) mainly exploits the ideas published in Nykänen (2009). A formalised (and simplified) notion of context-aware knowledge in terms of fuzzy logic programs is discussed elsewhere (Nykänen, 2007a), complemented with a precise definition of interpretation logics (Nykänen, 2007b). On the specification level of encoding semantic data, however, this discussion is implicitly present in the definition of the OKS Data Model (Nykänen, Salonen, Haapaniemi, & Huhtamäki, 2008).

7.1 Knowledge Management through Information and Data Sharing Systems

In order to consider evolution in the context of information management, it is helpful to review the abstract processes of knowledge management and data sharing. In a knowledge creating organisation, the following key roles and processes can be recognized in knowledge asset management (Nonaka & Takeuchi, 1995):

- Externalization (transforming knowledge from tacit to explicit, e.g. written reflective communication in small groups)
- Combination (transforming knowledge from explicit to explicit, e.g. integrating group's outputs on the level of the organisation)
- Internalization (transforming knowledge from explicit to tacit, e.g. exploring and studying the organisation memory)
- Socialization (transforming knowledge from tacit to tacit, e.g. existential "face-to-face" communication)

The practical but only partly technical challenges of knowledge management must also be acknowledged. When a knowledge management activity fails, it is not necessarily due to theoretical problems in abstract knowledge management. Rather, it may be due to not understanding or capturing the existing processes (e.g. human motivation and policy of participation), and shortcomings in everyday information systems (e.g. lack of explicit knowledge, management problems).

Also, sharing is an issue. When interests and values differ, an overly ideal model of collaboration might not work. Considering the practical challenges of information management, we can analyse (simplify) the setting further in terms of Data-Sharing Systems (DSS). According to Smith, Seligman, and Swarup (2008), the fundamental activities of DSS include:

- Discovery (e.g. learning what's in there)
- Access (e.g. being able to retrieve the interesting bits)
- Understanding (e.g. being able to translate/contextualize the retrieved data into something locally applicable)
- Policy-making (e.g. having clear rules for terms of use and giving and receiving credit)

While proper ICT tools are obviously a necessity, technologies constitute only a part of the big picture. In particular, while advocates of "semantic applications" typically highlight e.g., semantic search (and the related more sophisticated applications such as recommendation systems), semantic search alone does not provide an "evolutionary knowledge space". Analysing current data-sharing systems suggests that social factors and organisational policy-making are in a decisive role in successful applications.

Thus, because of its information sharing foundation, knowledge management, and hence evolutionary knowledge spaces, must identify and actively support the key activities of both knowledge management and data sharing systems. As usual, great challenges lie in the complexity of the socio-technical system.

7.2 Evolution in Information Systems

To clarify the discussion, it is helpful to focus on the concept of change. Simplifying the setting a bit, two main areas of change can be recognized: Changes in data, and changes in metadata. Note that this (idealistic) formulation suggests that information needs will dominate tool development in information systems. Another perhaps equally justified stance would be to say that the development of (legacy) tools is dominant in information systems.

Let us next consider changes with respect to the following six information representation layers in metadata (or semantic descriptions):

1. Conceptualization
2. Syntax
3. Knowledge Representation (KR) Language
4. Domain-Specific (Semantic) Schema(s)
5. Content Data
6. Interpretation

Intuitively, conceptualization denotes the modelling frame of reference and drives the whole process of capturing semantics (e.g. fuzzy vs. crisp entities, open vs. closed world assumption, abstract use cases, etc.). Syntax is required to capture descriptions (e.g. logical statements). KR language is needed for introducing domain-specific structures in terms of common fundamental concepts and well-established consequences (e.g. classes and

properties, logical entailment etc.). Domain-specific schemas are needed for carrying application-specific semantics (e.g. accepted structures and vocabularies such as the Dublin Core or RSS). Content data is obviously required as well, recording the asserted semantic descriptions with respect to the selected schemas (e.g. a particular date when a particular document was published, using a well-defined vocabulary). Finally, interpretation means how the available semantic descriptions and schemas are used in applications (e.g. using a publication date in a calendar application, perhaps not following the formal consequences of the underlying layers to the letter).

The relationship with Semantic Web technologies is obvious. It is also obvious that in principle changes may take place in any of the above six layers. Further, versioning might take place in all of these layers as well, both in data and in metadata. However, in the broadest sense, versioning requires a complete information system of its own: Simply introducing a particular "versioning vocabulary" requires fixing layers 1-3.

Obviously, typical information systems do not support arbitrary changes well. The reasons are partly related to the work needed for implementing suitable tools, partly related to the fact that all changes are not systematically archived with complete versioning information. Practical applications focus on versioning content, if any, with different versioning strategies. Versioning of schemata, KR languages, and syntax is much more difficult and typically takes seriously place only in substantial applications. In many cases, reasons are highly practical: The willingness to change layer X is inversely comparable to the complexity, number, and quality of the existing tools and data in that layer.

While acknowledging the complexity of potential changes in the worst-case scenarios (in particular, changes in the conceptualization and the KR language) it seems that in a reasonable setting, evolution mainly takes place in layers 3-6. Further, KR languages and domain-specific vocabularies which meet the technical and socio-economic requirements (e.g. licensing policy) of the P2P community can be considered to evolve very slowly. Note that these observations seem to speak strongly in favour of the W3C Semantic Web standards.

As a consequence, restricting evolution in levels 4-6 provides the opportunity to introduce common vocabularies. Of course, this is how applications indeed have evolved. In a P2P setting, common vocabularies may be used to implement efficient keyword-based systems and schema-based P2P systems, at least describing P2P resources from the perspective of a particular semantic search-use case. Further, dealing with evolution is seemingly easier when common versioning semantics can be used. In principle, this also allows the capturing (and perhaps explaining) of both the evolution of domain-specific

vocabularies and the evolution of content within a single common versioning system. However, common policies are needed for evolution management (in particular interpretation and tool support).

Finally, the observation of the interpretation layer is rather interesting. In particular, it turns out that this layer can be formalized to a certain extent, e.g. in terms of interpretation logics (Nykänen, 2007b). This provides a concrete - albeit abstract - strategy for implementing a unified search interface in a contributed knowledge system.

7.3 Knowledge

So far we have been mostly talking about information — what about knowledge?

According to the above characterisation, knowledge appears in implicit and explicit forms, in the heads of people, in their environment, in paper documents (etc.), and in computerized systems. Knowledge management requires both tools for recording, describing, and transforming explicit knowledge, but also visionary policies and places of fruitful internalization and socialization. Moreover, the requirements of data sharing systems must also be acknowledged[12].

Due to its information roots, our earlier discussion about the evolution of information directly applies to intermediated knowledge. In fact, an epistemological purist can always claim that externalized knowledge only exists as information (perhaps thus nullifying the very concept of formal externalized knowledge). However, in order to constructively characterize the ascent from information to knowledge, let us briefly consider a friendly application that exists on the blurred borderline.

A simple non-trivial approximation of explicit knowledge structures is provided by concept schemes (e.g. folksonomies or thesauri). Intuitively, the level and applicability of information is increased by weaving some of the individual fragments of consistent conceptual descriptions into common structure(s). Integrating the needs of discovery and understanding, concept schemes also establish an extensible, practical structure for performing subject-based semantic searches.

Unfortunately – to the relief of epistemological purists – using concept schemes in a P2P system again faces the same challenges and sources of single-point of failure as information evolution does. In particular, while a common schema for representing concept schemes might be easily agreed

[12]See also deliverables D6.1 and D10.5.

upon technically (e.g. SKOS), organising concepts basically leads to information management without centralized design.

In fact, this whole discussion indicates that in a decentralized system, inconsistencies and overlaps simply need to be tolerated.

7.4 Towards Contributed Knowledge Spaces: Semantic Search Revisited

An attractive strategy for addressing the concerns of mismatching views, evolutionary changes, and the lack of the global big picture, is to consider knowledge spaces in terms of contributed information systems with a long-term version memory (Nykänen, Salonen, Huhtamäki, & Haapaniemi, 2008). The basic idea is that a distributed evolutionary knowledge space is not a well-defined, strictly integrated information system with perfect semantics and referential integrity. Rather, it appears as a collection of contributing information systems that do not necessarily agree about the schemas, content data, or interpretations (or even the other underlying information layers).

As a consequence, the role of semantic search is again highlighted. In a very abstract technical sense, the semantic search interface more or less represents the whole knowledge space, perhaps itself denoting the unifying artefact: if a reference to a thing cannot be found via the search interface, that thing does not exist in a useful sense in the knowledge space. In a distributed system, the P2P network is ideally responsible for asserting this interface. Further, in order to manage trends at system level, past changes must also be remembered. In general, this may require archiving and versioning activities on several layers of information. Considering Semantic Web, this outlines the importance of proper namespace and semantic vocabulary management.

As an alternative, a complementary strategy is to recognise the assertion context and the interpretation context of semantic descriptions. In principle, this enables aspect-oriented perspectives to the contributed knowledge space. Whether this is feasible in practice is probably related to the degree of automation in managing and encoding contextual information. Further, dealing with assertion and interpretation contexts might be considered as the responsibility of the search interface, or be simply left to the user (thus recognising the contexts simply as "extra" search and knowledge translation criteria).

7.5 Taking Evolution into Account in Externalization and Combination

Accepting the idea of a contributed, evolutionary knowledge space, the processes of asserting and capturing must be designed with care. In particular, strictly global statements on the layer (5) of Content Data should be avoided. Rather, semantic statements should be technically captured as indirect statements where things are referenced with sufficiently rich characterizing properties rather than global identities. Also, for evolution-aware interpretation, asserted statements should be associated with sufficient contextual information. Of course, if these requirements are not met, evolution cannot be formally addressed. (While similar observations apply to other information representation layers as well, Content Data and Semantic Schemas provide the most important use cases in the evolution of formal languages.)

Finally, in order to optimize the expensive process of writing semantic descriptions, it is helpful to identify three main sources of data:

1. Human-authored descriptions (e.g. decisions and classification made by end-users or experts)

2. Machine-authored contextual descriptions (e.g. the date, the authoring tool, the MIME type, and the author of a document, automatically collected by the editor software)

3. Legacy description repositories (e.g. general-purpose adapted archives and other interoperable applications)

As suspected, the careful design of formal, evolution-aware knowledge spaces should take item 1 into account in the level of documented best practices, and items 2 and 3 in the level of technical system design.

8 Conclusion

Natural language as a foundation of knowledge representation and transmission changes over time. Hence, the mechanisms of change need first to be understood. The frequency-based approach presented and discussed in this deliverable helps to explain the driving forces of natural language change and variations. Since this is true not only for humans, but also for digital environments, we see a strong connection of the linguistic view of evolution to "formal" systems. One of the important considerations to be taken into account is that a system will always change. To make this process dynamic and natural (e.g. self-organised), the statistical properties of usage are most appropriate since they can be found in all the systems investigated here. The basic law *frequent use →simplification* has different kinds of appearance that mirror the power and economy of any system (language, network, digital environment).

Furthermore, the understanding of underlying laws and forces of language evolution helps with creating systems that are both adaptable to user needs and are self-organised.

This piece of collaborative research provides interdisciplinary models of language evolution that are relevant in different domains, such as social science, natural science and computer science.

The evolutionary framework plays various roles in this setting:

- Practical consequences (WP2, WP3, WP5, WP10)
- Theoretical consequences (WP6, WP10, WP11, WP12)

Practical consequences identify and hint, e.g., additional requirements for technology adoption and tool development (e.g. versioning), also suggesting best practices for authoring and collaboration processes (e.g. adding/inducing contextual descriptions). More generally, the research of the language evolution has a strong impact to the knowledge systems in DEs. Evolution of natural language implies changes in the conceptualization of a specific domain (or several domains within P2P environments). Consequently, this has effect on any knowledge system, its management and performance.

> The analyses of natural languages and design of interactive computer systems reveal many of the same pressures. In both communication media, these pressures lead to innovations in the structure of the medium, inconsistencies, and a continual tension between expressiveness, ease of use, ease of understanding, and ease of learning (Grudin & Norman, 1991).

As presented and discussed in this deliverable, the frequency usage underlies such pressures. Therefore, consideration of this aspect needs to be taken into account during the design and development of a dynamic system (OKS, knowledge systems, networks, and other), in general, and specific algorithms that build up such a system, in particular.

Theoretical implications on the other hand suggest considering language and representations as inherently dynamic properties of socio-technical systems. This has several more subtle consequences, related not only to linguistics (e.g. self-organisation in language), but also knowledge management (e.g. knowledge transformation, implicit versus explicit), and epistemology (e.g. knowledge versus representation, utility). Further analysis of the changes of linguistic patterns and structures could provide relevant insights and recommendations for the design of online communities, its communicative structures (including governance, knowledge production and management, etc.), and how to facilitate and support dissemination (e.g. successful integration of new members, sustainability aspects, and other). This work opens a future perspectives of research on distributed, self-organised and adaptable systems leading to integration and collaboration of different disciplines.

References

Aissen, J. (2003). Differential object marking: iconicity vs. economy. *Natural Language and Linguistic Theory*, *21*(3), 435–83.

Barabási, A., Dezso, Z., Ravasz, E., Yook, S., & Oltvai, Z. (2003). Scale-free and hierarchical structures in complex networks. *AIP Conference Proceedings*, *661*.

Berry, K., & Berry, C. (1985). *A description of Abun: A West Papuan language of Iranian Jaya*. Canberra: Australian National University.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences*, *7*(2), 173–222.

Brede, M., & Newth, D. J. (2004). Patterns in syntactic dependency networks from authored and randomised texts. In *Complex 2004: Proceedings of the 7th Asia-Pacific Conference on Complex Systems.*

Budinsky, F., Steinberg, D., Merks, E., Ellersick, R., & Grose, T. (2003). *Eclipse Modeling Framework*. Addison Wesley Professional.

Bybee, J. (1988). Morphology as lexical organization. In M. Hammond & M. Noonan (Eds.), *Theoretical morphology* (pp. 119–141). Academic.

Cancho, F. I. R., & Solé, 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.

Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Readings in english transformational grammar* (pp. 184–221). Waltham: Ginn.

Chomsky, N. (1972). *Language and mind*. New York: Harcourt Brace Jovanovich.

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sunderland, Massachusetts: Sinauer Associates.

Croft, W. (2000). *Explaining language change: an evolutionary appoach*. Harlow: Longman.

Croft, W. (2003). On iconicity of distance. *Cognitive Linguistics*, *19*(1).

Croft, W. (2008). On iconicity of distance. *Cognitive Linguistics*, *19*(1), 49–58.

Daum, B. (2005). *Rich-Client-Entwicklung mit Eclipse 3.1*. Heidelberg: dpunkt.verlag.

Dayley, J. P. (1985). *Tzutujil grammar*. Berkeley: University of California Press.

Digital Mars. (2009). *D programming language*. (`http://digitalmars.com/d/`, Last accessed on 6. May 2009)

Dini, P. (2008, October 7-8). Notes on relational biology and elementary category theory. In *Proceedings of OPAALS 2008.*

Dixon, R. M. W. (1991). A typology of causatives: Form, syntax, and meaning. Cambridge University Press: Cambridge.

Ferreira, A. A. A., Corso, G., Piuvezam, G., & Alves, M. S. C. F. (2006). A scale-free network of evoked words. *Brazilian Journal of Physics*, *36*(3), 755 - 758.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns*. Indianapolis: Addison Wesley.

Gendarmi, D., Abbattista, F., & Lanubile, F. (2007, May 8-12). Fostering knowledge evolution through community-based participation. In *16th International World Wide Web Conference*.

Givon, T. (2008). Isomorhism in the grammatical code: cognitive and biological considerations. *Studies in Language*, *15*(1), 85–114.

Gosling, J., Joy, B., Steele, G., & Bracha, G. (2000). *The Java Language Specification 2nd Ed.* Boston: Addison-Wesley.

Greenberg, J. (1966). *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.

Grudin, J., & Norman, D. (1991). Language evolution and human-computer interaction. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society* (pp. 611–616).

Haiman, J. (1980). The iconicity of grammar. *Language*, *56*, 515–40.

Haiman, J. (1983). Iconic and economic motivation. *Language*, *59*, 781–819.

Haiman, J. (1994). Iconicity. In R. E. Asher (Ed.), *The Encyclopedia of Language and Linguistics* (pp. 1629–33). Oxford: Pergamon PRess.

Haiman, J. (2000). Iconicity. In G. Booij, J. Mugdan, & C. Lehmann (Eds.), *Morphology: An international handbook* (pp. 281–288). Berlin: de Gruyter.

Haspelmath, M. (1999). Optimality and diachronic adaption. *Zeitschrift für Sprachwissenschaft*, *18*(2), 180–205.

Haspelmath, M. (2008a). Creating economical morphosyntactic patterns in language change. In J. Good (Ed.), *Language universals and language change* (pp. 185–214). Oxford: Oxford University Press.

Haspelmath, M. (2008b). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, *19*(1), 1–33.

Haspelmath, M. (2008c). Frequency vs, iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, *19*(1), 1–33.

Haspelmath, M. (2008d). Reply to Croft and Haiman. *Cognitive Linguistics*, *19*(1).

Haspelmath, M. (March 2008). *Why is language typology possible?* (Vol. Plenary Evening Lecture at Leipzig Spring School on Linguistic Diversity).

Hoekman, R. J. (2007). *Designing the Obvious: A Common Sense Approach to Web Application Design*. New Riders Press.

Isernhagen, R., & Helmke, H. (2004). *Softwaretechnik in C und C++* (4th ed.). Munich: Hanser.

ISO. (n.d.a). *ISO/IEC 9899 - Programming languages - C.* (`http://www.open-std.org/jtc1/sc22/wg14/www/standards.html`, Last access on 6. May 2009)

ISO. (n.d.b). *JTC1/SC22/WG21 - The C++ Standards Committee.* (`http://www.open-std.org/jtc1/sc22/wg21/`, Last access on 6. May 2009)

Jackendoff, R. (1977a). *X-syntax: a study of phrase structure*. Cambridge: MIT Press.

Jackendoff, R. (1977b). *X-syntax: a study of phrase structure*. Cambridge: MIT Press.

Jakobson, R. (1966). Implications of language universals for linguistics. In J. H. Greenberg (Ed.), *Universals of language* (pp. 263–278). Cambridge: MIT Press.

King, G. (1985). *Modern welsh*. London: Routledge.

Langacker, R. (2000). The meaning of of. In R. Langacker (Ed.), *Grammar and conceptualization* (pp. 73–90). Berlin: Mouton.

Lehmann, C. (1974). Isomorphism im sprachlichen Zeichen. In H.-J. Seiler (Ed.), *Linguistic Workshop II: Arbeiten des Kölner Universalienprojekts 1973/4* (pp. 98–123). Mnchen: Fink.

Liu, H., & Hu, F. (2008). What role does syntax play in a language network? *Europhysics Letters*, *83*.

Liu, J., Wang, J., & Wang, C. (2008). Research on text network representation. In *IEEE International Conference on Networking, Sensing and Control, ICNSC 2008* (p. 1217-1221).

Marcus, G. F. (2005). *Der Ursprung des Geistes*. Düsseldorf: Patmos Verlag.

Mayerthaler, W. (1988). *Morphologische Natrlichkeit*. Wiesbaden: Athenaion.

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*(6).

Mufwene, S. (2002). Competition and selection in language evolution. *Selection*, *3*, 46–56.

Newmeyer, F. (1992). Iconicity and generative grammar. *Language*, *68*, 756–96.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How japanese companies create the dynamics of innovation*. USA: Oxford University Press.

Nykänen, O. (2007a, October 5-8). Implementing context-specific views to distributed (rule) databases with fuzzy logic. In *Proceedings of the International IADIS, WWW/Internet 2007 Conference* (pp. 305 – 312). Vila Real, Portugal.

Nykänen, O. (2007b, November 26-27). Interpretation logics. In *Proceedings of the 1st OPAALS conference.* Rome, Italy.

Nykänen, O. (2009). Semantic web for evolutionary peer-to-peer knowledge space. In K. Birkenbihl, E. Quesada-Ruiz, & P. Priesca-Balbin (Eds.), *Monograph: Universal, ubiquitous and intelligent web* (Vol. X). CEPIS & Novtica.

Nykänen, O., Salonen, J., Haapaniemi, M., & Huhtamäki, J. (2008, October 7-8). A visualisation system for a peer-to-peer information space. In *Proceedings of OPAALS 2008.*

Nykänen, O., Salonen, J., Huhtamäki, J., & Haapaniemi, M. (2008). *OKS Data Model, version 1.01. A milestone specification for the OPAALS project (Contract number IST-034824)* (Tech. Rep.). WP10: Sustainable Research Community Building in the Open Knowledge Space. (Contribution to the Milestone M10.10: OKS Data Model (M24), 1 July 2008 (29 pages). A local copy is available at `http://matriisi.ee.tut.fi/hypermedia/julkaisut/20080701-oks-dm-v1-01.pdf`)

Oller, D. (2004). Underpinnings for a theory of communicative evolution. In D. Oller & U. Griebel (Eds.), *Evolution of communication systems* (pp. 49–65). Cambridge: MIT Press.

Pinker, S. (1994). *The language instinct: How the mind creates language.* NY: William Morrow and Company.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and brain sciences*, *13*, 707–789.

Plank, F. (1979). Ikonisierung und De-Ikonisierung als Prinzipien des Sprachwandels. *Sprachwissenschaft*, *4*, 121–158.

Ravasz, E., & Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E*, *67*(2).

Sereno, M. (1991). Four analogies between biological and cultural/linguistic evolution. *Journal of Theoretical Biology*, *151*, 467-507.

Shneiderman, B. (2002). *User interface design* (German ed.; mitp, Ed.). Bonn: mitp.

Smith, K., Seligman, L., & Swarup, V. (2008, September). Everybody share: The challenge of data-sharing systems. *IEEE Computer Magazine*.

Sole, R. V., Corominas, B., Valverde, S., & Steels, L. (2005). Language networks: their structure, function and evolution. *Trends in Cognitive Sciences*.

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of*

Communication, *1*(1), 1–34.

Steels, L. (2004). Analogies between genome and language evolution. In J. B. Pollack, M. Bedau, P. Husbands, T. Ikegami, & R. A. Watson (Eds.), *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems* (pp. 200–207). Cambridge, MA: The MIT Press.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78.

Stowell, T. (1981). *Origins of phrase structure.* Cambridge: MIT-Diss.

The Eclipse Foundation. (n.d). *Eclipse.* (`http://www.eclipse.org/`, Last accessed on 8. May 2009)

TIOBE Software. (n.d.). *Tiobe programming community index.* (`http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html`, Last accessed on 6. May 2009)

Tullo, C., & Hurford, J. (2003). Modelling Zipfian Distributions in Language. In S. Kirby (Ed.), *Proceedings of Language Evolution and Computation Workshop/Course at ESSLLI* (pp. 62–75). Vienna.

Vogt, P. (2007). Variation, competition and selection in the self-organisation of compositionality. In B. Wallace, A. Ross, J. Davies, & T. Anderson (Eds.), *The mind, the body and the world: Psychology after cognitivism?* (pp. 233–256). Exeter, UK: Imprint Academic.

Worden, R. (1995). A speed limit for evolution. *Journal of Theoretical Biology*, *176*, 137–152.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, *40*(1), 1–95.