NATURAL LANGUAGE AND DIGITAL ENVIRONMENTS
Evolutionary Perspective

Oxana Lapteva
Computational Linguistics Research Group, Department of English and Romance Languages
University of Kassel, Kassel, Germany

Keywords: Evolution, Frequency-based approach, Natural language, Digital environments.

Abstract: 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 work considers a different view on the evolution (in comparison with biological and cultural explanations) and its underlying laws by looking at usage frequency. It aims to explore the mechanisms and driving forces of natural language evolution and link them to the research of Digital Ecosystems involving design and implementation of systems. 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.). This paper presents and discusses a simple but very powerful mechanism of evolution: frequency.

1 INTRODUCTION

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. Starting with the investigation of natural language change and variation, this work aims to find the core-mechanisms of evolution and self-organisation and apply them to the domain of Digital Ecosystems (Dini et al., 2005; Briscoe and Wilde, 2006).

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. The linguistic view provides interesting insights into the research of evolutionary processes and mechanisms occurring in digital environments, not only in respect to the formal languages and formal representations, but also to a system's design. The frequency-based approach is a necessary prerequisite for interfaces, knowledge platforms, distributed systems, and others. Hence, the integration of "natural"(e.g. human language) and "formal" (e.g. formal languages, interfaces, formal knowledge spaces) constituents existing within Digital Ecosystems is the leitmotif of this research.

2 NATURAL LANGUAGE EVOLUTION

2.1 Overview

Within the scope 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).

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 (Steels, 2004). The cultural view which assumes learning as the underlying mechanism of language evolution still has a problem of explaining the mechanisms of sharing. The biological approach, by contrast, faces the problem of explaining the rapid rise and spread of new language items (e.g. new concepts) in human languages. 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*, and others. Another issue the biological view may face is the question of storage. According to Worden (Worden, 1995), humans do not have enough storage in genetic

Lapteva O. (2009).
NATURAL LANGUAGE AND DIGITAL ENVIRONMENTS - Evolutionary Perspective.
In *Proceedings of the International Joint Conference on Computational Intelligence*, pages 314-317
DOI: 10.5220/0002335003140317

form for such a huge amount of data (i.e. all types of linguistic information) they need to process day by day.

What is missing is an underlying mechanism that can explain the processes of change and variation in any system. Recent research of language evolution reveals that a simple (but very powerful) mechanism of frequency can explain the complex processes of change and variation (Haspelmath, 2008a; Haspelmath, 2008b).

2.2 Power and Efficiency

What are the potential triggers and motivations of language change? Can we find them in other systems? There are two important characteristics that exist 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. 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. Furthermore, we refer to the aspect of efficiency as a "Principle of Least Effort" or Zipf's law (Figure 1). According to this principle, "the rank of a word (in terms of its frequency) is approximately inversely proportional to its actual frequency" (Tullo and Hurford, 2003). The Zipfian distribution applied to the different aspects of natural language (e.g. syntax, morphology, symbolic reference, and others) can explain the "emergence of irregularities in language" (Tullo and Hurford, 2003) and "a certain combinatorial property of words, connectedness [...]" (Cancho et al., 2005). The last aspect is especially crucial in the context of dynamic knowledge systems (e.g. ontologies evolution, dynamic versioning, self-organised collaborative tagging).

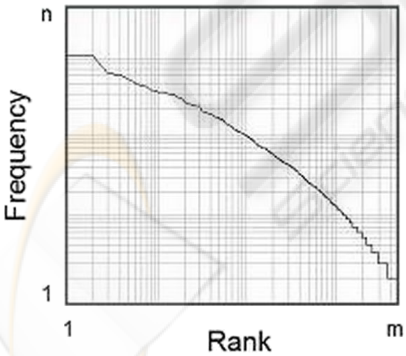

Figure 1: Zipfian Distribution.

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. To illustrate this aspect, three most common mechanisms of simplification can be named: abbreviation/acronyms (*TV – television, Radar – RAdio Detection And Ranging, Laser – Light Amplification by the Stimulated Emission of Radiation*), clipping (*Phone – telephone, Zoo – zoological garden, Fax – facsimile transmission*), and blending (*Vegeburger – vegetarian + hamburger, Smog – smoke + fog, Motel – motor + hotel*).

On the other hand, the power influences the language change and variation as well. Consider, for example, such mechanisms as metaphor (the word *mouse* as a "rodent" got a new meaning, a "computer device") and metonymy (*The White House tried to avoid the scandal* refers to the representatives of the White House). These are the mechanisms that lead to innovations, expressiveness, and understanding in our language system.

At first glance, the power seems to be independent of the frequency. However, the survival of complex structures depends on their statistical properties. The simplification is not equal to the structural reduction like shortening in word's length. It occurs in different fashions at different structural levels of organisation. Hence, natural language has an amazing capacity of balance between efficiency and power. And the underlying mechanism of this balance is the frequency.

3 DIGITAL ENVIRONMENTS: FREQUENCY AND EVOLUTION

3.1 Language System versus Digital Environments: Evolutionary Perspective

Can we apply the frequency-based approach and its findings to the digital environments? The proposed perspective 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. 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.

The frequency-based theory (Haspelmath, 2008b) leads towards a theoretical framework of evolution (not only language evolution, but also evolutionary processes occurring in digital environments) 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. 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. Furthermore, taking usage frequency into consideration helps to design an adaptive (e.g. to the user needs) system.

3.2 Language Networks and their Properties

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). Furthermore, since the network structure underlies the representation of knowledge in digital environments, further investigations of their dynamics are necessary.

In general, the language as a complex dynamical system can be analysed through the lens of network topology (Figure 2). As Sole and colleagues (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. 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 reveal two basic features (Liu et al., 2008):

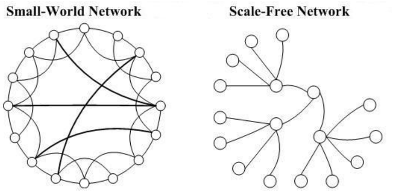

Figure 2: Network Topology (Huang et al., 2005).

- "small world" structure
  Cancho and Solé (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 et al., 2002; Liu and Hu, 2008; Barabási et al., 2003; Ferreira et al., 2006).

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. In Digital Ecosystems, the aspect of knowledge representation involving natural language becomes crucial.

4 CONCLUSIONS

The frequency-based approach 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 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 have to be investigated. The basic law *frequent use → simplification* has different kinds of appearance that mirror the power and economy of any system (language, network, digital environment).

In its infancy, this research opens a new dimension of critical questions. It is important to emphasize that the frequency-based approach does not focus on the origin of an element in a system (e.g. how and why an element occurred in a system). Rather – once it is there – how it develops, interacts with other entities, changes and influences more complex structures. From this perspective, the frequency-based approach seems to provide an explanation of the evolutionary processes occurring in any system. It opens a new perspective of looking at the problem of evolution in "natural" system and link the findings to digital environments. From the linguistic point of view, there is still a lot of research to be done. The role of frequency has been analysed in a variety of languages. However, an intricate question *"is this frequency-based hypothesis true for every language?"* still remains unanswered. Furthermore, the aspect of network topologies needs further investigations. The role of frequency in different types of networks would provide more general picture of the evolutionary processes.

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. As Grudin and Norman (Grudin and Norman, 1991) pointed out:

> 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.

These pressures seem to be the power and efficiency underlain by the mechanism of frequency. Therefore, consideration of this aspect needs to be taken into account during the design and development of a dynamic system (knowledge systems, networks, and interfaces), in general, and specific algorithms that build up such a system, in particular.

ACKNOWLEDGEMENTS

This research is conducted as part of the task "Evolutionary Framework for Language" within the interdisciplinary project (Network of Excellence) Open Philosophies for Associative Autopoietic Digital Ecosystems (OPAALS) (http://www.opaals.org).

REFERENCES

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

Briscoe, G. and Wilde, P. D. (2006). Digital ecosystems: Evolving service-oriented architectures. In *Conference on Bio Inspired Models of Network, Information and Computing Systems. IEEE Press*.

Cancho, F. I. R., Riordan, O., and Bollobás, B. (2005). The consequences of zipf's law for syntax and symbolic reference. In *Proceedings of the Royal Society*, pages 561–565.

Cancho, F. I. R. and Solé, R. V. (2001). The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265.

Dini, P., Rathbone, N., Vidal, M., Hernandez, P., Ferronato, P., Briscoe, G., and Hendryx, S. (2005). The digital ecosystems research vision: 2010 and beyond.

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

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

Haspelmath, M. (2008a). Creating economical morphosyntactic patterns in language change. In Good, J., editor, *Language Universals and Language Change*, pages 185–214. Oxford University Press, Oxford.

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

Huang, C.-Y., Sun, C.-T., and Lin, H.-C. (2005). Influence of local information on social simulations in small-world network models. *Journal of Artificial Societies and Social Simulation*, 8(4):8.

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

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

Motter, A. E., de Moura, A. P. S., Lai, Y.-C., and Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E*, 65(6).

Oller, D. (2004). *Underpinnings for a Theory of Communicative Evolution*, pages 49–65. MIT Press, Cambridge.

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

Steels, L. (2004). *Social and Cultural Learning in the Evolution of Human Communication*, pages 69–90. MIT Press, Cambridge.

Tullo, C. and Hurford, J. (2003). Modelling zipfian distributions in language. In Kirby, S., editor, *Proceedings of Language Evolution and Computation Workshop*, pages 62–75, Vienna.

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