Scale-free Networks in Dictionaries

Ágota Fóris
University of West Hungary

The aim of this paper is to show, through the application of the mathematical model of scale-free networks, how the scale-free network of language is represented in the information contained in dictionaries. Research conducted in the last few decades has proven that every phenomenon of nature and society-the relations of so many various systems-is organised into a complex system of networks. Research has also proven that complex networks can be analysed with the help of a common network model, and that the application of this network theory allows us to discover features of the analysed system that are not observable by other methods. After the discovery of the significance of networks, broad experimental and theoretical studies were launched to reveal the nature of networks and to apply the findings, and research on scale-free networks is the most outstanding among these. If we accept that the three components of the terminological unit may be modelled with the scale-free terminological network model (Fóris 2007), and that the language network is made up of at least these three networks, then we may suppose that dictionaries select and present various parts of this complex network from different approaches. The lexicographers' task-to put it simply-is to collect, record and make the data necessary for language use easily accessible. In order to meet this aim, dictionaries need to follow the three-sided structure of language networks. The various types of dictionaries compiled for different purposes developed a practical structure that reflects the structure of the language network. In the paper, I briefly touch upon the main characteristics of the scale-free network model that can be widely applied in linguistic research, and point out the lexicographic aspects of the model. Based on the network model we can draw conclusions concerning the practical structure of dictionaries. I demonstrate that the complex scale-free network structure of language containing three sub-systems enables us to use the language quickly and completely. I also illustrate and support the features of the language network model and its application with figures.

1. Introduction

The aim of this paper is to show, through the application of the mathematical model of scale-free networks, how the scale-free network of language is represented in the information contained in dictionaries.

Research conducted in the last few decades has proven that every phenomenon of nature and society—the relations of so many various systems—are organised into a complex system of networks. Research has also proven that complex networks can be analysed with the help of a common network model, and that the application of this network theory allows us to discover features of the analysed system that are not observable by other methods. After the discovery of the significance of networks, broad experimental and theoretical studies were launched to reveal the nature of networks and to apply the findings, and research on scale-free networks is the most outstanding among these (e.g. Amaral et al. 2000, Barabási 2002, Csermely 2006).

Previous linguistic research used various network models: for example graphs were used to model terminological systems and sentence structure in generative grammar, and the cob-web model was used to illustrate the structure of the mental lexicon in psycholinguistics. However, these simpler random networks are not suitable for the study of every aspect of linguistic features.

Researching the characteristics of networks brought several surprising results, the most important of which is the fact that the networks that exist in our environment are entirely different previously recognized networks. Computerised studies of models were conducted on the World Wide Web, which is a network that has web pages as its nodes and links as connecting edges, i.e. possibilities of making contact between the web pages (Albert et al. 1999).

Simple, random networks (graphs) were created by throwing dice, but networks of nature and society are created in a different way. Specifically, they self-assemble as follows: the network starts with one node to which newer nodes connect. While evolving, the new nodes have a higher possibility of connecting to nodes that have more relations. As a result, in such a network there are a few nodes (called *hubs*) that have a very large number of relations, and the rest of the nodes have a small number of relations. This fact means that there is such a large difference between the characteristics of various nodes that they are incomparable, meaning we cannot establish a scale within the network: ergo the term 'scale-free network'. Features of the nodes of a random network are situated within the narrow range of an average value, and their distribution can be illustrated by a *bell curve*. The distribution of the features of scale-free networks can be illustrated by a completely different curve; it does not have a maximum, but is rather the curve of a rapidly decaying *power law*. This means in relation to any given node, there are always nodes with more connections in an ever-decreasing quantity. This distribution can be formulated by the $Y=C\ v^{-n}$ power law, where y is the numeric value of a feature of the network, *v* is its range, and C and n are constants of the given network.

Hubs that have a large number of relations play a fundamental role in scale-free networks. It was shown on the model of the World Wide Web that even if we remove 80% of the nodes with a small number of relations, the network does not stop working. On the other hand, removing a small number of the hubs causes the network to cease functioning and break into smaller networks (Barabási 2002). In large scale-free networks the phenomenon called *small world* is always present, which means that any two selected nodes may be connected along just a few edges. This is what we find on the web, or in the case of airline routes. As every object, process, phenomenon, etc in nature and society is a member of several networks at the same time and because small worlds exist in every network, it follows that in our complex environment everything is related to everything else. This realisation forms the basis of, for example, the organisation and management of a complex economic system or the diagnostic and therapeutic methods of modern Western medicine that focus on the body as a whole.

The existence of networks was shown in nature, in society, in biological systems, in the web etc. Then came the realisation that network-like operation is not a unique phenomenon, but a universal feature of complex systems; that general laws regulate the various networks; and that the general laws of networks help us understand the details of various processes.

My hypothesis that the model of scale-free networks may be applied in linguistic research, terminology (see Fóris 2007b) and lexicography is supported by the previous successful use of random network models in linguistic research and by the fact that the theory of scale-free networks has proven to be universal.

In my paper I will summarise findings of my research on the applicability of scale-free networks in linguistics, and then I will show how language networks are reflected in dictionaries.

2. The findings of my preliminary studies

Before research into networks, there were significant linguistic studies proving the existence of language networks and the applicability of the knowledge of networks in linguistics. As early as the 1920's language statistics suggested such discoveries (e.g. Yule 1924, Zipf 1945, 1949). Several findings were published that may indirectly prove the existence of language networks (e.g. Ferrer Cancho-Sole 2001, for details see Fóris 2007a).

I examined the possibility of interpreting the findings of previous linguistic research with the help of the model of scale-free networks (ibid.). I showed that the original formulation of Zipf's law describing the distribution of word frequency is a special case of the general power law, where distribution is characterised by the exponent n=1. This settles the argument over Zipf's law, as in this interpretation any distribution where the n exponent of the power law is other than 1 is not a contradiction, but the consequence of the various characteristics of the given network. This statement was proven by my calculations on the data obtained from publications on quantitative linguistic research (Fóris 2007b).

Based on my own studies and the published data I consider it proven that language has the structure of a scale-free network. With the help of scale-free networks every detail of language can be modelled and described with a system of interconnected networks.

Detailed studies of terms proved that it is not sufficient to approach terms from the concept, because the term is a unit of three components—conceptual, linguistic and communicative—which is defined by Cabré as the *terminological unit* (Cabré 2003). Cabré's *theory of doors* follows from the development of the network model (Fóris 2007b). My terminological research with the help of the network model brought the following findings: the conceptual, linguistic and communicative components of the terminological unit belong to different sections of the language network, and therefore I have described a so-called scale-free terminological network model (ibid.). In actual language usage the communicative role of the terminological unit is determined jointly by the relations formed in the three networks.

3. Lexicographic aspects of the scale-free network model

If we accept that the three components of the terminological unit may be modelled with the scale-free terminological network model, and that the language network is made up of at least these three networks, then we may suppose that dictionaries select and present various parts of this complex network from different approaches. The lexicographers' task—to put it simply—is to collect, record and make easily accessible the data necessary for language use. In order to meet this aim dictionaries need to follow the three-sided structure of language networks. The various types of dictionaries compiled for different purposes developed a practical structure that reflects the structure of the language network.

Let us examine the structural elements of printed dictionaries that show the three-fold structure of the language network.

Dictionaries take the linguistic sign as their starting point and select a headword to represent the lexeme; this is how they present the information stored in the three-fold structure of the *language network*.

The place of the headword in the *conceptual network* is given in various ways and to various depth depending on the type of the dictionary. Providing the meaning with a definition directly describes the features of the denoted concept; in bilingual or multilingual dictionaries the foreign language equivalent is given and thus the conceptual systems of the two languages are linked through the linguistic network. Synonyms and references provide the relations of the nodes of the conceptual network and therefore significantly help to shed light on the communicational side of the relations of different nodes in the conceptual network. Example sentences and cross-references serve to clarify meaning, and to show the relations of the meaning in the communicative network. Cross-references reflect the existence of small worlds, enabling quick advancement through the conceptual system and the division of information, and so with their help we can follow the role of the headword on the network as a whole.

We can find most of the information concerning the *linguistic network* within the entries, and its amount, depth and form depend on the purpose of the dictionary. The part of speech of the headword, the domain of use, grammatical irregularities (such as a noun only used in the plural, or irregular verb conjugation) are basic linguistic information. We have to note that there are dictionaries that contain grammatical descriptions and tables in their appendix, and this bulk of information separated from the linguistic network is linked to the main body of the dictionary through references and symbols in entries.

Example (1): From English–Hungarian Dictionary of technology and science (1993)

Diamond gyémánt; *cipő* gyöngytű; *mat* deltoid, rombusz; *nyom* gyémánt, diamond, 4 pontos betű; *üveg* vágógyémánt; *vasút* (keresztezési) csúcsbetét; **~ antenna** rombuszantenna; **~ bearing** *óra* gyémánt csapágy; **~ bit** → ~ -drill bit; **~ bond** *ép* rombuszos falkötés, rombuszkötés; (…) [The details see Table 1 and Table 2.]

	English word	Hungarian equivalent 1	Hungarian equivalent 2	Hungarian equivalent 3
Domain 1 (General)	diamond	gyémánt	-	-
Domain 2	diamond	gyöngytű	-	-
Domain 3	diamond	deltoid	rombusz	-
Domain 4	diamond	gyémánt	diamond	4 pontos betű
Domain 5	diamond	vágógyémánt	-	-
Domain 6	diamond	(keresztezési) csúcsbetét	-	-

Table 1

	English word	Hungarian equivalent 1	Hungarian equivalent 2	Hungarian equivalent 3	Hungarian equivalent 4
Domain 7	diamond bearing	gyémánt csapágy			
Domain 8	diamond bit → drill-bit (cf. drill[1])	vésőfúró	fúróvéső	fúróél	fúrókorona
Domain 9	diamond bond	rombuszos falkötés	rombuszkötés		

Table 2

Example (2): From Wikipedia, the free encyclopedia (23.03.2008.)

Diamond – Domain: Mineralogy

In *mineralogy*, **Diamond** is the *allotrope of carbon* where the carbon atoms are arranged in an isometric-hexoctahedral crystal lattice. Its hardness and high *dispersion* of *light* make it useful for industrial applications and *jewelry*. It is the *hardest* known *natural material* and the third-hardest known *material* after *aggregated diamond nanorods* and *ultrahard fullerite*.

Diamonds are specifically renowned as a material with superlative physical qualities; they make excellent *abrasives* because they can be scratched only by other diamonds, *Borazon*, ultrahard *fullerite*, or aggregated diamond *nanorods*, which also means they hold a polish extremely well and retain their *lustre*. Approximately 130 million *carats* (26,000 kg) are mined annually, with a total value of nearly *USD $9 billion*, and about 100,000 kg (220,000 lb) are synthesized annually.[2] (…)

Diamond, Syn. Diamondsuit – Domain: Mathematics

In *mathematics*, and particularly in *axiomatic set theory*, (**diamondsuit** or **diamond**) is a certain family of *combinatorial principles*.

Definition

For a given *cardinal number* κ and a *stationary set* , the statement is the statement that there is a *sequence* such that

each

for every is stationary in κ

When $S = \kappa$, is written , and is written

Properties and use

It can be shown that ◊ ⇒ *CH*; also, ♣ + CH ⇒ ◊, but there also exist models of ♣ + ¬ CH, so ◊ and ♣ are not equivalent (rather, ♣ is weaker than ◊).

Charles Akemann and *Nik Weaver* used ◊ to construct a C^*-*algebra* serving as a *counterexample* to *Naimark's problem*.

For all cardinals κ and stationary subsets , holds in the *constructible universe*. Recently *Shelah* proved that for , follows from $2^{\kappa} = \kappa^{+}$.

Information stored in the *communicative network* partly appears in the elements of the dictionary already mentioned; such communicative information is provided in the example sentence or collocations, though not exclusively. The new type of learners' dictionaries provide information that is primarily communicative in nature; they take the communicative network as their starting point and guide the user of the dictionary through the linguistic network to the conceptual units without overloading the learner with detailed descriptions and classifications of the conceptual units.

Example (3): From *English–Hungarian learners dictionary* (Mozsárné Magay 2004)

Diamond (…)

1. I got an earring with a small diamond in it. *Kaptam egy fülbevalót, melyben van egy pici gyémánt*. * She wears a diamond necklace. *Gyémánt gyakláncot visel*.

2. Let's draw a diamond. *Rajzoljunk egy rombuszt*!

3. Diamonds (tsz) The 4 of diamonds is on the table. *A tök/káró 4 az asztalon van*.

Example (4): From Collins COBUILD English Dictionary for Advanced Learners (2001)

Diamond (…)

1. A diamond is a hard, bright, precious stone which is clear and colourless. Diamonds are used in jewellery and for cutting very hard substances. □ *…a pair of diamond earrings. …a sphere made of diamond without impurity or flaw*.

2. Diamonds are jewellery such as necklaces and rings which have diamonds set into them. □ *Nicole loves wearing her diamonds, even with jeans and a white T-shirt.*

3. A diamond is a shape with four straight sides of equal length where the opposite angles are the same, but none of angles is equal to 90˚: ◊. □ *…forming his hands into the shape of a diamond.*

4. Diamonds is one of the four suits of cards in a pack pf playing cards. Each card in the suit is marked with one or more red symbols in the shape of a diamond. □ *He drew the seven of diamonds*. ♦ A diamond is a playing card of this suit.

5. In baseball, the diamond is the diamond-shaped area of the playing field between the four bases.

As we have seen, empirical facts adjusted the traditional dictionary structure to the structure of the language network, but adjusting to this complex structure raised several issues. Computer-assisted linguistics, in particular lexicography and generative linguistics, uses many elements (in both research and application) that originate in the naturally evolved network structure of language. A good example would be the various joined corpora that contain different linguistic information, or the developing system of online dictionaries, which make it possible to access data found in several dictionaries at the same time (see the Example (4), where he cross-references are the links). Computer-based ontologies emphasize the conceptual aspect of the language network. The development of lexicography indicates that the complex dictionaries of the future will be linguistic databases joined by computers that model the language network. Knowledge on scale-free networks may efficiently contribute to the building of such databases (e.g. Miller et al. 1990, Prószéky-Miháltz 2002).

4. Summary

In my paper I briefly touched upon the main characteristics of the scale-free network model that can be widely applied in linguistic research, and pointed out the lexicographic aspects of the model. Based on the network model we can draw conclusions concerning the practical structure of dictionaries. I showed that the complex scale-free network structure of language containing three sub-systems enables us to use the language quickly and completely.

References

Dictionaries

Magay T.; Kiss L. (eds.) (1993). *Angol–magyar műszaki és tudományos szótár. English–Hungarian Dictionary of technology and science*. Vol. I-II. Budapest: Akadémiai Kiadó.

Collins COBUILD English Dictionary for Advanced Learners. 3rd ed. Glasgow: HarperCollins Publishers, 2001.

Mozsárné Magay, E. (2004). *Angol–magyar szótár kezdő és haladó nyelvtanulóknak* [English–Hungarian learners dictionary]. Szeged: Grimm Kiadó.

Wikipedia, *www.wikipedia.com* [on-line] [23.03.2008.]

Other literature

Albert, R.; Jeong, H.; Barabási, A. L. (1999). "Diameter of the World Wide Web". *Nature* 401. 130-131.

Amaral, L. A. N.; Scala, A.; Barthélémy, M.; Stanley, H. E. (2000). "Classes Of Small-World Networks". *Proceedings of the National Academy of Sciences* 97. 11149-11152.

Barabási, A.-L. (2002). *The New Science of Networks*. Cambridge: Perseus.

Cabré, M. T. (2003). "Theories of Terminology. Their description, prescription and explanation". *Terminology* 9 (2). 163-200.

Csermely, P. (2006). *Weak Links. The Universal Key to the Stability of Networks and Complex Systems*. Heidelberg: Springer Verlag. *www.weaklink.sote.hu/weakbook.html*

Ferrer Cancho, R.; Sole, R. V. (2001). "The small-world of human language". *Proc. Roy. Soc. B*. 268. 2261-2256.

Fóris, Á. (2007a). "A skálafüggetlen hálók nyelvészeti vonatkozásai" [Linguistic Applications of Scale-free Networks]. *Alkalmazott Nyelvtudomány* VII (1-2). 105-124.

Fóris, Á. (2007b). "Terminology and the Theory of Scale-free Networks". In Fóris, Á.; Pusztay, J. (eds.). *Proceedings of the international conference "Current Trends in Terminology" (Terminologia et Corpora IV)*. Szombathely: BDF, in press. [*http://www.termik.bdf.hu*]

Guiraud, P. (1954). *Les caractères statistique du vocabulaire*. Paris: PUR.

Miller, G. A. et al. (1990). "Introduction to WordNet: an On-line Lexical Database". *International Journal of Lexicography* 3. 235-244.

Prószéky, G.; Miháltz, M. (2002). "Automatism and user interaction: Building a Hungarian WoldNet". In *Proceedings of the 3rd International Conference On Language Resources and Evaluation. (LREC 2002)*. Vol. III. Las Palmas 957-961.

Yule, G. U. (1924). "A mathematical Theory of evaluation, based on conclusions of dr. J. Willis FRS". *Philosophical Transactions*. 213-221.

Zipf, G. K. (1945). "The meaning-frequency relationship of words". *The Journal of General Psycology* 33. 251-256.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.