Chapter 1

Complex Networks in Different Languages: A Study of an Emergent Multilingual Encyclopedia*

F. Canan Pembe[1,2] and Haluk Bingol[2]
[1] Dept. of Computer Engineering, İstanbul Kültür University,
34156 Bakırköy, İstanbul, Turkey
[2] Dept. of Computer Engineering, Boğaziçi University,
34342 Bebek, İstanbul, Turkey
canan.pembe@boun.edu.tr, bingol@boun.edu.tr

Abstract

There is an increasing interest to the study of complex networks in an interdisciplinary way. Language, as a complex network, has been a part of this study due to its importance in human life. Moreover, the Internet has also been at the center of this study by making access to large amounts of information possible. With these ideas in mind, this work aims to evaluate conceptual networks in different languages with the data from a large and open source of information in the Internet, namely Wikipedia. As an evolving multilingual encyclopedia that can be edited by any Internet user, Wikipedia is a good example of an emergent complex system. In this paper, different

* This work is partially supported by Bogazici University Research Fund under the grant number 06A105.

from previous work on conceptual networks which usually concentrated on single languages, we concentrate on possible ways to compare the usages of different languages and possibly the underlying cultures. This also involves the analysis of local network properties around certain concepts in different languages. For an initial evaluation, the concept “family” is used to compare the English and German Wikipedias. Although, the work is currently at the beginning, the results are promising.

1.1. Introduction

The study of complex networks has started to receive a great attention in recent years [Newman 2003]. Some of the types of networks mostly studied include the Internet, social networks and biological networks. Language has been another interesting area of this research.

There is some work in the literature regarding the language as a complex network. In [Motter 2002], a linguistic network is formed by connecting words about similar concepts using a thesaurus data, and this network is shown to be both small-world and scale-free which are the two of the commonly used properties in describing complex networks. In another such work, data from two online dictionaries are used to form conceptual networks [Batagelj 2002]. In that work, the network is constructed by connecting the entries to other entries used in their definition. The previous work on this subject mostly concentrated on single languages. The aim of this work is to investigate such conceptual networks in different languages and use some network properties to compare the usage or importance of the concepts in different languages.

One of the prerequisites of this work was to find suitable data to construct the networks. There are several dictionaries and thesauri available in different languages, some of which are available in electronic format, including English WordNet [Fellbaum 1998], Turkish WordNet [Bilgin 2004], dictionaries in several languages etc. The problem of using such data sources in a work of analyzing and comparing different languages is that, the format, content and completeness of each dictionary may be different, because these dictionaries may have been created by different organizations with different purposes in mind. The data to be used in such a work should be comparable both in coverage and format for the different languages involved. Therefore, Wikipedia project[1] is selected as the data source for the analysis.

The rest of the paper is organized as follows. The reason to select Wikipedia as the data source and the properties of it are given in Section 2. This is followed by the methodology of the network construction in Section 3 and some of the initial results obtained for English and German Wikipedias in Section 4. Then, the conclusion is given in Section 5.

1.2. Wikipedia

Wikipedia [Wikipedia 2006], started in 2001, is a multilingual web-based encyclopedia project and is considered as one of the greatest inventions in the World Wide Web after

[1] http://www.wikipedia.org

the invention of Google[2]. Some of the most important properties of this encyclopedia are both its free content and its openness. Any Internet user can create new entries or edit existing ones without even the necessity to register. This openness results in a dynamically evolving and developing source of data. There are almost no restrictions on the editing of the encyclopedia entries. Any user can even delete the passages written by others. In such a free environment, the control of the encyclopedia is maintained in a distributed fashion. This is achieved by making all the versions regarding the edits publicly available in addition to the current version of each entry. Users can correct wrongly written entries or recover intentionally deleted passages owing to the availability of versions. In this way, the encyclopedia evolves in a natural and fast way. Due to these properties, Wikipedia becomes a good example of a complex and emergent system and an interesting data source for research.

Using Wikipedia as the data source in this project has several advantages. First, all the Wikipedias in different languages have the same format where each encyclopedia entry corresponds to a page and users can make links to other entries from a page. This makes the comparison of it for different languages possible. Second, the encyclopedia is being created from scratch and is open to everyone. As a result, it is a large and natural data source for obtaining results on different languages.

Wikipedia is available in over 200 languages. Some of the Wikipedias are more active, including the English and German versions. In Table 1, some of the larger versions in different languages are given together with their current number of articles. In this work, English and German Wikipedias are used as data.

Table 1. Some of the largest versions of Wikipedia

Language	Number of Articles
English	937,803
German	343,612
French	226,032
Polish	189,106
Japanese	174,476

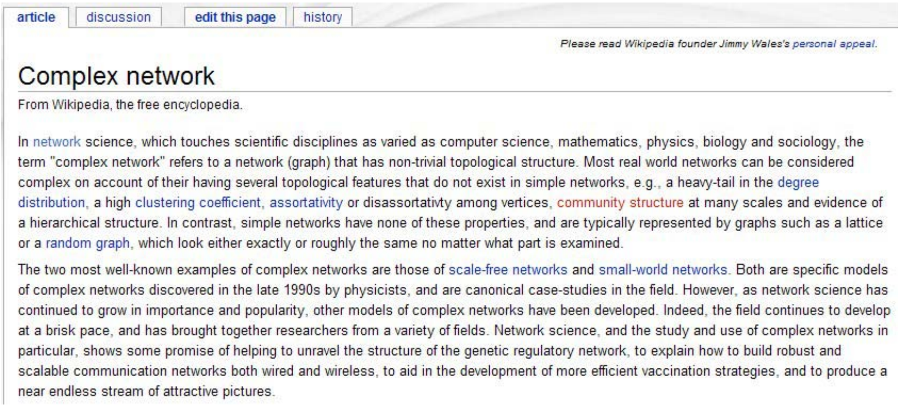
1.3. Construction of the Networks

The methodology is to construct conceptual networks for Wikipedias in different languages and comparatively analyze some of the properties of these networks. The data of both English and German Wikipedias is obtained as two XML files. Each XML file contains all the entries together with their content in that particular language. These data are collected at different time points and made available for research purposes at the Wikipedia site. In this work, data dumps obtained in December 2005 are used for both English and German.

Each entry page contains links to other entries which are embedded into the content. An example portion of an entry is given in Figure 1. The network can be formed in a natural way based on these data by connecting entry titles to the entry titles hyperlinked

[2] http://www.google.com

within the content of that entry; e.g. the entry "complex network" is connected to the entry "degree distribution" in the example of Figure 1.



Please read Wikipedia founder Jimmy Wales's personal appeal.

Complex network

From Wikipedia, the free encyclopedia.

In network science, which touches scientific disciplines as varied as computer science, mathematics, physics, biology and sociology, the term "complex network" refers to a network (graph) that has non-trivial topological structure. Most real world networks can be considered complex on account of their having several topological features that do not exist in simple networks, e.g., a heavy-tail in the degree distribution, a high clustering coefficient, assortativity or disassortativity among vertices, community structure at many scales and evidence of a hierarchical structure. In contrast, simple networks have none of these properties, and are typically represented by graphs such as a lattice or a random graph, which look either exactly or roughly the same no matter what part is examined.

The two most well-known examples of complex networks are those of scale-free networks and small-world networks. Both are specific models of complex networks discovered in the late 1990s by physicists, and are canonical case-studies in the field. However, as network science has continued to grow in importance and popularity, other models of complex networks have been developed. Indeed, the field continues to develop at a brisk pace, and has brought together researchers from a variety of fields. Network science, and the study and use of complex networks in particular, shows some promise of helping to unravel the structure of the genetic regulatory network, to explain how to build robust and scalable communication networks both wired and wireless, to aid in the development of more efficient vaccination strategies, and to produce a near endless stream of attractive pictures.

Fig. 1. Example portion from a Wikipedia entry

The dataset contains large amounts of data. In the network built, entry titles correspond to vertices whereas the edges correspond to distinct links from an entry page to other entry pages. Some of the properties of the dataset are given in Table 2.

Table 2. Properties of the datasets used

	English	German
Size of contents (in Gigabytes)	3.71	1.45
Number of vertices	1,587,127	518,070
Number of edges	16,355,115	7,682,508
Average degree	10.3	14.8

1.4. Analysis of English and German Wikipedias

In this work, some global properties of Wikipedia networks, such as the diameter, is not calculated due to the enormous size of the network and time restrictions. There is some research on Wikipedia showing that the degree distribution of the network shows scale-free properties [Voss 2005]. In another work, Jon Kleinberg's HITS and Larry Page and Sergey Brin's PageRank algorithms are used to extract the most central vertices from the English Wikipedia network [Bellomi 2005]. However, the research on Wikipedia is currently at the very beginning and most of the work concentrates on a Wikipedia in one language.

The aim of this work is not to concentrate on the network properties of Wikipedia in a single language. Instead, the aim is to comparatively analyze conceptual networks in different languages and this may also involve the analysis of local network properties around a certain concept in different languages.

First, the degree of vertices in English and German Wikipedias are considered. The average degree of the vertices is given for both of the languages in Table 2. As seen, although the number of vertices; i.e. entries, in German Wikipedia is much less than that of English, the average degree of vertices is higher. When the degrees of individual vertices are considered, it is seen that they reach 3000s in the English version and 2000s in the German one. However, such big numbers usually occur for entries which are lists in both of the languages, such as the entry with the title “List of airlines” and degree 1777.

Next, the concept of “family” (“familie” in German) is investigated in the Wikipedia networks for both of the languages. For this purpose, the 1-neighborhood and 2-neighborhood of the concept is considered. The number of distinct concepts in the neighborhoods of the concept and the clustering coefficient of “family” for both of the languages are given in Table 3. The clustering coefficient $C_2'(v)$ is calculated as in (1) and (2) where $|E(G^1(v))|$ is the number of edges among vertices in 1-neighborhood of vertex v, $|E(G^2(v))|$ is the number of edges among vertices in 1- and 2-neighborhood of v and Δ is the maximum degree of a vertex in the network [Batagelj 2002]:

$$C_2'(v) = \frac{\deg(v)}{\Delta} C_2(v) \tag{1}$$

$$C_2(v) = \frac{|E(G^1(v))|}{|E(G^2(v))|} \tag{2}$$

The initial results regarding the concept “family” are interesting. First, it was stated that the network size of English Wikipedia is much larger than that of German Wikipedia as given in Table 2. However, Table 3 shows that the number of concepts associated with the “family” concept is much larger in German version than the English one. Also, the “family” concept is more clustered in the German version. When considering the possible authors of the encyclopedia, the authors of the English version are expected to be more diverse. That is, this will include English-speaking Internet users from all around the world as well as those located in the U.S. or the U.K. In this manner, the number of authors in the German version of the encyclopedia is expected to be more limited when compared to the English version. Then, the question of whether the family concept is more important within the German culture. Although there may be several factors behind these initial results, extending this work to other concepts or network properties may be promising.

Table 3. Number of vertices in the 1- and 2-neighborhoods of “family” and clustering coefficient of “family”

	English	German
Number of vertices in 1-neighborhood	44	100
Number of vertices in 2-neighborhood	2530	3373
Clustering coefficient (*1000)	0.03	0.20

1.5. Conclusion

In this paper, an emergent multilingual encyclopedia, Wikipedia, is investigated as a large and evolving complex system. Different than previous work on conceptual networks which usually concentrated on single languages, we concentrated on possible ways to compare the usages of different languages and possibly the underlying cultures. For an initial evaluation, the concept "family" is used to compare the English and German Wikipedias. Although, the work is currently at the very beginning, the results are promising. As future work, the investigated concepts may be extended possibly with the use of categories; for example all the concepts in the category of "education".

References

Batagelj V., Mrvar, A., & Zaveršnik, M., 2002, Network analysis of dictionaries, Jezikovne tehnologije / Language Technologies, T. Erjavec, J. Gros eds., Ljubljana, **135**.

Bellomi F., & Bonato, R., 2005, Network Analysis for Wikipedia, in *Proceedings of Wikimania 2005*, Frankfurt, Germany.

Bilgin, O., Çetinoğlu, Ö., & Oflazer, K., 2004, Morphosemantic Relations In and Across Wordnets: A Preliminary Study Based on Turkish, in *Proceedings of the Global WordNet Conference*, Masaryk, Czech Republic.

Fellbaum, C. (Ed.), 1998, *WordNet- An Electronic Lexical Database*, the MIT Press.

Motter, A. E., de Moura, A. P. S., Lai, Y.-C., & Dasgupta, P., 2002, Topology of the conceptual network of language, in *Physical Review E,* 65, 065102.

Newman, M. E. J., 2003, The Structure and Function of Complex Networks, in *SIAM Review*, **167**, 45.

Voss, J., 2005, Measuring Wikipedia, in *Proceedings 10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm.