Evolutionary Aspects of Language Spreading: A Quantitative Research of Memetic Selection Processes in the Distribution of Virus Hoaxes.

PREVIOUS RESEARCH

As one of the founding members of the board of editors of the Journal of Memetics, the first peer-refereed journal in the field, Francis Heylighen has made several contributions to the field of memetics. He is the author of several papers on the subject (Heylighen 1997, 1998), has presented this work at international congresses, and was the first person to organise a symposium specifically on memetics. Recently he supervised a Master's thesis at the VUB about the possibilities of empirical testing of the memetic selection criteria (Chielens 2003).

INTRODUCTION

In 1976 Dawkins coined the term 'meme' as a cultural equivalent of the biological gene, a self reproducing information pattern. Examples of memes are inventions, ideas, traditions, melodies and chain letters. Each one of these information systems spreads by means of communication from one to several carriers. Thus, a successful meme can be compared to a cultural virus that exponentially spreads through ever large groups. Over the past decade, a growing number of publications has been devoted to memetics (Blackmore 2000, Aunger 2001,…) proposing explanations for phenomena from viral marketing to the popularity and spread of religion.

The memetic approach has been criticized by several authors (Aunger 2001) who have accused it of biological reductionism. Next to this rather ideologically influenced critique two main shortcomings were pointed out, the first one being that it is hard to define what exactly the unit of a meme is and the second one stating that the vague theoretical statements of memetics as yet have not passed any empirical tests (Edmonds 2002).

AIM

This research project aims to show that empirical studies within the field of memetics are possible and can lead to new insights in certain cultural phenomena.

By focusing on distinct entities, computer virus hoaxes (cf. Sophos), we will tackle the first problem with memetic modelling and by quantitative analysis of the spread of these units we will try to show that the theory of memetic selection can be empirically tested.

FOCUS

The focus of this research is on virus hoaxes, i.e. email messages warning the recipients for a non-existent computer virus, and urging them to pass on this warning to as many other people as possible. As such a virus hoax is an excellent illustration of a self-replicating message, that parasitizes the attention and computational resources of its recipients in order to maximally multiply itself. The wide expansion of electronic communication makes it necessary to take a closer look at the possible dangers of these virus hoaxes, which are threefold:

1) Virus hoaxes often propose methods of "protection" that are actually harmful (such as erasing essential program files).

2) They can create panic among naïve computer users by making them falsely believe that their computer is showing symptoms of a virus.

3) They produce economic damage by making their readers focus on the hoax instead of other activities, which results in a loss of time, energy, bandwidth and other resources.

By using memetics as a modelling framework we can analyse the qualities that a hoax needs to succeed in spreading. Thus, this project may not only lead to a deeper and better understanding of the spreading of memes in general but it may also contribute towards the defense against the threat of parasitic information.

Memetics is in the first place inspired by biological principles, but the observation of memes necessitates methods from the humanities and the social sciences. The study of virus hoaxes, which exist as a written corpus, allows us to use the tools of linguistics for their study.

RESEARCH HYPOTHESES

The core idea of memetics is that the success of a meme is determined by natural selection. Several memes are in competition for the attention of potential hosts and only those memes that are well adapted to the present socio-cultural environment will spread; the others will become extinct. This leads us to the generic prediction that "fitter" (i.e. better adapted) memes will be more widespread than less fit ones. To operationalize this as yet very abstract (and to some degree tautological) idea we need to formulate concrete fitness criteria or selection criteria that specify the degree to which a meme is adapted to its environment.

Heylighen (1997/1998) and Castelfranchi (2001) have proposed several such criteria on the basis of theoretical considerations. Some examples of these criteria are: simplicity (complex memes will have more difficulty to be accepted), novelty (an unexpected or original meme will attract more attention so have a higher possibility to be spread on) and authority (the apparent confirmation of a meme by an instance considered 'trustworthy' will increase its probability to be accepted).

We can expect that for each of the criteria separately, and for all criteria together, a high scoring hoax will appear more frequently than a low scoring one. If this hypothesis is confirmed this would allow us to make predictions as to which hoax will be more successful than other hoaxes. This would show that (at least certain aspects of) memetic models can stand up to quantitative, empirical testing.

METHODOLOGY

The advantage of using virus hoaxes is that they can be analyzed in several different ways. In the first place these memes are written which allows for linguistic analysis; moreover they spread through electronic communication so that the degree of spreading can be determined automatically (search engines); finally, the motivation of potential hosts can be measured using sociological methods such as surveys.

To test the hypotheses we will determine the statistical correlation between the score of a hoax on one of the criteria and an estimate of the degree of spreading of this hoax., taking into account that enough different hoaxes are analyzed so as to achieve statistical significance.

To be able to measure the degree of spreading (and thus the success) of a hoax, it is necessary to determine the exact content of the hoax text. Hoaxes are available in a number of specialized databases maintained by different organizations on the internet. By comparing the different sources it is not only possible to find the most prevalent form but it is also possible to compare the strength of different mutations of the hoax. This can then be used to uncover and recreate the evolutionary path that the hoax has followed, including the different mutations and the specific strengths and weaknesses of each mutation (Bennett 2003).

When the precise form is determined, it is possible to extract two or three distinguishing strings of text that only appear in this hoax. The goal of this operation is to make sure that every hoax can be characterized in a unique way by combining these fragments, so that every instance of this hoax will contain these fragments but no other online document contains them. Using these characteristic fragments, it is possible to determine the number of documents in which this hoax appears on the Internet by using one or more search engines (eg. Google, AltaVista, …). These search engines can not only find the number of instances of each significant string but can also determine how often these strings appear together on the internet, both on webpages or in newsgroups.

The scoring of the selection criteria can be done in two different ways: objective or subjective. Certain criteria can be measured objectively by applying linguistic tools directly on the hoax text. Simplicity for example can be scored with the aid of Flesch Kincaid or Gunning-Fog readability measures, or the average sentence or word length. Specificity can be determined by measuring the type-token ratio, lexical density or the number of low-frequency words (cf. Biber 1988, Dugast 1980).

Other criteria can only be measured subjectively. For example, the degree of novelty or danger will be estimated differently by different readers. To achieve sufficient reliability, the same hoax will be evaluated by a large group of subjects and the average score will be used. As an additional check, the same criteria will also be scored by a smaller group of experts.

Certain demographical data about the respondents will also be gathered during the survey, such as the level of schooling or the degree of knowledge of the language in which the hoax is written. We hypothesize that different categories of respondents

will accord a different order of importance to the criteria. For example, subjects with a higher level of education may give less weight to the "simplicity" criterion, but more to "specificity" or "authority". This applies especially to the subtleties of language. For example, the credibility of a hoax may drop significantly for a native speaker because of small grammatical errors that are not noticed by a foreign language speaker.

The analysis of the data will not only look for positive or negative correlations but will also search for a possible optimal score for certain criteria if there is no monotonic relation between scores and degree of spreading (for example, a hoax should be simple to understand, but not too simple in order not to lose credibility, so that there is an optimal degree of simplicity).

The empirical data, their interpretations and theoretical conclusions of this research will be presented in a doctoral dissertation at the end of the project.

REFERENCES

Aunger, Robert (Editor): Darwinizing Culture: The Status of Memetics As a Science, Oxford University Press; (2001)

Bennet, Charles H. & Li, Ming & Ma, Bin. "Chain Letters & Evolutionary Histories." Scientific American Jun. 2003: 64-69.

Biber, Douglas. Variation across speech and writing. Cambridge, Cambridge University Press, 1988

Blackmore, Susan The Meme Machine, Oxford University Press; (2000)

Castelfranchi, Cristiano. "Towards a Cognitive Memetics: Socio-Cognitive Mechanisms for Memes Selection and Spreading." Journal of Memetics – Evolutionary Models of Information Transmission 5, 2001. <http://jom-emit.cfpm.org/2001/vol5/castelfranchi_c.html>.

Chielens, Klaas. The Viral Aspects of Language: A Quantitative Research of Memetic Selection Criteria. Seniors Thesis VUB 2003.

Dawkins, Richard. The Selfish Gene. 2nd ed. Oxford, Oxford University Press, 1989.

Dugast, Daniel. La Statistique Lexicale. Geneva, Slatkine, 1980.

Edmonds, Bruce. Letter: "Three Challenges for the Survival of Memetics." Journal of Memetics - Evolutionary Models of Information Transmission 6, 2002. <http://jom-emit.cfpm.org/2002/vol6/edmonds_b_letter.html>.

Heylighen, Francis. "Objective, subjective and intersubjective selectors of knowledge" Evolution and Cognition. 3,1 1997. 63-67.

Heylighen, Francis. "What makes a meme successful? Selection Criteria for Cultural Evolution." Proc. 16th Int. Congress on Cybernetics. Namur: Association Internat. de Cybernétique. 1998.

Sophos describes hoaxes and scares. Sophos. 21 Jan. 2004 <http://www.sophos.com/virusinfo/hoaxes/>.