In: Proceedings of the International Workshop on the Emergence and Evolution of Linguistic Communication, Kanazawa, May 2004

Mapping, Measuring, and Modelling the Diffusion of Linguistic Material on the Internet

Gregor Erbach
German Research Center for Artificial Intelligence
Language Technology Lab
66123 Saarbrücken, Germany
erbach@dfki.de

(Extrended Abstract, 09 May 2004)

1. Introduction

This paper is concerned with methods for studying short-term linguistic evolution, i.e., language change taking place within recent history. Such short-term language change involves the diffusion of linguistic material over time across geographical space and different socio-linguistic communities. We propose to use the Internet - as a large-scale repository of recorded language use - as a data source for studying linguistic diffusion processes, and discuss the technical prerequisites of dating and geographic and social localisation of web pages, and of identifying linguistic communities. Finally, we report on preliminary investigations carried out by using a web corpus for a linguistic sub-community of technology experts.

The evolution of linguistic communication is to a large extent a process of the diffusion of linguistic material from one point of origin to different socio-linguistic communities [1,2]. Dawkins has first articulated the idea that a "meme" (a cultural artefact such as an idea, a melody, or a linguistic phenomenon) can play a similar role in cultural evolution as a gene in biological evolution [3].

Diffusion of memes can take place through movement of people, linguistic contact, and mass media such as books, newspapers, radio and television - and recently electronic communication networks, notably the Internet and the World Wide Web. The Internet - with its 3.5 billion publicly accessible web pages and 800 million usenet articles - constitutes a rich repository for observing the diffusion of linguistic material between different communities on a large scale, and in real time.

Using the Internet makes it possible to process a much larger mass of observable data than other methods of data collection, such as direct observation in the field. However, the disadvantage is that the temporal, geographical and social context in which the data originated is generally not known. Therefore, great care must be taken to make sure that the observable data are properly annotated with information about time of creation, geographical location, and social and linguistic context.

Linguistic diffusion has a temporal, a geo-spatial, and a social dimension. As communication networks interconnect the whole globe, geographical distance becomes less important for the diffusion of linguistic material, and the social aspect becomes more relevant. The social topology of cyberspace [4,5], as manifested in hyperlink connectivity between different sites, constitutes communities with common interests, and maybe also shared sublanguages.

In order to study the diffusion of linguistic material through time and (cyber)space - and to develop and verify/falsify theories - it is necessary to obtain a large corpus of documents. In order to study the process of diffusion, the documents in the corpus must be anchored in time and (cyber)space. The challenges involved in the dating and "localising" of documents are the topic of this paper.

This work was supported by the German Federal Ministry for Education and Research (BMBF) through the project COLLATE (01 IN C02).

1.1 Linguistic Material

The linguistic material that is diffused can be of a different nature:

- Lexemes
- Figurative language (metaphor, idioms, collocations)
- Morphology
- Syntax
- Stylistics (quoting conventions, emoticons ...)
- Semantic concepts
- Typographic, Markup and Layout conventions

Examples of the above phenomena are the use of the lexeme "handy" for mobile phone in German, the use of the metaphor "Frankenfood" for genetically modified food [6], and the use of the dative (example 1b) instead of the traditional genitive case (example 1a) after the preposition "wegen" in German.

(1) a. wegen seines Vaters [genitive] (because of his father)
b. wegen seinem Vater [dative] (because of his father)

While data relating to surface phenomena, such as lexical items or collocations, can be obtained by making use of keyword queries to search engines or WebCorp [7], the collection of data relating to other morphological or syntactic phenomena, such as example (1) above, would either require large-scale corpus collection and annotation, or the availability of a Linguist's search engine [8], which allows queries over morphological features or syntactic configurations in linguistically analysed documents.

2. Data Collection and Preparation

Various sources of data can be used as evidence for diffusion processes of linguistic material. The Internet comprises a wide variety of sources, including

- current web pages (obtained through search engines or focussed crawling)
- Internet Archive with historical snapshots of web pages [9]
- newspaper corpora and archives [10]
- publication repositories [11] and digital libraries
- archives of newsgroups [12] and mailing lists
- historical corpora (e.g. project Gutenberg, parliament debates)
- e-mail, chat

The Internet includes very diverse documents such as technical manuals, FAQs, commercial pages, personal homepages, research reports, legislation, weblogs etc. It is important to identify these genres in order to analyse the diffusion from one genre to another one, for example from informal communication to formal publications.

There are three important challenges in the preparation of the data

- accurate dating of the data
- geographic localisation of the data
- socio-localisation of the observed data within social and linguistic communities

Due to the heterogeneous nature of the sources, it is clear that there is no uniform method for dating and localisation, but that specific methods or combinations of methods are required for each source.

2.1 Dating of Pages

Dating is a challenge for two reasons. Firstly, electronic information is immaterial, and hence does not show signs of physical ageing, which could be used to determine its age. Secondly, it is problematic to use the very same linguistic material whose diffusion over time we want to study as an indicator of the time of document creation.

Fortunately, some documents are date-stamped, for example usenet postings, or newspaper articles. For other documents, some "archaeological" work is required, using as evidence metadata added by (different versions of) document creation tools which were only released at a certain time, or certain markup language constructs which were introduced at a certain time. Another useful dating technique is based on the fact that documents making reference to a certain unexpected events (such as the 9-11 terrorist attacks) cannot have been produced before the event. Other dating techniques make use of information extraction techniques for dates in case of dated documents, or analyse the use of dates in combination with tense. In example (2), the use of the past tense in connection with "September 2001" and the future tense in connection with "May 2003" provides evidence that the document was written between these two dates.

(2) The Treaty of Nice was signed in September 2001. In May 2003, the Convention on the Future of Europe will present a draft European constitution.

2.2 Geo-Localisation of Pages

An important problem is the geographic localisation of pages, which is useful in order to investigate whether there exists a significant influence of geographic proximity on the diffusion of linguistic material. Top-level domains and IP numbers are indicators for server location, but have only a weak relation to the geographical location of the page creators.

In cases where a page contains a reference to a geo-location, e.g. a postal address, this may be an indicator of the geographical origin of the page, but must also be treated with caution, according to the document type.

2.3 Socio-Localisation of Pages

"Localisation" means several things in this paper. It means assignment of pages to linguistic communities (e.g. speakers of American English) as well as assignment of pages to communities of practice (e.g. particle physicists or fans of classical music).

Assignment to linguistic communities involves identifying the language, and possibly the dialect, in which the page is written. Reliable language identification techniques exist for this task [13,14].

Next, pages must be assigned to sociolinguistic communities. Since we want to study the diffusion of linguistic material between communities, linguistic features should be used only with great caution for the identification of communities. It is problematic to using linguistic features for the identification of communities, for example by document clustering according to relative term frequencies (TD/IDF and related measures). Instead we propose to use connectivity information as non-linguistic features for the identification of communities. Communities are held together by "hub" pages with a large number of outgoing hyperlinks, and by "authority pages" with large numbers of incoming hyperlinks [15]. As a working hypothesis, we assume that communities that engage in linguistic communication are also strongly connected in through hyperlinks in cyberspace.

Information about media types can also used, in order to identify the influence of mass media, such as newspapers.

3. Preliminary Investigations

We have started a preliminary investigation related to a very specific linguistic community, the group of language technology experts. This limitation allows us to collect a comprehensive collection of web documents for one particular domain. We start out from a collection of manually annotated core documents for the domain (homepages of people, projects and organisations) and follow incoming and out going hyperlinks recursively from this core set. We have started to collect a web corpus for the area of Language Technology, which contains documents harvested from the web, as well as a database of hyperlinks between the documents [16]. We have also obtained the ACL anthology of conference and workshop proceedings in Computational Linguistics as a historical corpus. This will be used for an initial study of the diffusion of new concepts and technical terminology in sub-areas of the LT community.

4. Conclusion

The Internet offers unprecedented opportunities for studying language use within different linguistic and social communities. The available data range from formal publications (scientific articles, newspaper articles) to informal discussions and weblogs, and provide a much more comprehensive data base for the study of linguistic diffusion than has been available previously.

The Internet serves two functions, both related to our field of study. On the one hand, it serves as an archival medium for storing and publishing huge amounts of documents. In this respect, the Internet is a passive repository recording language use (e.g. scientific papers, newspaper articles) that takes place in the "real world". In this function, it is a mirror of language use that takes place outside the Internet. On the other hand, the Internet serves as a communication medium (e-mail, chat, forums, discussion groups), in which language use is shaped and diffusion processes take place. In some cases, both functions (communication and archival) occur simultaneously, for example in web-based discussion forums, and in archived mailing lists and newsgroups. We expect these cases to be the most productive sources for our study as they allow us to observe the diffusion processes very closely.

An open research question is whether hyperlinks also reflect communication pathways between communities, along which linguistic material is diffused. Strictly speaking, hyperlinks are not communication channels, but Park and Thelwall [17] argue that a hyperlink is an indicator of a communication relationship. A testable hypothesis is that an author who makes a reference to other authors by means of hyperlinks will also take over linguistic material from the referred author.

References

[1] Aitchison, Jean. *Language Change: Progress or Decay*. Cambridge University Press. 1991

[2] Britain, Dave. Geolinguistics and linguistic diffusion. In: U. Ammon et al (eds.) *Sociolinguistics: International Handbook of the Science of Language and Society*, Berlin: Mouton De Gruyter. in press

[3] Dawkins, Richard. *The selfish gene.* Oxford University Press. 1976

[4] Dodge, Martin and Rob Kitchin. Mapping Cyberspace. Routledge, 2000

[5] Dodge, Martin and Rob Kitchin. Atlas of Cyberspace. Addison-Wesley, 2002

[6] Hellsten, Iina. Focus On Metaphors: The Case of "Frankenfood" on the web. Journal of Computer-Mediated Communication. 8(4), 2003

[7] WebCorp: www.webcorp.or.uk

[8] Philip Resnik and Aaron Elkiss. *The Linguist's Search Engine: Getting Started Guide.* Technical Report: LAMP-TR-108/CS-TR-4541/UMIACS-TR-2003-109, University of Maryland, College Park, November 2003. http://lse.umiacs.umd.edu:8080/

[9] Internet Archive. www.archive.org

[10] Newspaper archive www.newspaperarchive.com (provides access to scanned historical newspapers, starting from the 18th century)

[11] e-print archive. www.arxiv.org

[12] Google groups (formerly DejaVu). groups.google.com

[13] Cavnar, W. B. and J. M. Trenkle. *N-Gram-Based Text Categorization.* In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, pp. 161-175, 1994

[14] Grefenstette, G. *Comparing two language identification schemes.* In Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data, Rome, Italy, 1995.

[15] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal Of the ACM*, 46(5):604-632, 1999.

[16] Erbach, Gregor. A Web Corpus for Language Technology. In: COLLATE: Final Report. DFKI and Saarland University. Saarbrücken (2003).

[17] Park, Han Woo and Mike Thelwall. Hyperlink analyses of the World Wide Web: A Review. Journal of Computer-Mediated Communication, 8(4), 2003