



Search


[Overview](#)
[Academics](#)
[Research](#)
[Careers](#)
[People](#)
[Contact](#)
[Home](#)[Future Students](#)[Current Students](#)[For Alumni](#)[For Employers](#)[News & Media](#)[Upcoming Events](#)[IU Home](#)[IUB Home](#)[IUB Computer Science](#)[IUPUI Home](#)[IUPUI Informatics](#)[IUPUI New Media](#)[IUSB Informatics](#)[IUS Informatics](#)

## Faculty Research Profiles

# John Paolillo

## Faculty Title

Associate Professor of Informatics

## Research Statement

My research activities are focused on two complementary areas of inquiry: the linguistics and sociolinguistics of computer-mediated communication, and the application of statistical models to formal and computational theories of language. Both of these areas of research support thriving research communities with diverse research agendas, so to see where my research fits, and even brings these two areas together, a bit of explanation may help.

In the past few years, the study of computer-mediated Communication (CMC), especially Internet communication, has emerged as a fast-growing area of research for members of the social science and humanities disciplines.

Linguistics is in a position to make a unique contribution to this new field of study, since language is involved in CMC in the most fundamental way: CMC, especially Internet CMC, is comprised overwhelmingly of textual representations of language. While it is true that new technologies, such as streaming video and audio, are anticipated to become increasingly important in the future, and while present-day technologies make significant use of graphical and audio forms of presentation, these are not likely to seriously diminish the importance of language in internet CMC in the long run.

This is particularly true if we consider the global impacts of internet CMC — the Internet often goes hand-in-hand with globalization in popular visions of the future. Whatever the relation between the two, at the very least, the Internet will bring speakers of the world's 6000+ languages into greater contact than was previously possible. A characteristic outcome



Our faculty res profiles highlig research intere accomplishmer select faculty n from the IU Sc Informatics. [Vi](#)

### SEARCH FOR PUBLICATIC

- [Google Sch](#)
- [CiteSeer](#)
- [Social Scier Research N \(SSRN\)](#)
- [DBLP Biblio Search](#)

### RESEARCH L

- [Informatics Areas](#)

of increased language contact is decreased linguistic diversity, so it is of great interest to Internet CMC research in what ways we can observe this process unfolding, and whether we can understand what its consequences will be.

For this reason I have been very interested, from an early time, in trying to ascertain the effects of Internet CMC on people's use of different languages. My efforts have been focused on (i) issues of online multilingualism, and (ii) large-scale surveys of popular Internet communications media. This latter area is a primary motivator for my research on probabilistic models in Computational Linguistics.

Somehow we must sift through enormous archives of CMC, and try to distill from it the understandings about language diversity that we seek. Fortunately, we are not alone in facing these problems, because this is in essence the same type of problem that is faced in many applications of natural language processing. For a whole litany of reasons, the entire field of computational linguistics/natural language processing has shifted over the last ten years from an almost exclusive focus on formal models and methods to an overwhelming focus on statistical models and methods. At the same time, some very surprising things have emerged, such as the technique called latent semantic analysis (LSA) in which a purely statistical model (of colossal scale) is coaxed into behaving much as a human would be expected to (e.g. on multiple choice tests). Of course, the formal models haven't gone away, as linguists still use them and create new ones. All the same, it has become increasingly clear that there is an acute need to reconcile these two approaches.

My research in this area focuses on the application of logit/logistic and log-linear models to language modeling, and the use of factor analysis to describe dimensions of language variation in a corpus of language. The overlap of these areas of interest with my interests in CMC arises from the need to develop corpus-based methods that can be used to analyze large CMC corpora.

### Select Publications

- "Analyzing Linguistic Variation: Statistical Models and Methods", CSLI Publications, 2002.
- "Language Variation in the Virtual Speech Community", *Journal of Sociolinguistics*, 2001.
- "Formalizing Formality", *Journal of Linguistics*, 2000.
- "Asymmetries in Universal Grammar", *Studies in Second*

*Language Acquisition*, 2000.

### **More Information**

- [Contact John Paolillo](#)
- View [John Paolillo's Web site](#)

### **Indiana University School of Informatics**

© Indiana University | [Site Map](#) | [RSS: News, Events](#) | [Informatics Technology](#) | [Comments: Webmaster](#)  
Page Last Modified: Monday, October 04, 2004