LESLI
Linguistic Evidence in Security, Law and Intelligence

Volume 1, No. 1 (2013) | ISSN 2327-5596 (online) | DOI 10.5195/lesli.2013.8 | http://lesli-journal.org

ILER: A Web-Accessible Resource for Research in Forensic Linguistics

Carole E Chaski PhD
Executive Director, Institute for Linguistic Evidence, USA

Abstract

ILER is a web-accessible platform for empirical research in forensic linguistics. It enables new researchers and practitioners to conduct mature research from literature review, experimental design, experiment review and approval, subject recruitment, human subjects protection and confidentiality agreement, subject demographic information collection, data collection, to data analysis.

Keywords: linguistic data collection, world wide web, Internet, human subjects protection, ground truth data, empirical linguistics, experimental design

Introduction

The World Wide Web provides an enormous amount of linguistic data, especially through social media sites, blogs, and other fora. It can be extremely tempting for linguists and other forensic examiners to simply "scrape the web" for linguistic data. This procedure has three problems. First, there are some ethical issues regarding human subjects (McEnery and Hardie 2012). Even though the web is public, not everyone who posts on the web has agreed to be part of a research study, and scraping the web provides no human subjects protection to the people whose words may be used, and possibly published, in research articles. Second, there is an inherent logical circularity of using unvetted web-data to solve the problem of anonymous web-data (Chaski 2013a). Most forensic issues involving web data arise because screen names can be invented and used at will; people are not who they seem to be on the web. Consequently, and third, the web-scraping procedure does not guarantee ground truth data. For instance, web-scraped data from blogs or Facebook or Twitter should not automatically be used for authorship identification, since it is highly possible that the screen name may not identify one person, or may identify a person whose characteristics are not accurately reported. But ground truth data, in which the data and demographics are accurate for known authors with known characteristics, is necessary for error rates to be correctly and accurately calculated (Chaski 2013b).

As more and more linguistic data are generated electronically, the need for collecting data from electronic media is obvious, since the medium of generation may influence the message, harking back to Marshall McLuhan's communication dictum "the medium is the message." Abbreviations, emoticons, email and texting are well-documented examples of how electronic media can affect language use (Baron, 2008; Crystal 2001, 2008). Further, as the forensic science community embraces "normal science" and experimental procedures (Mnookin et al 2011), the need for examining web-based linguistic data is a valid concern, especially when we consider that an experimental paradigm for validation testing requires data collection in a controlled way. In "normal science" linguistics, experimental research is well-established, with results from the 1960's still viable and valid. This essay presents a resource to researchers in forensic linguistics who want to bring "normal science" linguistics into forensic linguistics.

ILER (Institute for Linguistic Evidence Research) is a platform developed to address these issues, a research resource that enables an experimental paradigm in forensic linguistics to use web-generated data in a way that ground truth data can be collected.

ILER is a web-based platform that has several functions for different types of users. This essay focuses on the researcher's perspective, and describes ILER functions in the order of events that laboratory researchers go through during the process of conducting research. Generally, this process involves a literature review, experimental design, experiment review and approval, subject recruitment, human subjects protection and confidentiality agreement, subject demographic information collection, data collection, and data analysis.[1] ILER supports each of these phases in the process of conducting research in forensic linguistics.

Literature Review

Most researchers first need to understand the gaps in current research and what the community needs to test empirically. ILER includes two libraries of research literature: the scientific and the legal.

The ILER Research Library contains scientific literature in theoretical, applied, computational linguistics, psycholinguistics, sociolinguistics, and computer science as well as forensic linguistics. Each article can be annotated by readers, so that new researchers can see the commentary of previous readers. The annotation form enables the readers to provide both structured and unstructured responses; from these responses, articles are catalogued properly for information retrieval and new researchers are alerted to specific qualities of the research, such as whether an experiment did or did not use ground truth data, types of statistical analysis, etc.

[1] For an overview and examples of experimental designs related to linguistics and psycholinguistics, see Bennett, Hausfield, Reeve and Smith 1982, Campbell and Stanley 1970, Shadish, Cook and Campbell 2001. Federal Regulations for human subjects protection in the United States can be accessed at http://www.hhs.gov/ohrp/humansubjects/commonrule/.

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.8 | http:// lesli-journal.org

The ILER Legal Linguistic Evidence Rulings (LEGLER) Library contains approximately 300 published rulings that relate in some way to linguistic evidence in the United States court system[2]. Again, the annotation form enables both structured and unstructured information enabling efficient search as well as legal notes.

Experimental Design

ILER enables researchers and practitioners to design experiments, using a stepwise process that walks researchers through the design steps. The experimental designs currently coded into ILER include Stimulus-Response and Pre-Test/Post-Test, as these are most common in linguistic and psycholinguistic research, and Validation Testing of Specific Methods, as this kind of testing is crucial for the development of linguistics as a forensic science. Researchers and practitioners can create experiments using their own or ILER stimuli for the collection of relevant data. Stimuli can be written texts, audio snippets or videos. Subject responses can be either textual or metalinguistic. In textual responses, subjects write texts in response to stimuli. In metalinguistic responses, subjects provide scalar responses to stimuli. Scalar responses can be 1/0 (yes/no), 1 to 5 or 1 to 10.

Experiment Review and Approval

In the United States, research subjects must be protected from harm; the first step in this protection is the Internal Review Board (IRB). The Institute for Linguistic Evidence maintains an IRB whose members are a psychiatrist, psychologist, linguist, and attorney. The IRB can access each experiment designed in ILER and provide comments to the researcher for requited changes or immediately approve the experiment.

Subject Recruitment

Access to ILER is restricted; a user cannot sign up and create an account. ILER enables researchers to recruit subjects who access ILER through assigned accounts.

Users are vetted by researchers who are recruiting subjects, and ILE staff or regional co-directors create accounts for vetted users, solving the main problem of getting ground truth data from the Internet. Thus, ILER collects vetted data via the Internet.

As in laboratory experiments, researchers can provide rewards for subjects; ILER keeps track of how many stimuli and responses are required for successful completion of an experiment so that researchers know which subjects have earned rewards.

[2] Thanks to Harry L. Miles, Esq. of Green Miles Lipton LLP, Northampton, MA, USA who provided the initial research for LEGLER.

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.8 | http:// lesli-journal.org

Human Subjects Protection and Confidentiality Agreement

ILER includes human subject protections to solve ethical issues and protect the users. The web platform of ILER adds a level of security features to data collection, but generally ILER users (writers) are protected in ways that are similar and normal to any psycholinguistic or psychological laboratory environment, e.g. secure storage and identifier coding.

There are two aspects to secure storage in the electronic venue. First, user accounts are set to access certain portions of ILER depending on their security levels. Subjects (writers) can only access the experiment portion of ILER, and can only access their own folders for the experiments they have selected, and no other subject's folder. Second, ILER is served from a server facility in Atlanta where Google also has servers; it is a secure facility.

Security features within ILER are set up so that the profile for each ILER user can only be accessed by the user and the system administrator. Even researchers or research site administrators cannot access the user profiles. ILER recodes the writer's screen name into numerical codes so that a researcher or site administrator cannot guess which user is producing which text; this is a common laboratory procedure for human subjects protection translated into an electronic venue.

Further, in the same way that laboratory experiments are monitored, ILER is a closed system so that access is monitored, and the identity of the subjects can be monitored as closely as possible. This close monitoring also means that users from vulnerable populations can access ILER in a safe way. For example, ILE research associates are currently working out agreements with women's shelters and domestic violence shelters so that their clients can use ILER to write about their experiences in a safe environment. For these kinds of vulnerable populations, writing about traumatic experiences can be very healing, but can also be dangerous if their writing notebooks are found by their abusers, or if their screen names are guessed by their abusers. ILER recodes the writer's screen name into numerical codes so that even if a person were to hack the system and find a screen name, the person would not be able to access the written responses in experimental tasks.

When users log into ILER, writers select from a list of available experiments (or writing tasks). Before writers proceed to the experiment, they must agree to a confidentiality agreement that is specific to the selected experiment. Writers agree to this confidentiality agreement each time they access the experiment.

Subject Demographic Information Collection

After users are given their account information and sign into ILER, they are asked to provide demographic and linguistic information. These kinds of information can be used for sorting textual data in experiments for linguistic profiling and other forensic linguistic tasks and as grouping variables in statistical analysis. Writers cannot proceed to experiments unless they have completed the demographic information form.

Data Collection

ILER collects data in two ways. Writers who participate in experiments produce texts and metalinguistic judgments. These texts are stored in the DocData database within ILER for text analysis and statistical analysis. The DocData database also stores texts that have been donated to ILE for research or have been collected in laboratory

experiments prior to ILER or have been collected through other means such as Freedom of Information Act requests or literature searches.

ILER provides datasets for many different types of research in forensic linguistics. Current datasets include ground truth data for author identification, suicide note assessment, threatening communications assessment, complaint and other business related texts, financial documents, and handwriting identification. For most of these datasets, demographic information has been collected about the sources.

ILER enables data aggregation so that the community can share vetted linguistic data for experimental research and validation testing.

Data Analysis

ILER includes text analysis procedures so that non-linguists can access automated text analysis of the collected data. These text analysis routines provide pattern identification and quantification so that groups of texts can be analyzed statistically through statistical routines. Statistical analysis includes some procedures within ILER, but ILER quantification can be exported into Excel format for analysis by commercial and open-source statistical software.

ILER's text analysis routines, known collectively as TATTLER, provide linguistic analysis at the phonemic, phonotactic, lexical, syntactic and semantic levels. Research associates can request new routines for TATTLER, so that TATTLER continually grows as a text analysis system.

References

Baron, Naomi S. 2008. *Always On: Language in an Online and Mobile World*. New York: Oxford University Press.

Campbell, Donald T. and Stanley, Julian C. 1970. Second Edition. *Experimental and Quasi-Experimental Designs for Research.* New York: Rand McNally & Company.

Crystal, David. 2001. *Language and the Internet.* New York: Cambridge University Press.

Crystal, David. 2008. *Txtng: The Gr8 Db8.* New York: Oxford University Press.

Chaski, C.E. 2013a. "Best Practices and Admissibility of Forensic Author Identification." *Journal of Law and Policy.* Volume 21, No. 2. Brooklyn Law School.

Chaski, C.E. 2013b. "Five Data-Handling Issues in Forensic Linguistics." Linguistic Society of America Annual Meeting, Boston, MA, USA

McEnery, T. and Hardie, A. 2012. *Corpus Linguistics.* New York: Cambridge University Press.

Mnookin, J., Cole, S., Dror, I., Fisher, B.A.J., Houck, M., Inman, K., Kaye, D.H., Koehler, J.J., Langenburg, G., Risinger, D.M., Rudin, N., Siegel, J., and Stoney, D.A. 2011. "The Need for a Research Culture in the Forensic Sciences." *UCLA Law Review*, Volume 8.

Shadish, William R., Cook, Thomas D. and Campbell, Donald T. 2001. Second Edition. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Cengage Learning. www.cengage.com