Detecting Abstract Linguistic Properties Through the Study of Corpus Data

Anthony Kroch & Beatrice Santorini
University of Pennsylvania

For obvious reasons, the focus of much corpus linguistic research is on the surface word forms and strings that are available in all electronic corpora. As linguists, however, we are aware that language has structure which is not directly audible/visible on the surface. In order to study that invisible structure more effectively, we have been creating, in collaboration with others, a range of annotated historical and synchronic corpora of English and other languages. In this talk, we will briefly present the annotation scheme for these corpora and then will illustrate the value of such annotation with a case study on the loss of OV word order in English. In presenting this case study, we take advantage of the possibility of comparing the English development with parallel developments in French, Icelandic and Yiddish, three languages for which we now have annotated historical corpora in the Penn Historical Treebank style.