Patrick Juola, Ph.D.

Professor of Computer Science, Duquesne University (Pittsburgh)

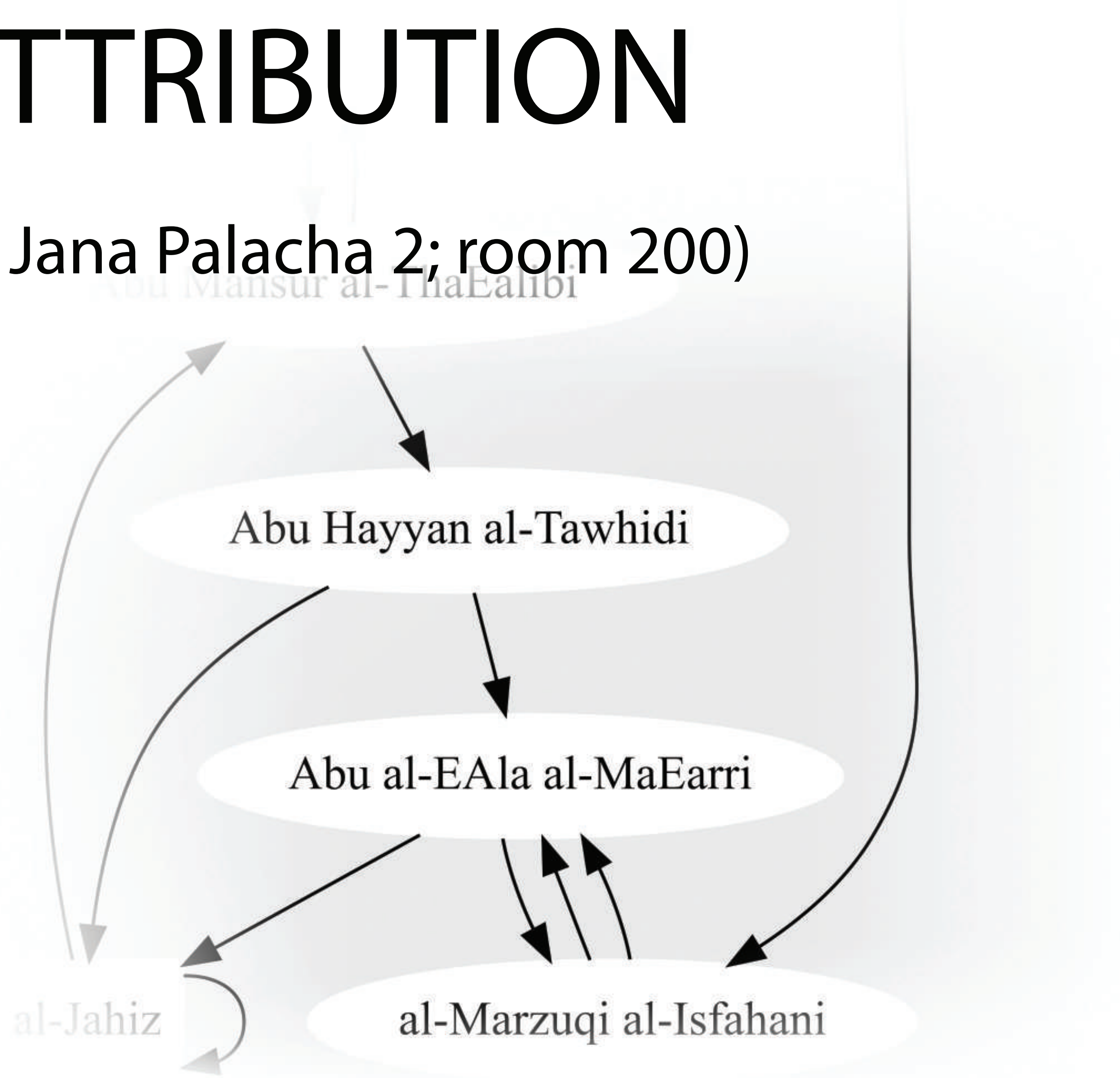
CROSS-LINGUISTIC REGULARITIES IN AUTHORSHIP ATTRIBUTION

(Monday 29 May 2017, 14:00; náměstí Jana Palacha 2; room 200)

Authorship attribution is a well-studied problem in corpus and quantitative linguistics, with many applications ranging across forensic science, history, journalism, and beyond. We begin by reviewing some case studies and discuss some of the major findings common to the field as a whole, as well as the psycholinguistic basis for authorship attribution.

One key methodological weakness is the need for closely matched samples. For example, to determine whether a particular novel written in English was by a specific person, we would ideally like other samples of novels, also written in English. Trying to analyze writings on the basis of a non-novel, or worse, a work in another language, would presumably introduce systematic errors, or worse, render the analysis impossible because many of the features (like words) would not be present in both samples! At the same time, cross-linguistic authorship attribution is a real problem, especially in Europe where nearly everyone is polylingual.

We discuss the use of cognitive and sociolinguistic universals as a basis for authorship attribution and describe two experiments in Spanish-English and Greek-English paired corpora. We identify several specific linguistic attributes that can be measured cross-linguistically and show how these features can be used for authorship attribution irrespective of language.


Mansur al-ThaEalibi
Abu Hayyan al-Tawhidi
Abu al-EAla al-MaEarri
al-Jahiz
al-Marzuqi al-Isfahani