

## Arabic Script-based Languages deserve to be studied linguistically

**Martin Kay**  
*Stanford University*

Arabic script-based languages are attracting increased attention for reasons that are regrettably far from their intrinsic linguistic interest. At the same time, statistical and corpus-based approaches to language processing are acquiring such dominance that it is becoming difficult for the advocates of other methods even to receive a hearing. I will argue that this is an alarming trend against which computational linguists, and especially those studying these languages, should resist with great determination. My argument for this position rests on the following observations:

1. Unless the role of quantum mechanics and chaos in the workings of ordinary language has been grossly underestimated, nothing about the subject is probabilistic in any fundamental sense.
2. The statistics are a surrogate for knowledge of the world and artificial intelligence and the performance of any system based on an approach that reduces this to numerical annotations on linguistic structures can only hope to reach a very low asymptote.
3. Thanks to Zipf's law, corpus annotation is subject to a severe law of diminishing returns to which the linguist's search for significant generalizations is not subject.
4. To the various levels of linguistic analysis and to the indefinite range of subjects and propositions that texts may treat, there correspond different notions of locality, each requiring its own statistical models.
5. Most importantly, most of the linguistic properties that must be considered for text processing are not emergent properties of the texts at all but crucially depend on *l'arbitraire du signe*, the arbitrary relation between a symbol and what it symbolizes.