Linguistic Structure and Short Term Memory

Emmanuel M. Pothos* (e.pothos@bangor.ac.uk), **Patrick Juola****, (juola@mathcs.duq.edu), & **Nick C. Ellis*** (n.ellis@bangor.ac.uk)

*School of Psychology, 39 College Road
Bangor, LL57 2DG, UK

**Department of Mathematics and Computer Science,
Duquesne University, Pittsburgh, PA 15282

Introduction

The induction of linguistic knowledge displayed by children is perhaps one of the most puzzling aspects of learning performance, because of the intricacies and complexities of language and also because of the unerring accuracy with which (apparently) most people seem to approximate the same linguistic structure (within a society). The latter observation suggests that language learning depends on universal features of the learning process. The feature we examine here is the short term memory (STM) span. If STM is relevant in language acquisition then we expect language structure to reflect the STM span. In other words, the STM span will be reflected in language only insofar that it is a relevant aspect of the language learning problem.

Mutual information and linguistic structure

All the analyses presented are based on the notion of "mutual information," a measure of relatedness between different probability distributions. Let "range" be the number of words between two given words, x, and y plus one. For instance, a range of 1 will indicate that words x and y are separated by only 1 other word. We are asking whether our expectation of obtaining word y at a particular location is affected by the knowledge that we have word x in an earlier location. A measure of this is the mutual information (MI) between $P(x)$ and $P(y)$, the probabilities of obtaining word x and word y respectively. Mutual information indicates how much the uncertainty involved in expecting y is reduced by knowledge that we have x, and is given by $\sum_{x,y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$. For different ranges, $P(x,y)$ is the probability of having both words x and y, separated by a number of words equal to the range. By MI profile, we mean the way MI varies with increasing range. This we take to be an indicator of statistical structure in language.

Analyses

We investigated samples from seven different languages, all from the CD-ROM database of the European Corpus Initiative Multilingual Corpus 1 (ECI/MC1), distributed by the Association for Computational Linguistics. Table 1 shows the number of samples and average number of words in each language.

Table 1: "SE" is the standard error of the mean sample size, for each language.

language	mean words	SE	samples
Bulgarian	1468	256	4
Czechoslovakian	27591	860	29
Dutch	181407	33483	35
English	97272	11805	12
Estonian	19944	15043	2
French	166620	202	26
Gae	200239	.	1
German	129270	96532	8

In this research our objective is to examine the evidence that different languages display a similar MI profile, regardless of sample differences. Therefore, our approach has been to standardize the average mutual information values for each language, for the different ranges. Standardization converts a set of variables so that the mean of each variable is 0 and the standard deviation 1, so that essentially all the variables are on the same scale. This means that the same differences in each of the variables are now directly comparable. Figure 1 shows the results of this calculation. While there is considerable noise, one can see that the mutual information dependence "elbows" at about four items for all the languages. This we take to indicate that STM is indeed a relevant aspect of language learning. With future work, we aim to extend our analyses so as to more accurately examine MI profile similarities.

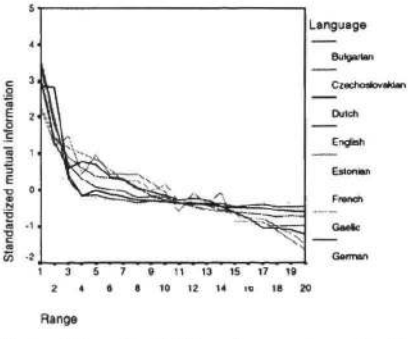

Figure 1: MI profile when MI values were standardized for the different languages.