Appendix 3

Equations Used for Statistical Measures and Example Calculation

N – corpus size,

f_A – number of occurrences of the keyword in the whole corpus (the size of the concordance),

f_B – number of occurrences of the collocate in the whole corpus,

f_{AB} – number of occurrences of the collocate in the concordance (number of co-occurrences)

T-Score $\frac{f_{AB} - \frac{f_A f_B}{N}}{\sqrt{f_{AB}}}$

MI-Score $\log_2 \frac{f_{AB} N}{f_A f_B}$

Church and Hanks, Word Association Norms, Mutual Information, and Lexicography, in Computational Linguistics, 16(1):22-29, 1990

MI^3-Score $\log_2 \frac{f_{AB}^3 N}{f_A f_B}$

Oakes, Statistics for Corpus Linguistics, 1998

log-likelihood $2 \cdot (xlx(f_{AB}) + xlx(f_A - f_{AB}) + xlx(f_B - f_{AB}) + xlx(N)$
$+ xlx(N + f_{AB} - f_A - f_B)$
$- xlx(f_A) - xlx(f_B) - xlx(N - f_A) - xlx(N - f_B))$
where $xlx(f)$ is $f \ln(f)$

Dunning, Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics 19:1 1993

Geometric Distance in the HCA Analysis

$$cos(\theta) = \frac{\vec{V}(terms_1) \cdot \vec{V}(terms_2))}{|\vec{V}(terms_1)||\vec{V}(terms_2)|}$$

Concrete example of calculating an MI score

To understand our process in greater detail, consider the following example. Suppose we want to measure the association between *xin* (心) and body terms (身, 體, 形) in all texts, using the MI score calculation. First we calculate the number of occurrences in all texts of *xin* (=8772). We then calculate the number of occurrences of our three body terms in

all texts (=11,579). Next, we need to continue to gather raw data about co-occurrences. Suppose we decide to use the word window of the sentence for this purpose. If so, then we will need to calculate the number of times within all texts *xin* co-occurs within the same sentence as any of our body terms (=946). Lastly, we need to know how many total characters are in the corpus (=5,742,539). We can now calculate the MI Score representing these relationships.

A=frequency of target word, i.e. body terms (11,579)
B=frequency of collocate, i.e. *xin* (8772)
C=characters in corpus (5,742,539), and
D=# sentence-level collocates of (*xin* & body terms) (946)

Xin x Body terms **MI Score:**
=Log2((D*C)/(A*B)
=Log2((946*5,742,539)/(11579*8772))
=Log2(5432441894/101570988)
=Log2(53.48)
=5.74