DR. GOOD said that the theory of clumps is bound to be largely experimental so perhaps it should be called "clumpology". The need for experiment arises because there seems to be no theoretical reason for choosing a unique definition of a clump. One has to begin by selecting a measure of relevance. If A and B are words or documents, etc., and if R_{AB} is a measure of relevance, such as P(B/A)/P(B) (for suitably defined probabilities), then other measures of relevance can be defined in terms of the matrix $R = \{R_{AB}\}$ (which need not be symmetrical). For example, we could first define $S_{AB} = R_{AB}/\Sigma_B R_{AB}$, in order that the row totals of the new relevance matrix should add up to 1. Then we could define a revised relevance matrix such as

$$I = \lambda \underline{S} + \mu \underline{S}^2 + \nu \underline{S}^3$$

where $\lambda + \mu + \nu = 1$, in order to-give some weight to nodes at distance 2 and 3 from any given node. Even when the measure of relevance is settled, there is still a vast choice of definitions of a clump. For example, apart from the ones mentioned in Miss Sparck-Jones' paper, a triple infinity of clumps can be defined thus: Call ζ a(α, β, γ)-clump if

$$\min_{A \in \zeta} \sum_{B \in \zeta} T_{AB}^{\alpha} > \beta N^{\gamma}$$

where α, β, γ are constants and N is the number of nodes in ζ

Finally, when clumps are defined, one can see if they break up into smaller clumps, and also if they fall into clumps of clumps, when they are themselves regarded as nodes in a weighted oriented linear graph (and so on). In this way the vocabulary should break up into a hierarchical classification. (Some words will deserve to be promoted to the status of a clump of the first or higher order). But this classification is not necessarily a tree, and not even a lattice, though it can be converted into one by means of a device of inserting imaginary categories, as suggested by the Cambridge group.

With luck one would find that the hierarchies of two different languages would be roughly isomorphic. This is the hope of those who are working on the thesaurus approach to mechanical translation. A danger in automatic clump-finding is that the two hierarchies may not be nearly isomorphic.

Accordingly, Dr. Good wanted to raise a question. Is it possible to produce an automatic method for ***encouraging*** the hierarchies of the two languages to be roughly isomorphic? The idea is to put the two vocabularies into a single weight oriented linear graph. There would be three kinds of relevance to define: those within the vocabularies of each language, and those across from words of one language to the other. The latter could be defined in terms of a statistical analysis of human translation, and perhaps also with the help of a probabilistic

dictionary. (Such dictionaries do not yet exist). By finding a suitable definition of a clump it may be possible to get both vocabularies to fall together into a hierarchy of bilingual clumps in a satisfactory manner for the application. It would be largely a matter of giving appropriate weight to the interlinguistic measures of relevance.

DR. SHERRY asked what was the distinction between clumps and rows.

MISS SPARCK-JONES replied that a clump is simply a set of rows subject to a similarity condition.

DR. SHERRY asked further if a better thesaurus was the aim of clumping.

MISS SPARCK-JONES answered that that was their aim.

J. McDANIEL