Statistical Discrimination of the Synonymy/Antonymy Relationship Between Words

P. A. W. LEWIS, P. B. BAXENDALE AND J. L. BENNETT

IBM San Jose Research Laboratory, San Jose, California

ABSTRACT. A basic hypothesis is stated about the contextual and co-occurrence properties of synonymous words. On the basis of this hypothesis, several statistics are derived for use in discriminating between pairs of words which are synonymous and pairs of words which are nonsynonymous. The discriminating power of these statistics is tested on a corpus consisting of titles of physics theses. The tests indicate that two of the derived statistics have relatively high discriminating power. The results are interpreted and the possibility of obtaining better discriminating power is discussed.

"Without speculation there is no good and original observation."

—Charles Darwin

1. *Introduction*

The index to a document collection specifies the subject content of the collection and at the same time serves as the device through which information about the collection is sought. An index, therefore, must bridge the semantic gap which exists between the language of an author and the language of a searcher. Whether this function is imposed on author, indexer or searcher is immaterial: The need to span the semantic gap cannot be circumvented. To this end, specific semantic relationships are customarily imposed on index units and are made explicit as "see" or "see also" references or as some form of thesaurus.

An interest in automatic indexing procedures has motivated several attempts to specify automatically the semantic associations between units of an index vocabulary [1–5]. In these instances, language is viewed essentially as a statistical phenomenon, and the aim is to define quantitatively a measure of association between words (or alternatively the descriptors of an index) using some function of the frequency with which words occur or co-occur within a document collection.

The customary result from such investigations has been a list of ranked associated words. Regardless of the analytical method used or the different features of text on which the quantification is based, all investigations must include consideration of the same basic question: What useful semantic interpretation is made plain by the derived statistical associations? Further, can a reliable criterion of "useful associations" be delimited, in terms of a threshold, by a given statistical measure? It is not enough to say words are "associated"—in the extreme, all words of a text are associated in that they are all drawn from English. If we are interested in using lists of associated words as part of an index, we must be able to specify the type of association indicated by the measure. Such a delineation has not been possible in the work reported in the literature to date.

In this paper we deal with the derivation of a statistical measure of association

but one which is capable of separating word pairs expressing "equivalent" or "contrary" meanings from word pairs expressing other semantic relationships such as "inclusion" or "property." The starting point for this discrimination was a hypothesis as to how the synonymy association is inferred from text. This basic hypothesis is applicable either to descriptors occurring as indices to a document or to individual words occurring in a set of sentences.

The development of the paper follows the order in which the research was done. We believe that this format enables a clear exposition of the rationale and development leading to the results—an important need in this field. The outline is (1) a statement of the hypothesis, (2) the design of statistical measures of synonymy on the basis of the hypothesis, (3) precategorization of selected word pairs from a data base as synonyms or nonsynonyms, (4) a test of the statistical measures of synonymy using the precategorization as a standard, and (5) modification of the measures in the light of our empirical findings. The paper is concluded with an interpretation of the results.

2. *The Basic Hypothesis and Its Statistical Interpretation*

The basic hypothesis on which the statistical measures of synonymy described in this paper are based [6] is the following: *If two words (descriptors) are synonymous, then they very infrequently, or never, co-occur as words in the same sentence (as descriptors for the same document), but in their separate occurrences they tend to have similar contexts.*

This is a broad and intuitively inferred linguistic statement, but it does not encompass all the linguistic aspects of synonymy. If only single words from text are considered, then a resulting measure will not be able to indicate that "cybernetics" is a synonym for "automata studies." Since this type of paraphrase expresses synonymy, our initial model is defective to this extent. Furthermore, synonyms do co-occur in such sentences as "Thiamine, popularly referred to as vitamin B_1, ..." But we believe that these co-occurrences are relatively infrequent, as stated in the hypothesis.

It also should be observed that the basic hypothesis could apply as well to pairs of words which are antonymous. Measures which we derive therefore do not provide any discrimination between the relationships of synonymy and antonymy.

With these qualifications in mind, we are now ready to examine the statistical implications of the basic statement. Note that there is an inherent statistical element in the hypothesis. The statement says: Given that a pair of words is synonymous, the probability of their co-occurring is not equal to zero but is small. By small we mean that the probability is small relative to the probability of co-occurrence under a hypothesis that the two words are not synonymous.

The hypothesis also states: In their separate occurrences synonyms tend to have similar contexts. A rough statistical way of putting this might be the following. Consider the words which have occurred with either one or the other or both of the two words to be a set. Then we have a fraction of the set which contains words occurring with both of the given words. If this situation is now considered for many pairs of words, a distribution of fractions is obtained. For those pairs of words which are synonymous, it is expected that this distribution will be concentrated on

high values, say values greater than one-half, and otherwise it will be concentrated on low values.

Roughly, then, we are saying that the quantities referred to in the statement of the hypothesis have distributions, and moreover distributions which depend on the conditioning statement of synonymy or nonsynonymy. These ideas are made more precise in what follows.

There is another quasistatistical element in the hypothesis in that it is not in general possible to say in a dichotomous way that a pair of words is synonymous or not synonymous. To the extent that this lack of dichotomy holds, the problem of discrimination becomes a type of statistical ranking and selection procedure. We assume, for the purposes of this experiment, that any given pair of words *can* be categorized as synonymous or nonsynonymous. We are hence dealing with a relatively standard statistical classification problem [7]. In this the idea is to find a statistic or measure based on the attributes of a pair of words that will allow a decision to be made as to whether the given pair of words is synonymous or not, and which in some way minimizes the two types of error probabilities: (1) α, the probability of classifying a synonymous pair as not synonymous, and (2) β, the probability of classifying a nonsynonymous pair as synonymous.

The statistical methodology which has been developed for classification problems can then be used in this minimization problem. It should be noted, however, that because of the inherent statistical nature of the hypothesis there will always be some (hopefully small) probability of misclassification.

Statistical classification theory shows how to find not one but a set of admissible decision procedures which correspond to a convex curve in the (α, β)-plane, and which are such that for a decision procedure corresponding to a particular (α, β) point on this curve, any attempt to decrease $\alpha(\beta)$ leads to an increase in $\beta(\alpha)$. One way of choosing a particular decision procedure is to minimize the overall probability of misclassification. This requires a knowledge of the prior probabilities of a pair being synonymous or not synonymous. This overall measure needs to be used with care; moreover, the prior probabilities are usually not known exactly. However, if we know their relative magnitudes, we can use the overall measure to determine the relative sizes of α and β which we can tolerate. In particular it is known that of all the possible pairs in a given corpus, the overwhelming majority would be nonsynonymous. *Consequently, in the discrimination of nonsynonymous and synonymous pairs in this situation, it is possible to infer that the error probability β must be kept much smaller than α in order that the set of pairs selected as synonymous does actually contain a high percentage of synonymous pairs.*

3. *Definitions Used in the Design of the Statistical Measures*

Some definitions are now given which enable us to express more precisely what is meant by the "context" of a given word. The definitions are given within the framework of a corpus which consists of a set of sentences. These definitions can easily be translated to the case where the corpus is a set of descriptors which have been assigned to the documents in a collection. In the empirical verification of the model, we actually used a corpus consisting of a set of titles.

Let n be the number of sentences in the corpus under consideration and let X

denote the set of distinct words which occur in the n sentences:

$$X = \{x_1, \cdots, x_{\mu(X)}\},$$

where $\mu(X)$ is the number of elements in X. In general capital letters are used to denote sets and lower case letters to denote the number of elements in the set.

Now let K be the set of indices of X:

$$K = \{1, \cdots, \mu(X)\}.$$

Let the letters i, j, k stand for variables running through the set K, so that x_i or x_j or x_k denotes the function from K to X, i.e., they can denote any of the members of X. Furthermore, let the letters a, b, c stand for fixed elements in K, so that x_a, x_b and x_c denote specific members of X.

Let n_i denote the number of sentences in the collection in which x_i occurs at least once. This might be called the sample size for x_i and is often called the frequency of occurrence of x_i. Similarly let n_{ij} denote the number of sentences in the collection in which both x_i and x_j occur at least once, or the sample size for the doublet (x_i, x_j).[1]

Now by the context of x_i is meant the set consisting of those words in X which have occurred in at least one sentence with x_i :

$$\begin{aligned} D_i &= \{x_k \mid n_{ik} \neq 0, \quad k \neq i\} \\ &= \{x_k \mid x_k \text{ co-occurs in at least one sentence with } x_i\}. \end{aligned}$$

The size of this set is denoted by

$$d_i = \mu(D_i).$$

By the *mutual context* of a pair of words (x_i, x_j) is meant the set consisting of those words in X which have occurred in at least one sentence with x_i and in at least one sentence with x_j :

$$\begin{aligned} D_{ij} = D_{ji} &= \{x_k \mid n_{ik} \neq 0; \quad n_{jk} \neq 0; \quad i \neq j \neq k\} \\ &= D_i \cap D_j \\ &= \text{the } \textit{mutual context} \text{ of } x_i \text{ and } x_j. \end{aligned}$$

The size of D_{ij} is denoted by d_{ij}. The sets D_i, D_j and D_{ij} and their mutual relationships are shown in Figure 1.

The set D_{ij} can be further broken down in a way which is informative as to its nature and which facilitates its use. Let

$$\begin{aligned} J_{ij} &= \{x_k \mid n_{ik} \neq 0; \quad n_{jk} \neq 0; \quad n_{ijk} \neq 0; \quad i \neq j \neq k\} \\ &= \text{the } \textit{joint context} \text{ of } x_i \text{ and } x_j, \end{aligned}$$

and

$$\begin{aligned} P_{ij} &= \{x_k \mid n_{ik} \neq 0; \quad n_{jk} \neq 0; \quad n_{ijk} = 0; \quad i \neq j \neq k\} \\ &= \text{the } \textit{pairwise context} \text{ of } x_i \text{ and } x_j. \end{aligned}$$

[1] Some of these relations have been expressed in matrix notation by previous workers in the field.

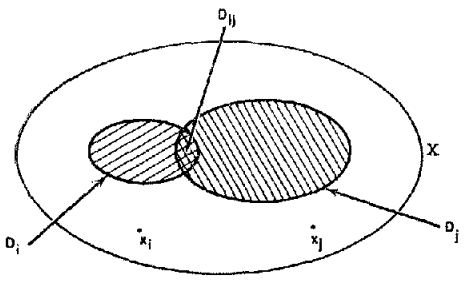

FIG. 1. Contexts of x_i and x_j, and mutual context of x_i and x_j

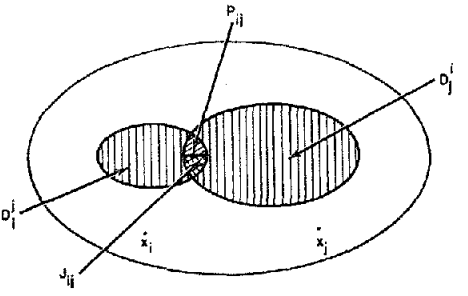

FIG. 2. Difference sets and the pairwise and joint contexts of x_i and x_j

The added condition is on the frequency of occurrence of triples, n_{ijk}, which is just the number of sentences in the collection in which x_i and x_j and x_k have all appeared. The size of J_{ij}, denoted $\mu(J_{ij})$, is the number of terms x_k which have occurred at least once in a sentence in which *both* x_i and x_j *also* occur. The size of P_{ij}, $\mu(P_{ij})$, is the number of terms x_k which have not occurred in any sentence with both x_i and x_j but which have occurred in some sentence with x_i and in some other sentence with x_j.

From the definition of J_{ij} and P_{ij} it is clear that

$$P_{ij} \cup J_{ij} = D_{ij} \quad \text{and} \quad P_{ij} \cap J_{ij} = \phi, \quad \text{the empty set,}$$

so that $\mu(P_{ij}) + \mu(J_{ij}) = p_{ij} + j_{ij} = d_{ij}$. These sets and their relationships are shown in Figure 2. Note that

$$n_{ij} = 0 \Rightarrow n_{ijk} = 0$$

and

$$\Rightarrow J_{ij} = \phi,$$

but that

$$n_{ij} = 0 \nRightarrow P_{ij} = \phi.$$

Furthermore, if $n_{ij} = 0$ and consequently J_{ij} is empty, then $P_{ij} = D_{ij}$. (Note also that if $n_{ij} \neq 0$, then it is not possible to split D_{ij} into P_{ij} and J_{ij} if only doublet information is available.) In fact n_{ij} and consequently n_{ijk} may be zero but $\mu(P_{ij}) = p_{ij}$ may be relatively large. *It is precisely this situation which is exploited in what follows in designing a measure of synonymy of two terms x_i and x_j.*

Difference sets are here defined as follows:

$$D_i^j = D_i - D_{ij}.$$

The size of D_i^j, denoted by d_i^j, is given as

$$d_i^j = d_i - d_{ij}.$$

These difference sets are shown in Figure 2. Other definitions are given as needed in what follows.

4. Quantification of the Hypothesis—Derivation of Measures

In terms of the above definitions the basic hypothesis may be paraphrased as the following: If a pair of words, x_a and x_b, is synonymous, then n_{ab} will be small, perhaps zero, and the pairwise context, P_{ab}, will be large relative to both D_a and D_b. The condition of n_{ab} being small also implies that P_{ab} will be large relative to J_{ab}.

The problem now is to determine a measure or statistic, a function of quantities such as d_a, d_b and p_{ab}, which allows discrimination between synonymous and nonsynonymous pairs. Not having parametric hypotheses and distributions, as is the case in many statistical classification problems, this *determination must be done on intuitive grounds and then verified and possibly modified empirically.* The basic intuitive idea is of course the hypothesis stated above, plus the following points which indicate the type of quantification we need.

(i) Referring to Figure 2, we want a strong indication of synonymy if P_{ab}, D_a and D_b essentially overlap. This is a reiteration of the basic hypothesis that x_a and x_b will have similar contexts if they are synonymous. It can be considered also as a requirement to normalize the size of P_{ab} so as to define what is meant by "large" mutual context.

(ii) If as in (i) P_{ab}, D_a and D_b essentially overlap, then the value of the measure should increase and the indication of synonymy should increase as d_a (and therefore p_{ab} and d_b) increases.

This latter is an extremely important point in the design of the measure. In effect, even when there is a fixed corpus size (i.e., the n sentences under consideration), every pair of words which can be formed from the collection X has its own sample size. This sample size is $(n_i + n_j)$ if $n_{ij} = 0$, and approximately $(n_i + n_j)$ if n_{ij} is small. Now in the attempt to determine synonymy it is absolutely necessary to deal with this multi-sample-size problem. Quite clearly, if the sample size for a pair is the minimum of two, there is virtually no evidence on which to base a decision. Consequently, any measure which assigns as much evidence for or against synonymy on the basis of this minimum sample size as it does, for instance, for a pair whose sample size is $n/2$, will be inadequate.

The classification problem is then a sequential one, even for a fixed corpus size, and as in any sequential decision problem, a region in the decision space is required in which the decision reached is that more evidence is needed before a classification is assigned to the pair of words. In order to do this, without requiring different cutoff points for different sample sizes, the measures are designed to be roughly modal about zero. In other words, "don't know" decisions and decisions of nonsynonymy are based on negative values of the measure. It is to be expected then that the measure for any pair will be slightly negative for very small sample sizes and then will go strongly positive or negative as the sample size increases, depending on whether the pair is synonymous or not. This behavior will, of course, be subject to random fluctuations in the measures.

It should also be noted that from empirical evidence obtained, d_i increases roughly linearly as n_i, so that the size of the contextual measures may be used to represent sample size. One would not, however, anticipate that this initial linear relationship would hold indefinitely as corpus size increased.

(iii) If D_a and D_{ab} essentially overlap, so that d_a and p_{ab} are approximately

equal, but $p_{ab} \ll d_b$, then the chances are that either the mutual context of the two words is only incidental or that x_b is a broad term which (semantically) includes x_a. The case when $d_b \approx p_{ab}$ but $p_{ab} \ll d_a$ is similar.

As with the basic hypothesis, this point is purely speculative and is subject to empirical verification.

(iv) The quantities d_a, d_b and p_{ab} only indicate that context has been shared, not how frequently it has been shared. For instance, for $x_c \in D_a$ we might have $n_{ac} = 1$ but for $x_d \in D_a$ we might have $n_{ad} = 100$. These contextual measures therefore could not distinguish the following importantly different cases:

(a) For $x_k \in P_{ab}$, all the n_{ak}'s and n_{bk}'s are *large*;
for $x_k \in D_a^b$, all the n_{ak}'s are *small*;
for $x_k \in D_b^a$, all the n_{bk}'s are *small*.

(b) D_a^b, D_b^a and P_{ab} are the same as in (a), but
for $x_k \in P_{ab}$, all the n_{ak}'s and n_{bk}'s are *small*;
for $x_k \in D_a^b$, all the n_{ak}'s are *large*;
for $x_k \in D_b^a$, all the n_{bk}'s are *large*.

A stronger indication of synonymy would naturally be desired in case (a), but measures which use only the contextual quantifiers d_a, d_b, p_{ab} could not, by their very nature, provide this.

We now proceed to the design of a measure of synonymy. Such a measure or statistic will be a function, G, of all or some of the frequency and contextual parameters associated with a given pair of words. *For simplicity the case where* $n_{ab} = 0$ *and consequently* $p_{ab} = d_{ab}$ *is considered.* If a measure of synonymy based on similarity of context and infrequent co-occurrence can be developed, it should hold for the special case of synonyms which do not co-occur. Modification of the measure to the case where $n_{ab} \neq 0$ is indicated in Section 8.

(1) The contextual measure $p_{ab} = d_{ab}$ is one possible function, but it meets none of the above four requirements.

(2) By using the normalized measure

$$G_1 = \frac{p_{ab}}{d_a + d_b - p_{ab}} = \frac{p_{ab}}{d_b^a + d_a^b + p_{ab}},$$

we satisfy requirements (i) and (iii) above, but not requirement (ii) because the effect of the size of n_a and n_b is masked out in the ratio. This measure is bounded by 0 and 1, which can be an advantage, but it is nonlinear, and this is undesirable. In consequence, this measure was not tested in the experiment.

A corresponding normalized but linear measure is

$$G_2 = p_{ab} - d_a^b - d_b^a = 3p_{ab} - d_a - d_b.$$

This meets the requirements of (i), (ii) and (iii), since for given x_a and x_t with pairwise context p_{ab} the measure G_2 is maximum and equal to p_{ab} when $p_{ab} = d_a = d_b$. This is condition (i) above. Condition (ii) is satisfied because when there is a maximum overlap of p_{ab} with d_a and d_b, then G_2 increases linearly with p_{ab}. In a similar way it can be seen that condition (iii) holds.

Another way of writing G_2 which may help to clarify its nature is the following:

$$G_2 = \sum_{k=1,\ k\neq a,\ k\neq b}^{\mu(X)} S_k, \qquad \text{where} \quad S_k = \begin{cases} 1 & \text{if} \quad n_{ak} > 0, \quad n_{bk} > 0; \\ -1 & \text{if} \quad n_{ak} > 0, \quad n_{bk} = 0; \\ -1 & \text{if} \quad n_{ak} = 0, \quad n_{bk} > 0; \\ 0 & \text{if} \quad n_{ak} = 0, \quad n_{bk} = 0. \end{cases}$$

(3) The measure G_2 does not meet condition (iv), but it can be modified as follows in order to do so. We define G_3 as

$$G_3 = \sum_{k=1,\ k\neq a,\ k\neq b}^{\mu(X)} H_k, \qquad \text{where} \quad H_k = \begin{cases} \min(n_{ak}, n_{bk}) & \text{if} \quad n_{ak} > 0, \quad n_{bk} > 0; \\ -n_{ak} & \text{if} \quad n_{ak} > 0, \quad n_{bk} = 0; \\ -n_{bk} & \text{if} \quad n_{ak} = 0, \quad n_{bk} > 0; \\ 0 & \text{if} \quad n_{ak} = 0, \quad n_{bk} = 0. \end{cases}$$

This is a straightforward generalization of G_2 to take care of condition (iv); i.e.' to weight the measure by the number of times words in D_a, P_{ab} and D_b have occurred with x_a, x_a and x_b, and x_b, respectively. A reason for using $\min(n_{ak}, n_{bk})$ in the case where $n_{ak} > 0$ and $n_{bk} > 0$ is a feeling that other possibilities such as the average or maximum of n_{ak} and n_{bk} would weight the measures too strongly. In particular, this would occur when a word x_c goes into P_{ab} from D_a or D_b as the sample size increases. The final determination of the relative utility of these weightings must be made empirically, but for an initial determination of any superiority of G_3 over G_2 the minimum will serve as well as the other possibilities.

Another problem which must be settled empirically is that of modifying the measure so that it will be modal about 0. For instance, if d_a and d_b are approximately equal, then G_2 will be approximately 0 if $d_a \approx d_b \approx \frac{3}{2}\ p_{ab}$. We are therefore saying that we expect, approximately, more than two-thirds overlap if two terms are synonymous. It may well be that in most corpora there is, on the average, too much extraneous context for any given word for the two-thirds overlap to be obtained. The measures will then have to be modified, for instance by using instead of the G_2 defined above,

$$G_2 = Kp_{ab} - d_a^{\,b} - d_b^{\,a},$$

where the constant K is to be determined empirically.

The measures G_2 and G_3 have been formulated on the basis of intuitive ideas without reference to a particular body of data. Consequently, if they can discriminate the synonymy relationship in one corpus, then they should apply to any corpus. This would not be the case if the measures lacked a conceptual foundation and were designed solely on an examination of a particular corpus.

5. *Empirical Investigation of the Measures*

General Observations. The general plan for comparing the performance of measures is as follows. First it is necessary to categorize, qualitatively and dichotomously, as "synonymous" or "not synonymous" pairs of words taken from the test data base. This sample of pairs of words (which do not co-occur) is used as a standard to avoid the tendency to classify pairs after the fact and on the basis of

measures themselves. A measure value is then calculated for each of these pairs, and the pairs are listed in descending order on the value. A rough evaluation of the α and β misclassification errors associated with a cutoff point is then made. In this way, G_2 and G_3 will be tested, and we will expect to evaluate other G's—modifications of G_2 and G_3—to see if performance can be improved. While points (i)–(iv) outlined in Section 4 continue to serve as guidelines, further indicators of desired behavior are developed as we examine the test data in what follows.

PRECATEGORIZATION OF WORD PAIRS. The preselection and intellectual classification of pairs of words from the sample was performed as follows. It was recognized that there would be difficulty in obtaining a consensus as to whether a given pair of words expressed a given relationship, so definitions were formulated which were as operational as we could make them. Categorization then took place according to the following definitions:

(a) One word was considered synonymous with another word if it met *any one* of the following criteria:

(i) If meaning was preserved within some syntactically appropriate sentence frame when one word of a pair was substituted for the other, e.g.,

$$\text{Light from a constant} \left\{ \begin{matrix} \text{bright} \\ \text{intense} \end{matrix} \right\} \text{source} \cdots .$$

(ii) If one word of the pair was the plural form of the other, e.g.,

spectrum : spectra.

(iii) If a word of the pair was designated a "see" reference in the *International Dictionary of Physics* or as a "synonym" in *Webster's International Dictionary*, e.g.,

conductivity : see : resistivity,

conversion : syn : transformation.

(iv) If there was a consensus of interpretation of the definition given in the *International Dictionary of Physics*, e.g.,

decay : disintegration.

The definition of radioactive decay in the dictionary was "radioactive disintegration."

This category of synonymous words is referred to by an A in the test sample listed in Table 1.

(b) *Antonymous relationships* were determined on the basis of *either* of two criteria:

(i) A word whose opposite meaning is formed by adding a prefix with negative connotation such as anti-; un-; non-; dis-; etc. For example,

stable : unstable.

(ii) A word given as an antonym in *Webster's International Dictionary* or by definition in the *International Dictionary of Physics*, e.g.,

bound : free.

TABLE 1. THE TEST SAMPLE OF 120 WORD PAIRS* WITH PRECATEGORIZED ASSOCIATION: A, SYNONYM; B, ANTONYM; C, ASPECT OR PROPERTY; D, INCLUSION; E, ARBITRARY

Pair Number	Word-pair	Category	Pair Number	Word-pair	Category
1	absolute : relative	B	61	emission : scattering	E
2	absorption : adsorption	D	62	emulsion : mixture	D
3	absorption : diffusion	B	63	emulsions : mixtures	D
4	acceleration : velocity	C	65	energy : transition	E
5	acoustics : sound	A	66	excitation : excited	D
6	afterglow : glow	D	67	film : films	A
7	alloy : alloys	A	68	film : layer	D
8	alloy : mixture	D	69	flow : flux	A
9	altitude : altitudes	A	70	fluid : fluids	A
10	amplitude : attenuation	C	71	fluid : liquid	D
11	amplitude : energy	E	72	fluid : solid	B
12	angle : angles	A	73	fluids : solids	B
13	anisotropic : anisotropy	A	74	fluorescence : ionization	E
14	annihilation : disintegration	D	75	fluoride : fluorine	D
15	anode : cathode	D	76	force : forces	A
16	atom : atoms	A	77	force : pressure	C
17	band : bands	A	78	formation : generation	A
18	band : line	D	79	frequencies : wavelength	D
19	bands : lines	D	80	frequencies : wavelengths	D
20	beams : rays	A	81	frequency : wavelength	D
21	binary : divalent	C	82	friction : heat	C
22	boundary : range	D	83	generation : growth	A
23	bound : free	B	84	graphs : plots	D
24	bright : intense	A	85	halide : halogen	D
25	bromide : bromine	D	86	heat : loss	C
26	charge : coulomb	D	87	heat : resistance	C
27	circuit : circuits	A	88	heat : thermal	D
28	cloud : condensation	D	89	infrared : ultraviolet	B
29	collision : collisions	A	90	law : principle	A
30	collision : impact	A	91	level : transition	E
31	conduction : conductivity	D	92	lifetime : lifetimes	A
32	conductivity : resistivity	A	93	lifetime : tellurium	E
33	constant : fixed	A	94	lifetimes : neutrino	E
34	conversion : transformation	A	95	liquid : liquids	A
35	correlation : relation	D	96	mass : momentum	C
36	counter : counters	A	97	mass : weight	C
37	cross-section : lifetime	E	98	measurement : measure	A
38	decay : disintegration	A	99	measurement : measurements	A
39	decay : emission	C	100	solid : wavelength	E
40	decay : lifetime	D	101	neutrino : neutron	E
41	decay : lifetimes	D	102	noise : sound	A
42	decay : scattering	E	103	perturbation : relaxation	E
43	density : mass	C	104	pressure : shock	D
44	deuterium : tellurium	E	105	process : technique	A
45	deuterium : zinc	E	106	propagation : transmission	A
46	device : mechanism	A	107	soft : weak	E
47	diffraction : disintegration	E	108	solid : solids	A
48	diffraction : dispersion	E	109	sonic : supersonic	E
49	dioxide : monoxide	D	110	spectra : spectrum	A
50	discharge : discharges	A	111	stable : unstable	B
51	discharge : disintegration	E	112	supersonic : ultrasonic	E
52	disc : disk	A	113	voltage : volt	D
53	disintegration : lifetime	D	114	conduction : level	E
54	disintegration : particle	C	115	conductivity : solid	E
55	disintegration : photodisintegration	D	116	density : spectra	E
56	distribution : distributions	A	117	frequency : transition	E
57	doublet : singlet	E	118	absolute : zinc	E
58	dynamic : static	B	119	band : density	E
59	emission : fluorescence	D	120	emulsion : fluid	E
60	emission : frequency	C	121	collision : zinc	E

* The number 64 was not assigned to a pair.

TABLE 2. WORDS DELETED FROM THE CORPUS DURING PROCESSING

the	are	what	said
of	we	would	may
and	his	who	about
to	but	when	over
a	they	them	some
in	or	her	these
that	which	any	before
it	will	more	could
is	from	now	such
I	had	its	upon
for	has	up	every
be	our	do	how
was	an	out	come
as	been	can	us
with	their	than	shall
on	there	only	should
have	were	made	then
by	so	other	like
not	my	into	well
at	if	must	say
this			

The category of antonymous words is designated by a B in the test sample listed in Table 1.

(c) *Neither* synonym nor antonym. Two other common types of semantic associations frequently found in a thesaurus are those expressing an aspect or property of a concept and that of inclusion. These, together with "arbitrary" pairs showing no obvious relationship, comprised the nonsynonymous categories designated as C, D and E, respectively, in Table 1. The actual criteria, which are not as operational as with the synonym and antonym categories, were the following:

(i) One word was considered an aspect or property of the other when it expressed a resultant or common feature of the concept expressed by the first, e.g.,

"heat" is an aspect resulting from "friction."

(ii) One word of the pair was said to be included in another when it was semantically subsumed under the other, e.g.,

"alloy" is subsumed under "mixture."

(iii) One word of the pair had no discernible relationship to the other—what was referred to above as "arbitrary," e.g.,

collision : zinc.

THE CORPUS. The corpus selected for this investigation was 6000 titles taken from Marckworth's *Dissertations in Physics* [8]. This corpus was used because the subject-content was relatively homogeneous (nuclear physics) and because titles were already available in computer-readable form. Computing simplicity also dictated the use of single words, even though a set of descriptors (word strings) for a collection of documents might have been preferable. However, it was felt that if the measures could adequately discriminate the synonyms from the nonsynonyms using single words, they would work as well, if not better, on a corpus of descriptors.

Words whose grammatical function predominates over the semantic function, such as "the" and "an," were eliminated from the titles. A list of such words,

deleted from the corpus during processing, was compiled from Dewey's data on the frequency of occurrence of words in English [9]. The list is given in Table 2.

THE TEST SAMPLE. Using the definitions given under "Precategorization of Word Pairs," the authors compiled a list of 120 word pairs from the corpus for which they could agree on class assignments. These word pairs are given in Table 1. For the experiments reported here, where the aim has been to design a measure which is adequate to separate the synonym/antonym relationship from the other three types of relationship, the first two were put together in one category and in subsequent figures are marked by an asterisk. The aspect, inclusion, and arbitrary relationships comprise the nonsynonymous category and are left unmarked in subsequent figures.

Of the 120 word pairs, 45 belonged to the first category. This virtually exhausted the pairs with $n_{ij} = 0$ for which a consensus could be obtained for inclusion in this synonym/antonym category. Nonsynonymous pairs represent no problem, and these were chosen at random so as to obtain a representative sample of individual word-pair sample sizes. The number was set at 75 so as to give a test sample which was big enough to give reasonable error probability estimates and at the same time be computationally manageable.

EXAMINATION OF THE DATA IN TERMS OF THE MEASURES. The measures G_2 and G_3 were computed for the test sample of 120 word pairs for various values of the constant K, and for each run they were listed in descending order on the measures. Two examples, one each for G_2 and G_3, are shown in Figures 3 and 4. The center columns show the pair number, an asterisk or no asterisk according to whether the pair was precategorized as synonymous or nonsynonymous and then two columns of numbers which are the measure values for the two categories. The two categories are split for graphical display purposes; the pluses and minuses on the right of these numbers are explained later.

In this way one can make a rough appraisal of the value of K which gives the best discrimination based on the particular measures. The two figures show the measure values resulting from the optimal values of K: $K = 8$ for G_2 and $K = 5$ for G_3. Variation of K by ± 1 does not make much difference.

It is immediately obvious that G_3 is a better classification statistic than G_2, thereby bearing out point (iv) in Section 4. If we accept all pairs with measure values above zero as synonyms, it is seen that we have good discrimination against nonsynonyms, but relatively poor discrimination of synonyms.

MODIFICATIONS TO G_2 AND G_3. Next G_3 was modified to form G_4 and G_5, which used the average and maximum, respectively, of (n_{ak}, n_{bk}) for words in P_{ij} instead of the minimum. The use of the average improved the discrimination slightly and the use of the maximum produced no improvement over the average.

The size of the constant K in G_3 and its probable dependence on the particular corpus was felt to be a drawback of the measure, and another and seemingly more natural way of reducing the influence of extraneous context was tried. This was to eliminate all words from D_a^b and D_b^a for which n_{ak} or n_{bk} had the value 1, giving reduced contexts for x_a and x_b. These are denoted by R_a^b and R_b^a and their sizes by r_a^b and r_b^a. A measure G_6, corresponding to G_2, but using the reduced contexts,

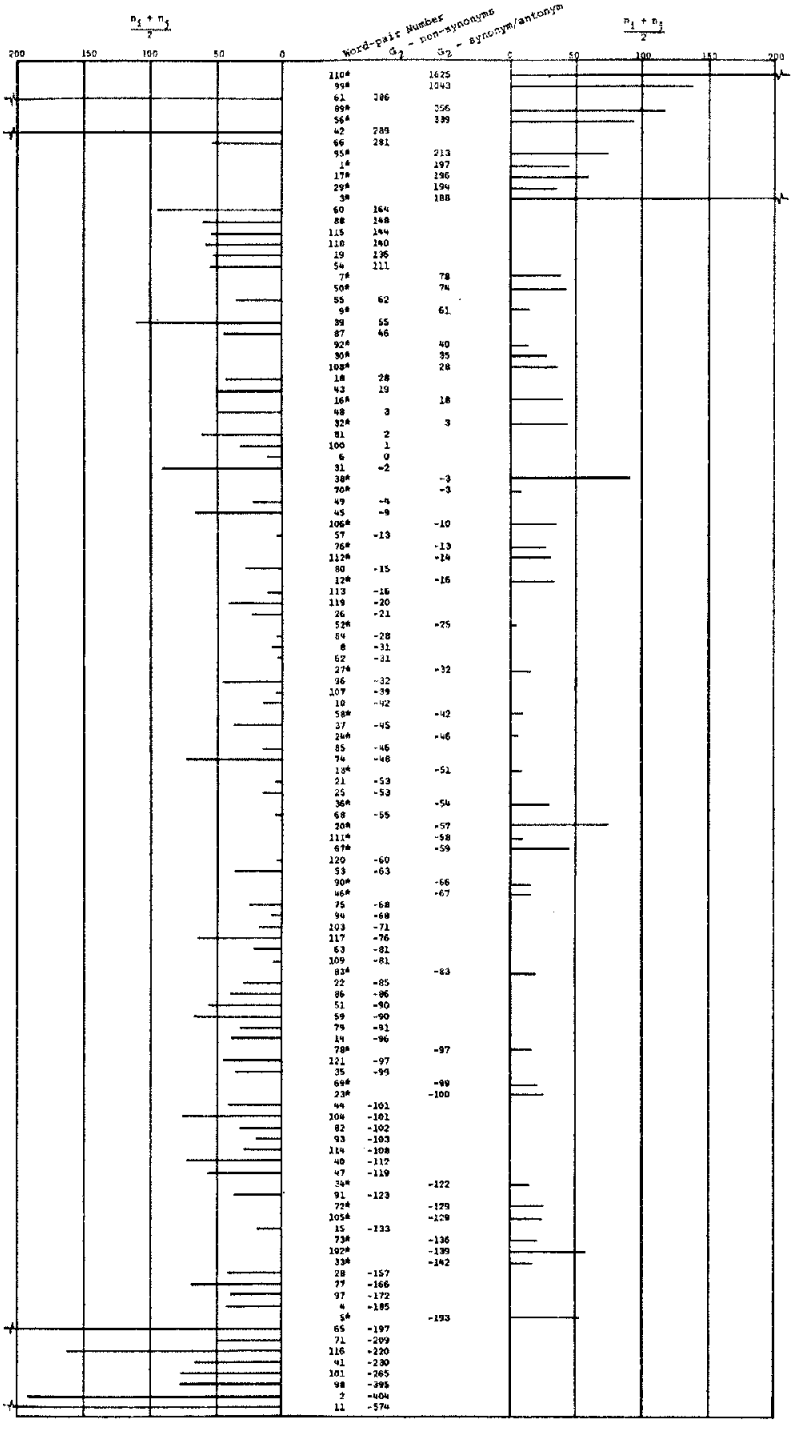

FIG. 3. Word pairs listed in order on G_2 with $K = 8$. The center columns give the word-pair number and the value of G_2. The average frequency of each word pair is shown as a horizontal bar.

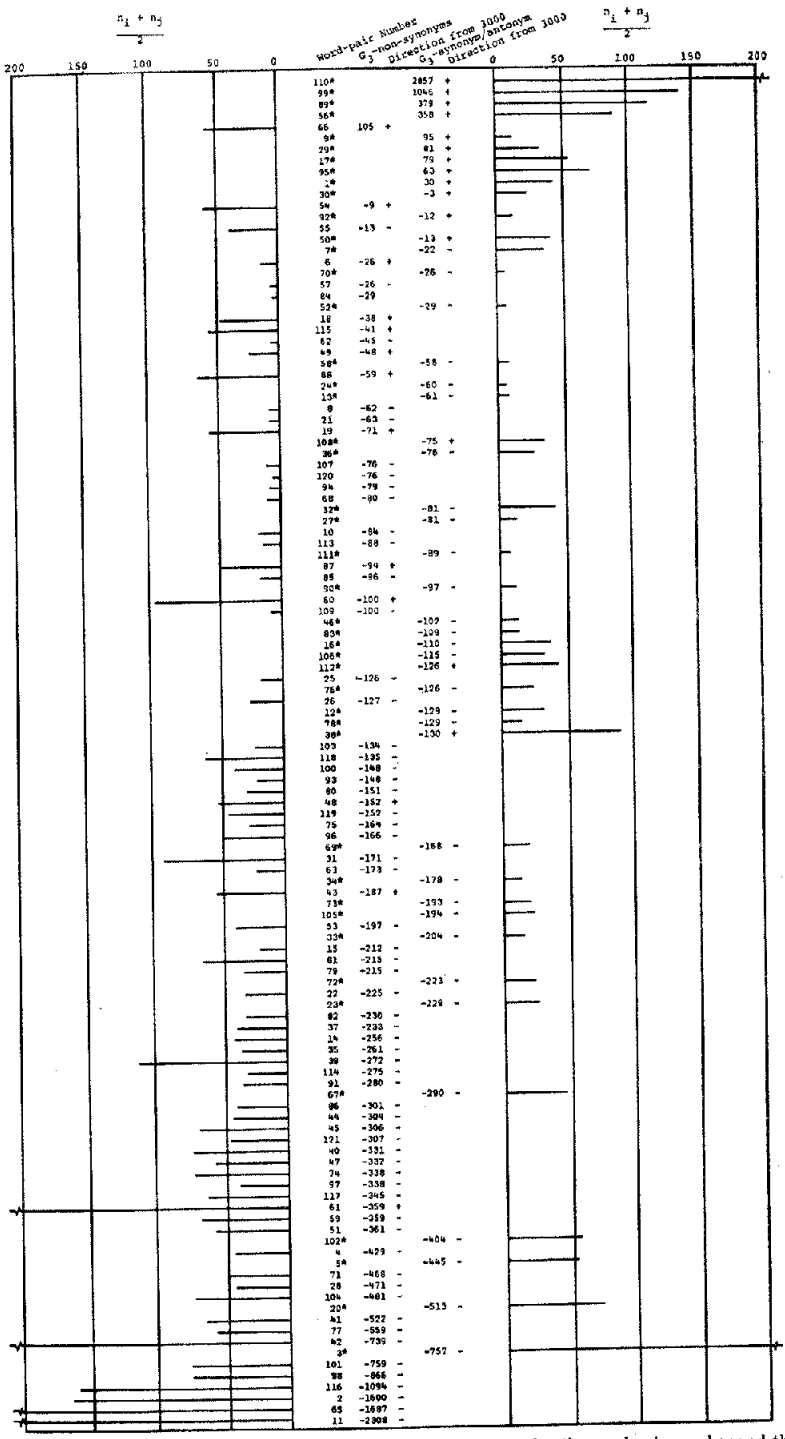

FIG. 4. Word pairs listed in order on G_3 with $K = 5$. The center columns give the word-pair number and the value of G_3 followed by an indication of the increase or decrease of that value when compared with the sample of 3000 titles. The average frequency for each word pair is shown as a horizontal bar.

was tried. Similarly a measure G_7, corresponding to G_3, again using reduced contexts, was tried. As with G_2 and G_3, the measure G_7 was found to be more discriminating than the measure G_6.

The measure G_7 can be written as

$$G_7 = \sum_{k=1,\ k\neq a,\neq b}^{\mu(X)} H_k', \qquad \text{where} \quad H_k' = \begin{cases} K_{\text{avg}}(n_{ak}, n_{bk}) & \text{if} \quad n_{ak} > 0, \quad n_{bk} > 0; \\ -n_{ak} & \text{if} \quad n_{ak} > 1, \quad n_{bk} = 0; \\ -n_{bk} & \text{if} \quad n_{ak} = 0, \quad n_{bk} > 1; \\ 0 & \text{otherwise.} \end{cases}$$

The ordering of the test pairs on the measure G_7 is shown in Figure 5. There is a slight improvement over the discrimination produced by G_3; the optimum value of the constant K, however, has now been reduced to the value 1. Values of K from 0.8 to 1.2 give roughly equivalent results. The measures G_8 and G_9 used, respectively, the maximum and minimum of (n_{ak}, n_{bk}) for words in P_{ij}. They appeared not to perform as well as G_7 and were not considered further.

Before examining the test results achieved with the measures G_3 and G_7 in more detail, we consider the error probabilities and overall probability of misclassification for the scheme based on G_7.

Figure 6 shows the estimated error probability curve with the cutoff value of G_7 as parameter. For instance, the point marked -20 means that the classification rule used was to accept as synonyms all pairs with G_7 values greater than -20. The error probability $\hat{\beta}$ is then estimated as the ratio of the number of pairs which were precategorized as being nonsynonymous and which had G_7 values above -20, to the number, 75, of nonsynonymous pairs. The error probability $\hat{\alpha}$ is estimated similarly. An error probability curve which is a straight line connecting the extreme points (1, 0) and (0, 1) would represent purely arbitrary classification or no discrimination for the measure at all.

Figure 7 shows, for the same cutoff values, the estimated overall probability of misclassification for two cases:

(a) prob(not synonymous) = 0.99; prob(synonymous) = 0.01;

(b) prob(not synonymous) = 0.50; prob(synonymous) = 0.50.

The overall probability is estimated as

$\hat{\alpha} \times$ prob(synonymous) $+ \hat{\beta} \times$ prob(not synonymous).

Note that in case (a) an overall probability of misclassification of 0.01 can be obtained by rejecting synonymy for all pairs regardless of the value of the measure, illustrating that this criterion of discrimination must be used with care. The misclassification probability is reduced to 0.0078 by taking the cutoff point as 0. The prior probabilities of case (a) are quite realistic; if anything, even smaller prior probabilities of synonymy would be expected in practice.

More Detailed Investigation of Selected Measures. In order to decide which of the two measures G_3 and G_7 give better discrimination, it is necessary to examine the test results in more detail. We begin by examining point (ii) in Section 4. That is, we are interested in whether a stronger indication of synonymy or nonsynonymy is obtained as the available evidence (sample size) for a pair increases. For this purpose we have plotted horizontally in Figures 4 and 5 the average of the frequencies for the individual words in a pair, i.e., $(n_i + n_j)/2$. This is shown on

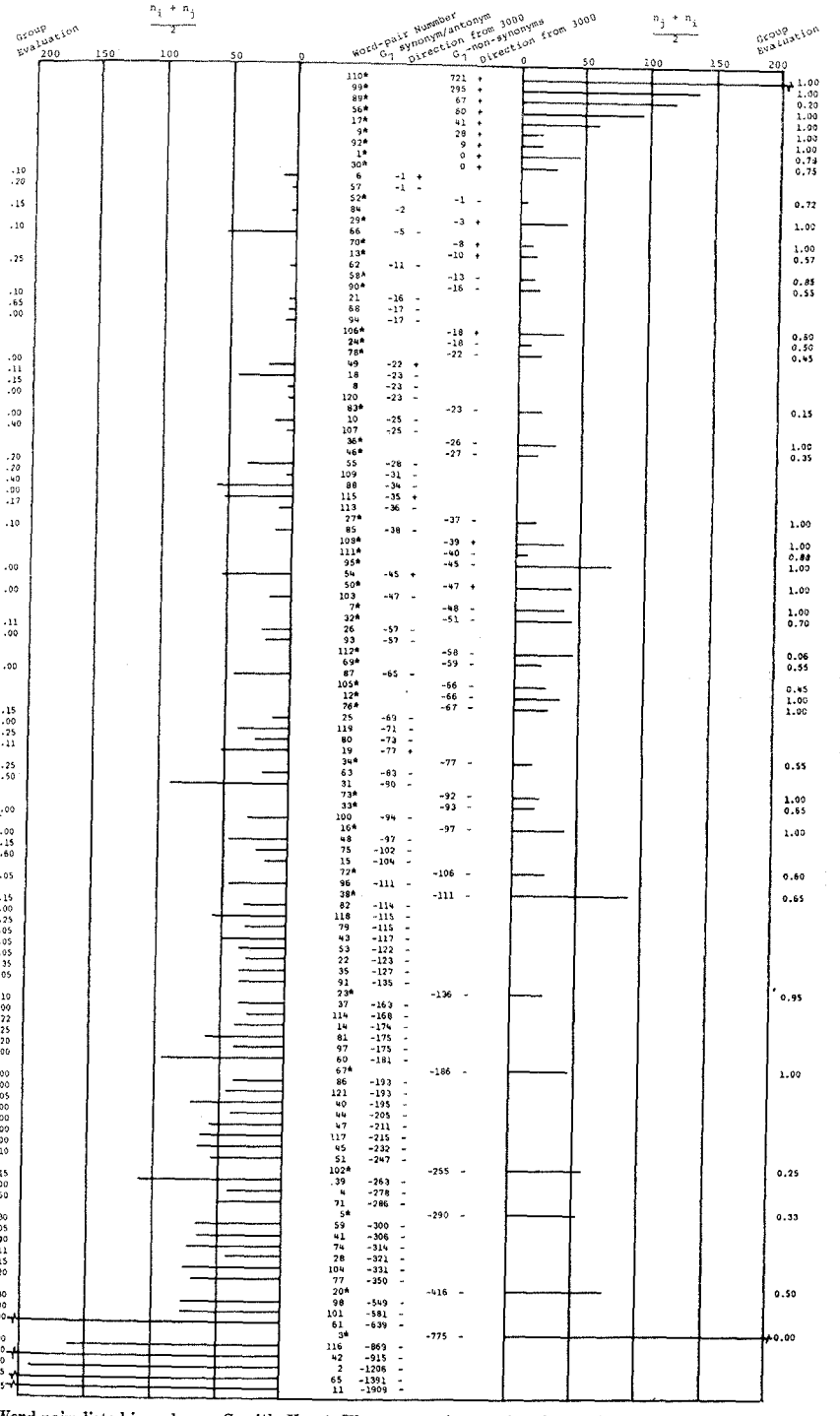

FIG. 5. Word pairs listed in order on G_7 with $K = 1$. The center columns give the word-pair number and the value of G_7 followed by an indication of the increase or decrease of that value when compared with the sample of 3000 titles. The average frequency for each word pair is shown as a horizontal bar. The outer columns give a quantification of the synonymy ratings of the word pairs by a group of physicists.

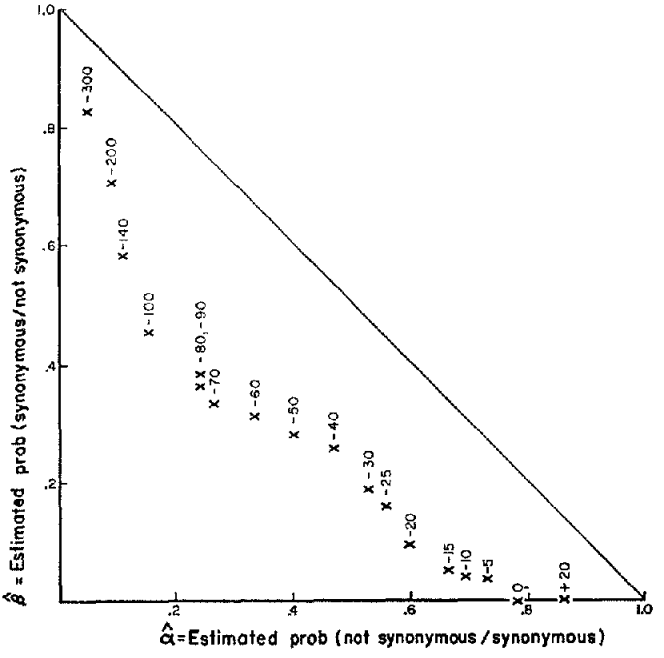

FIG. 6. Estimated error probabilities for discrimination using the measure G_7 with $K = 1.0$; G_7 values as parameter

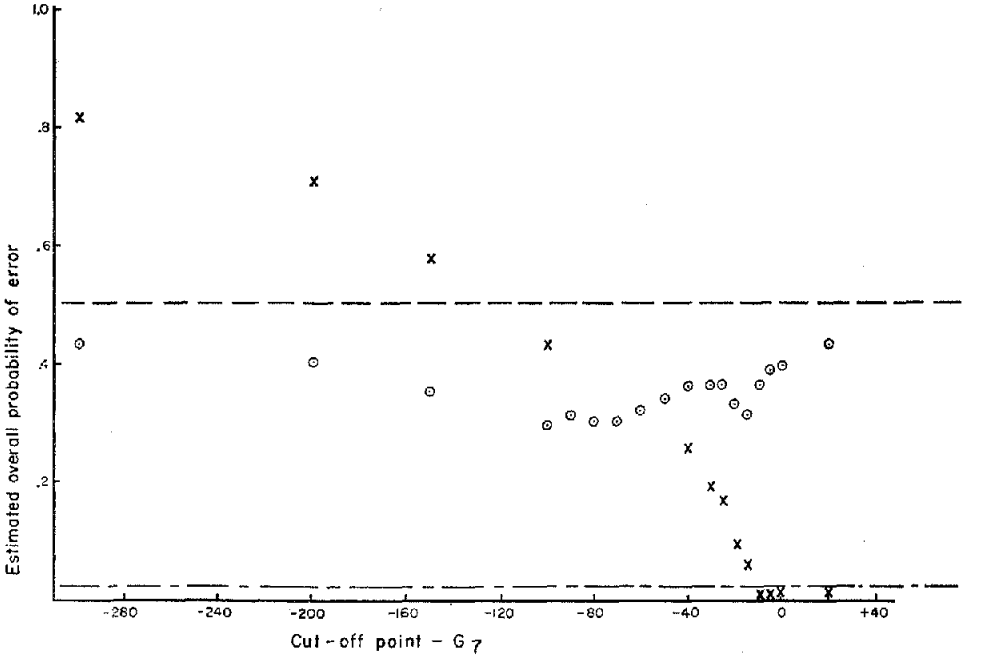

FIG. 7. Estimated overall error probabilities for G_7 with $K = 1$: (a) prob (not synonymous) = 0.99 is indicated by ×, (b) prob (not synonymous) = 0.50 is indicated by ⊙

the left for pairs precategorized as nonsynonyms, and here a definite trend to high average frequencies as values of the measure decrease is expected. For pairs in the "don't know" region—just below zero—on the average, low average frequencies are expected. Similarly, average frequencies for pairs precategorized as synonyms are shown on the right, and they would be expected to increase as the measure values increase in the "accept" region—above 0.

These effects are seen most strongly for nonsynonymous pairs in Figure 5. The high-frequency pair number 66, "excitation: excited," is an anomaly; it is classified as a synonym by almost all the measures which have been investigated. For synonyms the expected trend appears in both Figures 4 and 5 in the "don't know" and "accept" regions. There are also pairs precategorized as synonyms which have large negative values of G_3 and G_7 and large average frequencies. We consider these latter pairs in Section 6. The measure G_7 would appear to be better than G_3 merely on the basis of the more pronounced trend in average frequency for nonsynonymous pairs.

Two additional questions immediately present themselves.

(1) The first is whether synonymous pairs in the "don't know" region would move into the "accept" region and the nonsynonymous pairs in the "don't know" region would become more negative as the number of titles, and therefore the evidence for each pair, increased. It is not possible to answer this question completely across an ensemble, i.e., by looking at a cross-section of test pairs for a given number of titles, although the trend in average frequencies would suggest this to be true. Ideally one would want to add more titles and recompute the measure values; e.g., increase the sample to 12,000 titles. Since this would have involved considerable time and computation, it was decided instead to investigate the measure values for a sample of 3000 titles from the 6000 titles.

The results of the analysis of 3000 titles are also shown in Figures 4 and 5. On the right of each measure value there is a plus or minus, according to whether the measure value increased or decreased in going from 3000 to 6000 titles. In Figure 5 it can be seen that some of the values for synonymous pairs in the "don't know" region are increasing while almost all of the values for nonsynonymous pairs in the "don't know" region are decreasing. It can therefore be concluded that the measure is behaving as expected and that slightly better discrimination would be obtained as the number of titles increased.

(2) The second question, somewhat related to the first, is whether the optimum value of the constant K will be the same when the sample size for pairs goes up, either within a given set of titles or because the set of titles is getting bigger. The constants were found to be the same for the 6000 titles as for the 3000 titles for both G_3 and G_7. However, one could argue that as the number of titles from a homogeneous sample increased, and consequently the n_i's increased, then the set D_i, the context of x_i, would eventually receive very few new members. Also for $x_k \in D_i$, the frequency n_{ik} would increase and eventually the reduced context of x_i would be virtually the same as D_i. The measures G_3 and G_7 would then have eventually the same quantities in their equations, but different constants K, i.e., $K = 5$ and $K = 1$, respectively. One, or both, of these constants would not be optimal.

To investigate this point further an additional sample of 109 word pairs was

selected. For this sample each word in the pair had a frequency, n_i, greater than 49. Only 2 pairs of this sample were precategorized as synonymous. Listing these pairs on G_3, we found that a value of $K = 3$ was optimal; for $K = 5$, 12 of the 107 nonsynonymous pairs were above 0. This is consistent with the appearance of nonsynonymous pairs with high average frequencies high on the list in Figure 4.

The classification by G_7 was optimal with $K = 1$, with a value of $K = 0.9$ being about as good. For both these K values all the nonsynonymous pairs had G_7 values below 0, while one of the synonymous pairs had a very high positive value of G_7 and the other a value just below 0.

The evidence drawn from this additional sample is therefore that the measure G_7 gives better discrimination than the measure G_3.

6. *Empirical Modification of a Measure to Account for Imbalance in Contexts*

Up to this point the data has been used to verify the discriminating power of several intuitively derived measures; of those investigated, G_7 appears to be the best measure. Now we are interested in examining in the corpus the numerical characteristics of particular sample pairs to see whether any of the shortcomings of the measures can be obviated. The main drawback of the scheme based on G_7 is the poor selection of pairs precategorized as synonyms. Figure 5 shows that there is a tendency for some synonymous pairs with high average frequencies to have strongly negative values of G_7. Thus, although frequency has a strong effect, there is also some other factor (or factors) influencing the value of G_7 for synonymous pairs.

One factor to investigate is imbalance in the contexts of the individual words in a pair—point (iii) in Section 4. This is the only one of the four points which has not yet been justified. The effect of imbalance was masked out by the use of average frequency, $(n_i + n_j)/2$, in the investigation of the measures to this juncture.

In investigating imbalance we attempted to eliminate the effect of average frequency by taking from Figure 4 the synonymous pairs and nonsynonymous pairs with average frequencies between 20 and 60. This is a larger spread than is desirable, but was necessary to obtain a sufficiently large sample for the investigation. In Table 3 these pairs are listed in the order of their G_7 values, separately for synonyms and nonsynonyms, and a measure of their imbalance, in terms of the reduced contexts R_i and R_j, has been computed and is shown in the final columns. The synonymous pairs show a definite trend of increasing imbalance with decreasing values on G_7, which is not evident for the nonsynonymous pairs. This indicates that point (iii) of Section 4 is erroneous and that synonymous pairs may have imbalanced contexts.

To overcome this we modified G_7 with an imbalance correction factor to obtain G_{10}:

$$G_{10} = G_7 + C \mid r_i^j - r_j^i \mid .$$

It was found empirically that values of $K = 0.8$ and $C = 2.4$ gave optimal discrimination on the sample pairs from 6000 titles; this is shown in Figure 8. The sample pairs using 3000 titles were also examined and the results are indicated by a plus or minus on the G_{10} values in Figure 8. A plus indicates that the G_{10} value obtained from the 6000 titles was greater than the G_{10} value obtained for the same

TABLE 3. INVESTIGATION OF IMBALANCE EFFECT FOR THE MEASURE G_7

Synonyms							Non-synonyms						
Pair Number	G_7	Av.(n_i,n_j)	r_i^j	r_j^i	Total $= d_{ij}+r_i^j+r_j^i$	$\% = \frac{\lvert r_i^j - r_j^i \rvert}{\text{Total}}$	Pair Number	G_7	Av.(n_i,n_j)	r_i^j	r_j^i	Total $= d_{ij}+r_i^j+r_j^i$	$\% = \frac{\lvert r_i^j - r_j^i \rvert}{\text{Total}}$
1	0	42.5	33	16	110	.155	49	-22	23.0	26	3	46	.500
30	0	24.0	20	10	56	.179	18	-23	45.0	24	24	85	.000
29	-3	33.0	16	23	86	.081	55	-28	39.0	33	3	63	.476
106	-18	32.0	22	9	58	.224	115	-35	53.0	34	28	114	.053
36	-26	26.5	23	6	47	.340	54	-45	56.5	23	32	102	.088
108	-39	35.0	32	12	76	.263	103	-47	20.5	0	21	29	.724
50	-47	40.0	43	9	89	.382	26	-57	24.5	28	6	57	.386
7	-48	35.5	2	37	67	.522	87	-65	45.5	28	36	105	.076
32	-51	41.0	44	3	73	.562	119	-71	41.5	30	18	76	.158
112	-58	49.0	3	45	73	.575	80	-73	28.5	24	18	66	.091
105	-66	22.5	15	19	46	.065	19	-77	54.0	21	47	116	.224
12	-66	32.0	51	0	74	.689	100	-94	35.5	35	23	87	.137
76	-67	24.5	15	20	55	.091	48	-97	50.0	38	21	95	.179
16	-97	39.0	7	57	96	.521							
72	-106	23.0	2	46	57	.772							
23	-136	24.5	8	51	70	.614							
67	-186	44.0	3	74	99	.717							
102	-255	57.5	8	78	109	.642							
5	-290	52.0	0	85	95	.895							

pair from 3000 titles; vice versa for the minus. It is clear that some discrimination of synonymous pairs has been gained at the expense of a slight loss of discrimination against nonsynonymous pairs. The estimated error probabilities for G_{10} are shown in Figure 9. There is a slight improvement over G_7, but there is now another constant in the measure. However, the frequency effect becomes clearly visible for synonymous pairs, except for the few pairs with large negative G_{10} values. The measure G_{10} was tried on the set of 109 word pairs with average frequencies above 49 and, using $K = 0.8$ and $C = 2.4$, only 2 of the 107 nonsynonymous pairs were classified as synonyms. The two pairs which had been precategorized as synonyms were correctly classified as synonyms. This is consistent with the results from the first set of sample pairs.

7. *A Check on the Degree of Synonymy*

Since it is not in general possible to say in a dichotomous way that a pair of words is synonymous or not synonymous, there is an upper limit to the discrimination which any method can achieve. Therefore, before attempting to improve on the discrimination obtained using G_7 and G_{10} by looking for new measures, we felt we should obtain more information on what this upper limit might be.

To this end, ten physicists were given lists of the 120 sample word pairs (with 23 singular-plural word pairs excluded). Without any guiding information (no operational definitions), the physicists were asked to decide whether pairs were:

(a) synonyms,
(b) near approximation to synonymy,
(c) antonyms,
(d) near approximation to antonymy,
(e) neither synonym nor antonym.

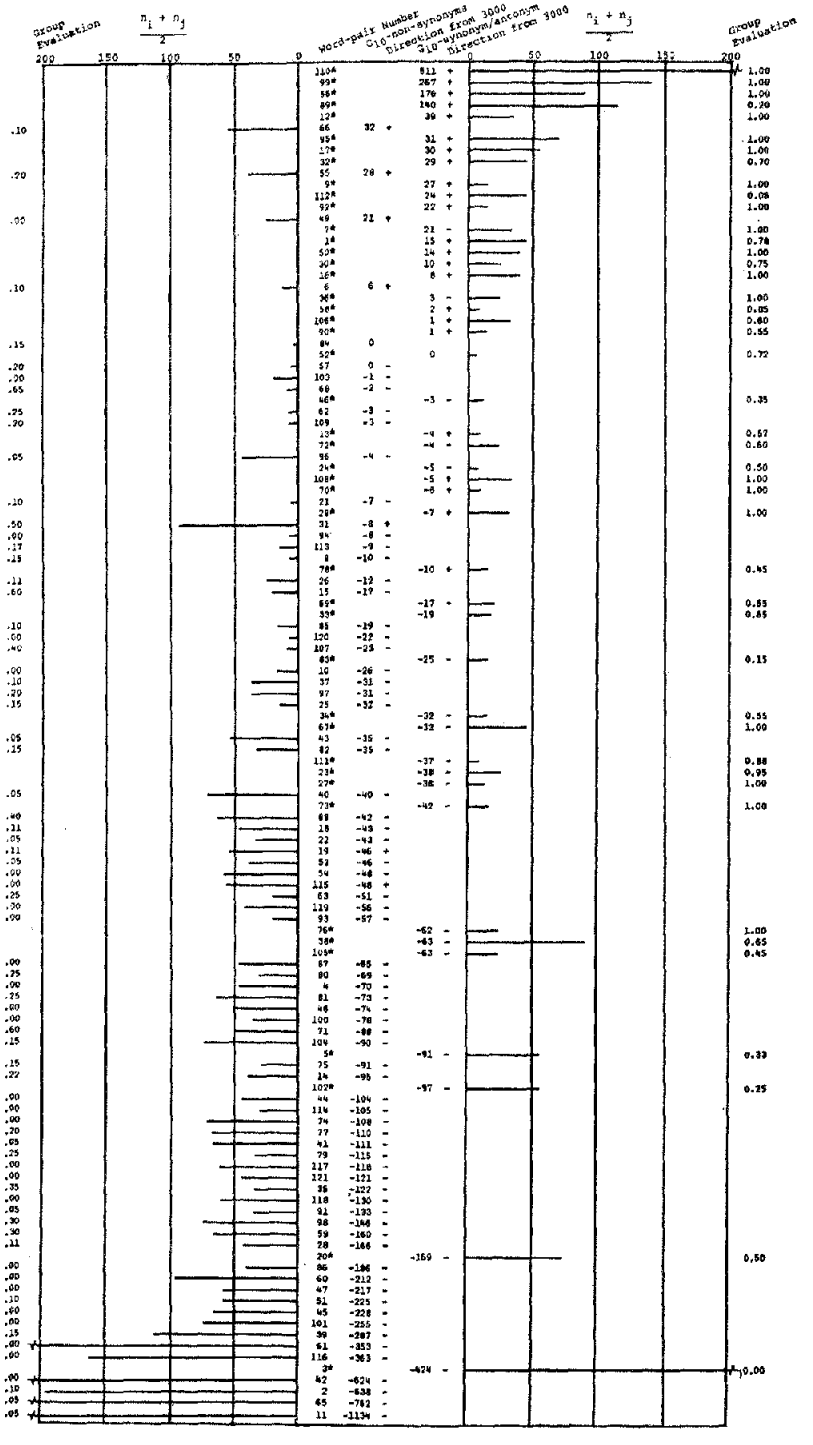

FIG. 8. Word pairs listed in order on G_{10} with $K = 0.8$ and $C = 2.4$. The center columns give the word-pair number and the value of G_{10} followed by an indication of the increase or decrease of that value when compared with the sample of 3000 titles. The average frequency for each word pair is shown as a horizontal bar. The outer columns give a quantification of the synonymy ratings of the word pairs by a group of physicists.

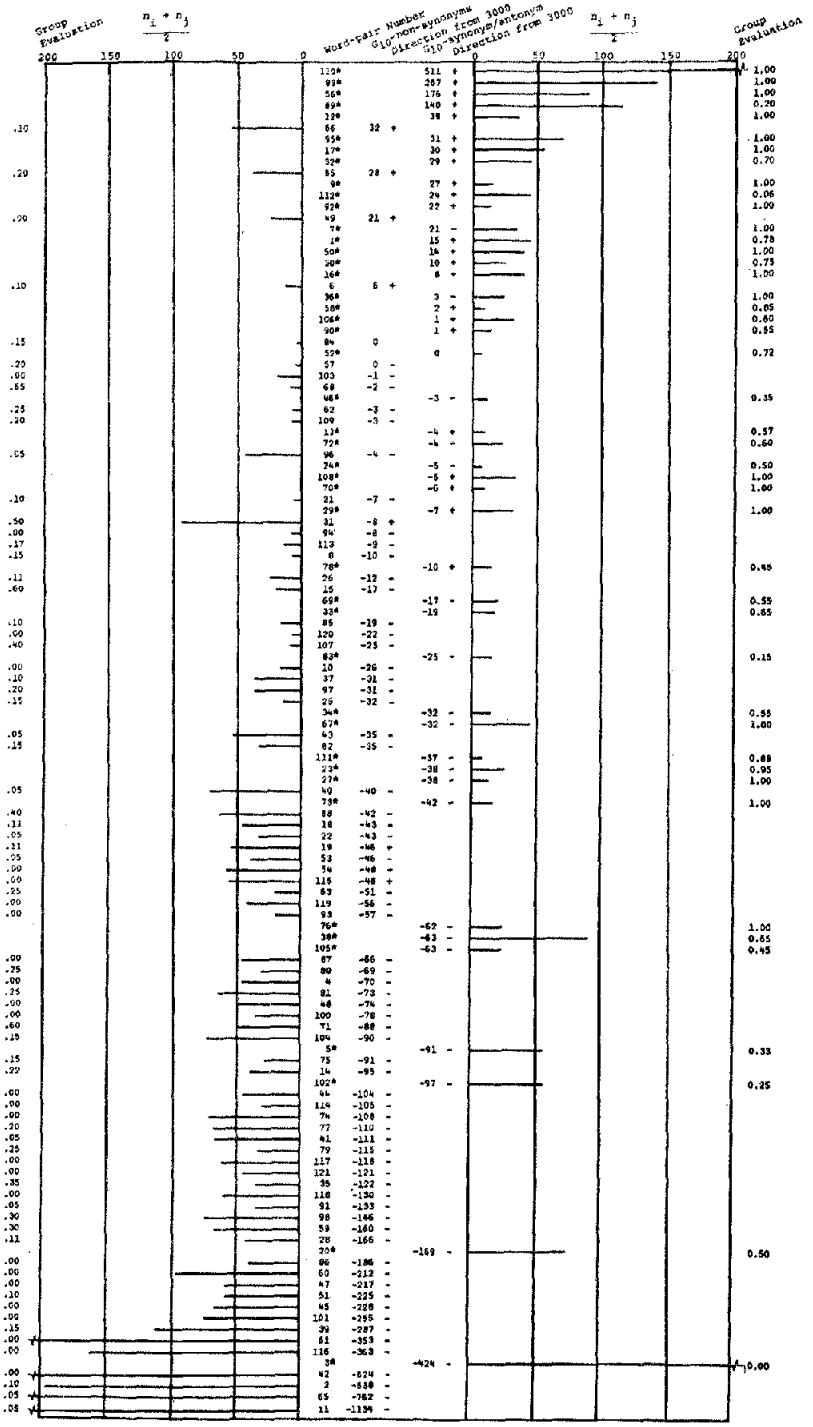

FIG. 8. Word pairs listed in order on G_{10} with $K = 0.8$ and $C = 2.4$. The center columns give the word-pair number and the value of G_{10} followed by an indication of the increase or decrease of that value when compared with the sample of 3000 titles. The average frequency for each word pair is shown as a horizontal bar. The outer columns give a quantification of the synonymy ratings of the word pairs by a group of physicists.

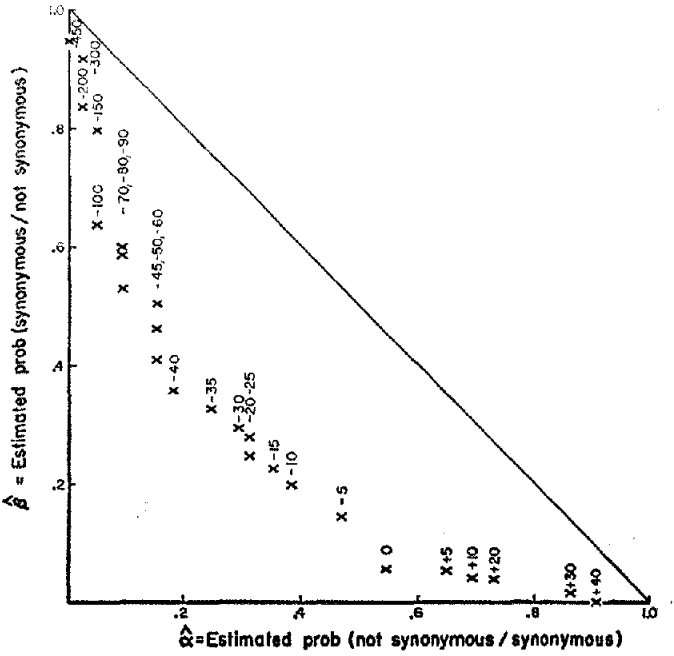

FIG. 9. Estimated error probabilities for discrimination using the measure G_{10} with $K = 0.8$ and $C = 2.4$; G_{10} values as parameter

If either (a) or (c) was selected by an individual physicist, a score of 1 was given to the pair; a score of $\frac{1}{2}$ was given for (b) or (d), and a score of 0 was given for (e). A score of 1 was assigned by the experimenters to singular-plural word pairs. Averaged over the ten physicists, this gave a measure of the degree of synonymy/antonymy for each pair.

The results showed disagreement by the physicists with our precategorization of synonyms/antonyms on the basis of the operational definitions given in Section 5. For those 45 pairs precategorized as synonym/antonym, the average of the physicists' scores was approximately $\frac{3}{4}$. One can therefore estimate an α for the physicists, using the precategorization as a standard, as being $\hat{\alpha} = \frac{1}{4}$. The corresponding β was estimated to be approximately $\hat{\beta} = \frac{1}{10}$. These are rough figures, but they verify that there is an upper limit other than $[\alpha = 0, \quad \beta = 0]$ which can be achieved by any method of discrimination. However, if the point $[\hat{\alpha} = \frac{1}{4}, \quad \hat{\beta} = \frac{1}{10}]$ is plotted in Figures 6 and 9, it is evident that the statistical measures have not achieved this upper limit. This is true even if the relatively small number of physicists polled is allowed for.

Another comparison of the results of this test with the discrimination achieved by the measures is obtained by listing the average of the physicists' scores for each pair on the extreme right and left in Figures 5 and 8. If the pairs in the "don't know" region—for example, G_7 or G_{10} values from 0 to -20—are ignored, there is a rough correlation between the rankings induced by G_7 and G_{10} and the rankings of the physicists. This is especially true for the ranking induced by G_{10}. In particular, note that of the pairs precategorized as synonym/antonym, the four ranked lowest by G_7 and G_{10} were strongly rejected by the physicists.

8. *Unexamined Features of the Empirical Results*

Whatever further gains can be achieved in the discrimination of synonymous from nonsynonymous pairs by statistical means will probably come from one of two sources:

(a) The first would be a detailed examination of the contextual environment of word pairs which have behaved anomalously in the study. For instance, pair 66, "excitation : excited," is classified as a synonym by almost all measures and will clearly not be classified as a nonsynonym no matter how large a sample size is obtained. On the other hand pair 67, "film : films," is definitely synonymous but is rejected as such by all the measures even though it has a large sample size.

Questions arise such as: What effect does the length of title or the variation in the structure have on the specification of context? What effect has a difference in syntactic role of each word of a pair? For example, in a word pair such as "film : films," one word may be used exclusively as a noun (films) whereas the other may be used as both adjective and noun. Do the adjectival uses tend to introduce extraneous subject matter with respect to synonymy—"photographic film" versus "fluid film lubrication"? Investigation of these questions might lead to a strengthening of the model by inclusion of appropriate language features.

(b) Another possibility is improvement on intuitive grounds of the measures we have designed. In going from G_3 to G_7 we used as a measure of "extraneous" context the frequency of occurrence of a word, x_c, in D_a or D_b. We said in effect that if x_c occurred only once with x_a or x_b, then this occurrence could be treated as random with no semantic significance. More formal measures of statistical association, which hopefully reflect semantic association, are available; an example is the quantity $(n_{ac} - n_a n_c/n)$. This will be positive if the joint frequency of occurrence, n_{ac}, is greater than its predicted value, $n_a n_c/n$, under a hypothesis of randomness. Thus a measure using this idea could be formed as follows:

$$G_{11} = \sum_{k=1,\ k\neq a\neq b}^{\mu(X)} H_k'',$$

where

$$H_k'' = \begin{cases} \operatorname{av}\left\{\max\left[0, \left(n_{ak} - \dfrac{n_a n_k}{n}\right)\right],\ \max\left[0, \left(n_{bk} - \dfrac{n_b n_k}{n}\right)\right]\right\} & \text{if } n_{ak} > 0,\ n_{bk} > 0; \\ -\max\left[0, \left(n_{ak} - \dfrac{n_a n_k}{n}\right)\right] & \text{if } n_{ak} > 0,\ n_{bk} = 0; \\ -\max\left[0, \left(n_{bk} - \dfrac{n_b n_k}{n}\right)\right] & \text{if } n_{ak} = 0,\ n_{bk} > 0; \\ 0 & \text{otherwise.} \end{cases}$$

The utility of the measures derived in this paper will also be clearer if they are tested on different samples. By different samples here is meant not only titles from disciplines other than physics, but also sets of descriptors from a collection of documents. Of particular interest is the stability of constants such as K in the measures.

The question of discrimination of word pairs for which the condition $n_{ij} = 0$ does

not hold has also not been examined empirically, but here it is felt that the measures would carry over with d_{ij} replaced by p_{ij}, the size of the *pairwise* context of a pair of words. This conjecture would require empirical verification.

9. *Summary and Conclusions*

An assessment of the adequacy of the discrimination of synonymous from nonsynonymous pairs of words achieved by the measures G_7 and G_{10} can only be made if prior probability of a word pair being synonymous or if costs of making the two types of error are available. If the lack of dichotomy in the synonymy relationship is neglected, then all of the information which is needed to assess the utility of the measures for a given system is contained in the estimated error probability curves given in Figures 6 and 9. Ideally these curves should remain as close as possible to the vertical and horizontal axes, and it can be seen that in Figures 6 and 9 this is far from the case.

Let us now evaluate the experimental results.

In the original sample of word pairs, shown in Table 1, and the additional 109 word pairs with average frequencies above 49 (Section 5), there were in all 182 nonsynonymous pairs, all of which were correctly classified by the measure G_7 with $K = 0.9$ and a cutoff point of zero. Therefore an upper 99 percent confidence limit for β is approximately $1/182 \simeq 6/1000$. Furthermore, with the same measure and cutoff point, the estimated error probability $\hat{\alpha}$ is $\frac{4}{5}$.

As for prior probabilities, we have remarked previously that for most systems the probability of a word pair being synonymous is very small, and under this assumption, some evaluation of the measure can be made. As a specific example, assume that this probability is 1/1000.

Using this assumption, we can now see what the error probabilities achieved in our experiment would mean to a user if the corpus of titles were regarded as a document collection. The inquirer would supply one word and the system would search through the stored list of corpus words which do not co-occur with the given word. Then it is expected that with a high probability no more than 6 out of every 1000 words examined would be incorrectly selected as synonyms/antonyms. With the prior probability assumed above, only 1 in 1000 of the words examined would, on the average, be synonymous/antonymous with the given word, and the chance of selecting this word as a synonym/antonym is only 1 in 5.

Two further points should be borne in mind.

(a) The empirical evidence shows that the error probability, α, decreases with increasing corpus size, although it can be seen that it is never lower than approximately $\frac{2}{3}$. This figure was obtained by assuming that all of the synonyms/antonyms in Figure 5 which are marked with a plus eventually move into the "accept" region. By the same reasoning, four nonsynonyms would move into the "accept" region and thereby have a slight adverse effect on β.

(b) As noted in Section 2, classification theory presumes that the words being examined are placed in either of the two classes on a binary basis. The procedures which we have used to achieve classification offer the additional possibility of a ranking of the pairs within a class on the basis of measure value. If in the synonym/antonym class of words derived for the user (see above), the word with the highest measure value is selected, then the probability that it is a synonym/antonym should

be much greater than if a word in the class at random is selected. Figures 5 and 8 support this conjecture.

Mention should also be made of the practical matter of the expense of calculating the measures for a corpus—though this was not a point of direct concern in this research. The requirement for storing and searching current data on the co-occurrence of all words in a growing corpus would not be trivial to implement. It is clear that this would be a factor in any application of the statistical measures we have discussed.

In conclusion the following can be said. The measures do have power to discriminate between synonymous and nonsynonymous pairs. The lack of complete dichotomy in the definition of synonymy places an upper limit on the discriminating power which can be achieved, but the test with the physicists shows roughly that this upper limit is well above the discriminating power achieved by the measures. Whether the discrimination which has been attained is adequate for any particular system can only be assessed in terms of the a priori probability of a pair being a synonym and the costs of error associated with the system.

RECEIVED MARCH, 1966; REVISED JUNE, 1966

REFERENCES

1. STILES, H. E. The association factor in information retrieval. *J. ACM 8* (April 1961), 271–279.
2. GUILIANO, V. E., AND JONES, P. E. Linear associative information retrieval. In Howerton and Weeks (EDS.), *Vistas in Information Handling, Vol. 1*, Ch. 2, Spartan Books, Washington, D. C., 1963.
3. BAKER, F. B. Information retrieval based on latent class analysis. *J. ACM 9* (Oct. 1962), 512–521.
4. DOYLE, L. B. Semantic road maps for literature searchers. *J. ACM 8* (Oct. 1961), 553–578.
5. SPIEGEL, J., BENNETT, E., HAINES, E., VICKSELL, R., AND BAKER, J. Statistical association procedures for message content analysis. Information System Language Studies No. 1, Rep. No. SR-79, The Mitre Corp., Bedford, Mass., Oct. 1962.
6. BAXENDALE, P. B. An empirical model for computer indexing. In *Machine Indexing*, American U., Washington, D. C., 1962, pp. 207–218.
7. BIRNBAUM, A., AND MAXWELL, A. E. Classification procedures based on Bayes' formula. *Appl. Statist. 9* (1960), 152–169.
8. MARCKWORTH, L. M. *Dissertations in Physics; An Indexed Bibliography of All Doctoral Theses Accepted by American Universities, 1861–1959*. Stanford U. Press, Stanford, Calif., 1961.
9. DEWEY, GODFREY. *Relativ Frequency of English Speech Sounds*. Harvard U. Press, Cambridge, Mass., 1923, p. 19.