[*International Conference on Machine Translation of Languages and Applied Language Analysis*, National Physical Laboratory, Teddington, UK, 5-8 September 1961]

PROCEDURES FOR THE DETERMINATION OF DISTRIBUTIONAL CLASSES*

by

K.E. HARPER

1. INTRODUCTION

STUDIES in Distributional Semantics are now underway at The RAND Corporation based on the 250,000 word corpus of Russian Physics text.** For present purposes, it is important to note that this text has been subjected to machine translation and human post-editing, and that a glossary of the forms found in the text and a syntactic description of each sentence is preserved on magnetic tape. The syntactic description is based on dependency analysis, which specifies the dependency pairs in a sentence and arranges them in a tree-like structure.***. This information will be subjected to automatic analysis for a number of language data processing purposes, including the study of word correlation, which we term distributional semantics. The present paper describes some of the typical procedures that will be used in automatic analysis, and discusses some of the problems involved. Data derived from current, semi-automatic analysis are also presented.

The *Syntactic Combination* in the RAND system of sentence-structure analysis designates a two-term combination (governor and dependent). The capability of dealing with multiple-term combinations is inherent to the system, but will not be considered here. The combinations in our text (approximately 240,000 in number) are derived from two sources: (i) the glossary, which is a list of all the forms in our text, together with identifying codes ("word numbers") and Grammar Codes specifying the morphological and, to some extent, syntactic characteristics of each form; ii) the Dependency Table, which is a list of the Grammar Code pairs or allowable syntactic combinations, together with an indication of the Direction of dependency. For purposes of machine translation, Grammar Codes are deficient when they contain, for a given form, less syntactic information than a human user of the language has at his disposal. This deficiency arises both from the inadequate description of syntax that is characteristic of all grammars, and from the slighting of semantic factors.

* This study was supported by the U.S. Air Force Office of Scientific Research. This paper is their Technical Note TN-149.

** The procedures followed in the processing of this text are detailed in the series of papers, "RAND Studies in Machine Translation", Nos. 1-10.

*** Procedures for automatic determination of sentence structure are given in the RAND paper, RM-2538, *Studies in Machine Translation - 10: Russian Sentence-Structure Determination* (D.G. Hays and T.W. Ziehe). See also RM-2068, *Studies in Machine Translation - 8: Manual for Post-editing Russian Scientific Text* (K.E. Harper, D.G. Hays, and B.J. Scott).

For example, if our Grammar Code, specifies merely that a given word is an adverb, and the Dependency Table allows all adverbs to depend on all verbs and adjectives, we will have an overly-generalized grammar. In Russian, OCHEN' (very; very much) will not be found to qualify all verbs or adjectives: the combinations, ON OCHEN' NACHAL (he began very much) or OCHEN' ORTOGONAL'NYJ (very orthogonal) do not occur. Which verbs and adjectives fall into the categories of possible or impossible governors of OCHEN'? The establishment of such categories, or Distributional Classes, is a part of the business of Distributional Semantics. The immediate purpose is the improvement of the specificity of the syntactic codes used in machine translation. A more general purpose is the establishment of broad semantic classes based on combinatorial possibility or probability.

In our usage, a *Distributional Class* (DC) is a list of words bearing a specified syntactic relation to another group of words (or, rarely, to another single word). The DC is *derived, i*.e., it is obtained by "testing" against another category of words; in this sense, the DC is secondary and the words "tested against" are primary. Since the relationship is a syntactic one, the DC is essentially a morphological class, although it may include more than one part-of-speech; further restrictions (such as case, number, tense) may be placed upon the DC during the testing process. The primary class may be formed in accordance with any criteria desired. Examples are: (i) a morphological class (e.g., nouns, plural nouns, future-tense perfective verbs), (ii) a syntactic class (e.g., governors of the infinitive, verbs governing both the accusative and dative case, prepositional equivalent determinants*), and (iii) classes formed in an *a priori* manner (e.g., animate nouns,** "abstract" nouns, verbs and deverbative nouns of motion, adjectives denoting color, or derived distributional classes). Further restrictions may be placed upon the primary class in the combination to be tested, such as English equivalent selection or word order.

Even within the framework of the two-term syntactic combination, a very large number of DC's can be derived, once it is assumed that primary and secondary categories are of a tentative, experimental nature. Since DC's based on such criteria as graphic features or frequency of occurrence are less likely to result in meaningful or usable information, one may choose to work with more conventional categories. The following is

---

* Cf. RAND paper, P-1941, *Machine Translation of Russian Prepositions* (K.E. Harper).

** It may also be maintained that this class is morphologically derived.

a sample of the DC's derivable from combinations of the type enumerated in the preceding paragraph:

Primary categories	Secondary categories (DC's)
1. adjectives (G) *	1. adverbs uniquely depending on adjectives (D)
2. verbs that govern the infinitive; active (G)	2. Noun subjects (D)
3. animate nouns as actors (D)	3. active verbs (G)
4. adjectives that govern nouns(G)	4. nouns (D)
5. adjectives that do not possess the comparative form (D)	5. nouns (G)
6. verbs, future-tense, perfective, passive (G)	6. instrumental nouns of agency (D)
7. DC 3, above (G)	7. nouns, direct object (D)
8. "abstract" nouns (G)	8. prepositional phrases (D)
9. nouns derived from verbs of motion (G)	9. genitive nouns (D)
10. the preposition U, with the translation "for" (G)	10. nouns, object of the preposition (D)

In some instances, the DC's established above could result in immediately useful syntactic codes in our glossary. In other instances, the utility of the information is obscure; in view of the randomness of the sample and the extremely small size of the text, one may doubt that the information has meaning. It seems advisable to begin collecting data against firmly established morphological or distributional classes. To a degree, this depends on the judgement and intuition of the investigator who must take advantage of his knowledge of the language, and, at the same time, regard his deep-seated beliefs with healthy suspicion.

A Distributional Class is merely a list of words; as such, it is of extremely limited value until it becomes the subject of further abstraction or generalization. It is common knowledge that not all nouns can be the subjects of all verbs. The task, of recording all subject-verb combinations in a huge block, of text seems enormous; lacking a suitable generalization, applicable to sentences not yet written, it may be of doubtful utility. Not possessing a huge block of text, but believing that the problem of meaning can be attacked by data processing techniques, we have adopted the following set of principles in the study of Distributional Classes: (i) priority should be given to the development of a methodology, (ii) we should take advantage of the limited area of discourse (in this instance, physics texts), and attempt less than a complete semantic classi-

"G" - governor; "D" = dependent.

fication of the language, and (iii) we should rely heavily on the evidence in text, generalizing tentatively and with care. In some instances, we would hope to derive automatically a DC that we could have formed intuitively or with the aid of grammars, dictionaries, or a thesaurus. Corroborative evidence of this kind should certainly be utilized, since our hope is to develop procedures for forming DC's of far greater number and complexity than we have been able to form by traditional methods.

2. THE FORMATION OF DISTRIBUTIONAL CLASSES

WE have said that a logical place to begin this study is by looking at the members of well-established syntactic combinations. Further, in dealing with a text of limited size, it is important to deal with combinations that occur frequently. Below we cite examples of three DC's, formed on the basis of such combinations: subject-verb (DC l), noun-genitive noun (DC 2), and adjective-noun (DC 3). To repeat: these DC's are formed by semi-automatic analytic procedures, and are purely experimental. In each instance, the characteristics of the primary category and the conditions of the relationship within the combination are given in summary form. For purposes of automatic analysis, these specifications are, of course, couched in terms of the RAND Grammar Code and sentence-structure determination system. A brief discussion of the advantages and problems of transformational analysis is included. Discussion of the possible significance and use of these DC's is deferred until Sec.3.

2.1 The DC of Verbs Having Only Animate Actor Nouns as Dependents

Primary Categories: (i) animate nouns or first-person pronouns, nominative case; (ii) animate nouns, instrumental case; (iii) zero nominative dependent. Relation: dependent.

Secondary Categories (respectively): (i) active verbs; * (ii) passive verbs or passive participles; (iii) first-person verb. Relation; governor. Verbs meeting these requirements are also subject to the restriction that an inanimate noun did not appear as the subject of their active forms.

A search of existing analytic reports covering a text block of 120,000 running words resulted in a list of 103 words in this class, which we shall call DC 1. A by-product of the search was a list of 46 verbs (DC 1.1) which sometimes have animate actor dependent nouns. An obvious means of increasing the number of words in DC l, and of increasing the total frequency of occurrence, is the use of transformations, particularly

* For this purpose, "active" signifies all non-passive, finite verbs, and the infinitive used with such words as ESLI and SLEDUET.

that of the subjective genitive dependent of a deverbative from a DC 1 verb. Thus, obtaining the deverbative NABLYUDENIE (observation) from the DC 1 verb NABLYUDAT' (to observe), and finding the combination NABLYUDENIYA IVANOVA (the observations of Ivanov) to be of the subjective genitive type, (transformable into Ivanov observes), we can assert that additional evidence of DC 1 has been found.

For a number of reasons, we have deferred attempts to utilize this transformation in establishing DC 1: (i) certain DC 1 verbs do not possess a deverbative; (ii) for many verbs a semantic shift is involved, or suspected, in the deverbative; (iii) determination of the subjective-genitive relationship is often difficult, or merely arbitrary in ambiguous circumstances; (iv) this relationship is not coded during the process of sentence-structure determination. It is clear that in our text an animate genitive noun dependent of one of these deverbatives is a positive indication of the subjective genitive. The relationship with respect to the inanimate noun dependent is less clear, and apparently depends upon a more precise classification of deverbatives than we now possess.

2.2 The DC of Governors of Nouns Naming Physical Particles

An *a priori* group of nouns naming physical particles,* called DC 2, was formed. This class was selected because of a presumed semantic homogeneity and because of high frequency of occurrence in text. Three main classes of governors have been distinguished, and under the noun governors, further sub-classification has been made. The data was retrieved from analytic reports for 120,000 words of text.

DC 2.1 Noun governors of DC 2 (genitive case)

Primary Category: DC 2, genitive case. Relation: dependent.

Secondary Category: noun, any case. Relation: governor.

A total of 202 words were found for DC 2.1. These were arbitrarily classified as follows:

DC 2.1.1. Deverbatives from transitive verbs that are passive in -SYA. 42 words. Example: ISPUSKANIE (emission).

DC 2.1.2. Deverbatives, other. 46 words. Example: DVIZHENIE (movement).

---

* e.g., MEZON (meson)

DC 2.1.3. Quantifiers. 9 words. Example: CHISLO (number).

DC 2.1.4. Other non-deverbatives. 105 words. Example: MASSA (mass).

DC 2.2. Verbs for which DC 2 is the object of the action.

Primary Category: DC 2, (i) accusative case; (ii) nominative case. Relation: dependent.

Secondary Category: verb, (i) transitive, active form, or active participle; (ii). transitive, passive form, or short-form participle. Relation: governor.

31 words. Example: SCHITAT' (consider).

DC 2.3. Verbs (active, for which DC 2 is subject.

Primary Category: DC 2, (i) nominative case; (ii) any case. Relation: (i) dependent; (ii) governor.

Secondary Category: verb, (i) active, finite; (ii) active, non-reflexive participle. Relation: (i) governor; (ii) dependent.

33 words. Example: OBRAZOVAT' (form).

Given any set of two DC's, we can of course establish another DC with respect to their combination. As an illustration, we cite the following three DC's formed on the basis of DC 2 and DC 2.1.3.

DC 2.1.3.1. Noun governors of the combination 2.1.3 + 2 (genitive case) *and* of 2 (genitive case). 14 words. Example: IZMENENIE (change).

DC 2.1.3.2. Noun governors of 2 (genitive case) *but not* of 2.1.3. + 2 (genitive case). 188 words. Example: DELENIE (fission).

DC 2.1.3.3. Noun governors of 2.1.3 + 2 (genitive case) *but not* of 2 (genitive case). 14 words. Example: OPREDELENIE (determination).

2.3 The DC's of "Word Governors" of the "Modifier", REZKIJ (Sharp)

(The term, "word governor", as adopted here is to designate words of a common derivational family and meaning, but possibly of different parts-of-speech. "Modifier" in this context denotes a word attributed

to any variant of the word governor. In the following expressions, the first member is a modifier, and the last a word governor: "sharp difference", "sharply differs", "sharply different", and "sharpness of difference". This relationship is represented in different ways on the syntactic level; the terms are adopted as a means of dealing with their likenesses on the semantic level.)

Primary Category: The adjective, REZKIJ, as (i) attributive* (dependent); (ii) short form, masculine, feminine, or plural (governor); (iii) short form, neuter (dependent); (iv) predicative instrumental (dependent).

Secondary Categories: (i) noun (governor); (ii) noun (dependent); (iii) verb or adjective (governor); (iv) noun dependent of the verb governor. (The governor-dependent relationship in (ii) and (iv) is due to conventions in the RAND sentence-structure determination program).

Automatic procedures for establishing DC's of this type of course depend upon a system of classifying words into word "families". These families must be large enough to include OTLICHIE (difference) and RAZLICHIE (difference), but exclusive enough to omit NALICHIE (presence). We do not presently possess a system capable of making these distinctions; indeed, it appears that distributional semantics is one of the prerequisites for such a system.

The "word governors" of REZKIJ will be called DC 3. In a text of 210,000 words, 33 such governors were found; the combination occurred 85 times. The distribution for each corpus (30,000 running words of text) was as follows:

	Corpora						
	1	2	3	4	5	6	7
Number and frequency of new governors in each corpus	10/13	7/14	1/5	5/6	3/6	2/3	2/2
Total number and frequency of governors in each corpus	10/13	13/21	7/9	8/9	10/18	4/5	8/10

* This characteristic is determined by two factors; the Grammar Code of the adjective, and the relative position of the governing noun. The noun, REZKOST' (sharpness) has not occurred in our text; its specifications as a member of the primary category have, therefore, been omitted.

3. THE SIGNIFICANCE OF DISTRIBUTIONAL CLASSES

IN the preceding section, procedures have been given for the establishment of three distributional classes: verbs that have only animate nouns as subjects (DC 1), the governors of nouns naming physical particles (DC 2), and word governors of a given adjective (DC 3). With one or two minor exceptions, the procedures set forth can be carried out automatically over any extent of text that has been processed. At this early point, however, we can only begin to assess the utility of these arbitrarily-formed classes and the fruitfulness of the process as a whole.

A number of problems immediately present themselves:

(i) The criteria for forming these DC's are too obvious or irrelevant. The criterion of animacy (for DC 1) is an obvious one, but one that should eventually result in such well-known sub-classifications as verbs of cognition, communication, etc. On the other hand, the criterion may be so loose as to be meaningless: almost any verb may have an animate subject (one soon comes to the conclusion that the latter objection has little relevance to physics texts). The criteria for DC's 2 and 3 may be objected to on grounds of over-specificity: how can meaningful distinctions be made between the governors of such a small group of words? Should not these distinctions be founded on more substantive grounds?

It would appear difficult either to support or to refute these objections in advance. Distributional classes represent facts of language, but the optimum procedure for analysis of this very large number of facts must probably come from experience. We must simply begin; if we knew how to begin, it would be unnecessary to begin at all.

(ii) Membership or non-membership in a given class is based on chance, and is therefore a poor basis for generalization. DC l, for example, contains a number of verbs that can also be used with inanimate subjects. As more text is processed, the class will probably decrease in size, because of the criterion of "exclusiveness" (i.e., the exclusion of verbs which also have taken inanimate nouns as subjects). Classes 2 and 3 can only increase in size, since none of their present members is subject to exclusion. From this point of view the categories for forming classes 2 and 3 are more satisfactory. From another point of view, the fact that DC 1 will lose a part of its membership does not affect its utility in studies of the present text. We are not so much interested in absolute properties of classes as in their combinatorial properties; all DC 1 verbs in the present text presently possess a commonness that will be useful in the study of their dependents.

It is difficult to predict the significance of DC's 2.1-2-3 (governors of particles). The list is relatively large (comprised of some

200 members), but we do not yet possess data that would indicate the degree of "coverage" for these governors (i.e., the proportion of new governors in new blocks of text). The relatively small number of verb governors of particles is interesting, suggesting either that words like "mesons" can act or be acted upon in very limited ways, or that physicists prefer to write about them as if this were true. At any rate, particles have a disproportionately large number of deverbative noun governors. These characteristics will be compared with the behaviour of other groups of nouns, subject always to the restriction that the homogeneity of DC 2 is a matter of conjecture. In any event, it seems reasonable to presume that if the number of new members of a class markedly decreases with new text, the element of chance membership in a class is less important. The great question is, of course, the size of the corpus required for such computations. One is perhaps not surprised that, in 1,000 pages of text, physicists have applied the adjective REZKIJ (sharp) to only 33 words, or that the number of new word governors of this adjective grows smaller in each block of this text. But neither is one reassured by the small frequency of this adjective (85 occurrences), when attempting to classify the "potential" governors.

(iii) Words that are alike in one respect (i.e., in belonging to a DC) may be unlike with respect to a large number of different combinations; their differences may be more significant than their likenesses.

This is, of course, a real problem. One method of dealing with it is a statistical procedure for assigning "association coefficients" to the members of a DC. We are aware, for example, that the verbs in DC 1 are traditionally classified as verbs of "cognition", "communication", "modality", etc., and that these distinctions may be more essential than the commonness that they share distributionally. Rather than assign syntactic or "semantic" codes on this basis, however, we might proceed statistically: a computer program will be devised for counting the number (and considering the frequency) of certain syntactically equivalent dependents that each pair of verbs has in common. The degree of alikeness is represented by a number (the association coefficient), and through these numbers classes of words are distinguished.

We do not know what the results of such a procedure will be. (It would appear that frequency of occurrences is a more critical factor than the statistical method employed.) We can, of course, compare the machine-generated classes with traditional classes. When the results are different, we may either inquire into the reasons for failure of the statistical procedure, or question the relevancy of traditional classifications to word correlations. In any event, further investigation of the statistically-derived classes may be justified. For example, the traditional classification may not have made a distinction between the verbs, VYCHISLIT' (to compute) and IZUCHIT' (to study). Distributionally, these can be distinguished, with respect to their governorship of

the class of physical participles used as object of the verb: particles can be studied, but not computed.* The human editor can easily verify the result: this is information that "everyone knows", although no one may have ever built it into a syntax. In a real sense, one of the functions of language data processing is the automatic formulation of hypotheses which may be accepted or rejected by the investigator.

It would, of course, be a great over-simplification to assume that our two sample verbs are permanently distinct. On some other distributional basis they may again become members of the same class. Great variation in DC-membership must certainly be characteristic of natural languages. It is by no means certain that a system of major and minor classification if meaningful, nor can we predict the relation of any such system to any existing system. One would hope that a manageable number of classes could be derived, since it is manifestly impractical to record and predict the syntactic combination of each word in our glossary with every other word. If the combinatorial possibilities can be dealt with in our limited universe of physics texts, they may also be dealt with in broader contexts.

(iv) Many words have multiple connotations, as is evidenced by the numerous definitions, synonyms, and special usages cited in dictionaries and thesauri. How can such words be classified, and how can multiple systems of classification be integrated?

The extent of this problem is somewhat hidden in machine translation glossaries, which are characterized by general, "cover-all" target language equivalents. Although the adjective REZKIJ has been adequately translated as *sharp,* it appears to have had at least four connotations in our text: (i) *abrupt* and (ii) *rapid,* when applied to processes *(a sharp rise in temperature);* (iii) *marked* (or clear), when applied to characteristics *(a sharp dependence);* and (iv) connotations of physical configuration *(a sharp edge).* This kind of information suggests that the word governors of the adjective (DC 3) are, in fact, non-homogeneous in different ways. The same conclusion ought to be derivable from a comparison of the appropriate association coefficients of the word governors: if we can intuitively form three classes of noun-governors of *sharp* (processes, characteristics, and objects), we may be able to arrive at the same result via a statistical procedure. (We may also take advantage of morphological characteristics of the word-governors, although with a certain risk of imprecision).

---

* This distinction is derived from the fact that VYCHISLIT' is a transform of a member of DC 2.1.3.1 and IZUCHIT' is a transform of a member of DC 2.1.3.3. (See Sec. 2.2. above).

The problem here is essentially complementary to the preceding (iii), where the dissimilarity of words in a DC was caused by their different combinatorial properties. In this instance, the dissimilarity is due to the multivalence of the word in the primary category. The complications and correction procedures would appear to be the same.

(v) Two-term combinations are often an inadequate basis for studying word correlation; other contextual items will also have to be considered.

Preliminary study has already established this point. In at least two instances we have, in fact, been dealing with three-term combinations: (a) animate subject + DC 1 verb + noun dependents of the verb, and (b) noun governors + quantifiers (DC 2.1.3) + DC 2 (genitive case). (Sec.2.2 above). In the latter case, the quantifier may be considered a "neutral" element in some combinations *(emission of a number of mesons),* and non-neutral in others *(determination of the number of mesons).* The syntactic role of "number" may, in some sense, be cancelled out by its semantic role. Certainly the inter-relation of the different members of a multiple-term combination must be considered in the formulation of distributional classes.

The foregoing discussion has brought to light more problems than answers. In seeking for answers to procedural problems, we would postulate some kind of priority for the following:

(1) Studies should be made of the effect of DC's on choice of English equivalent during the machine translation process. This information is available for automatic analysis. For example, among the noun governors of DC 2 (genitive case) there are ten multiple-equivalent words whose translation was invariant in combination with DC 2 (96 occurrences). Again, the translation of DC l.l verbs (which may or may not take an animate subject) in some instances depends directly upon the animacy of the subject. The meaning of facts such as these is obscure.

(2) Tests should be made of procedures for establishing "association co-efficients", using appropriate criteria.

(3) The advantages and disadvantages of using *a priori* classes as a check against DC's should be determined.

(4) Frequency studies should be made to determine the predictability of combinations of words or of word classes in new text.