25

1965 International Conference on Computational Linguistics

ON A CERTAIN DISTRIBUTION OF SEMANTIC UNITS

Wojciech Skalmowski

Department of General Linguistics
Jagellonian University
35, Krupnicza
Kraków, Poland

SUMMARY. A remarkable regularity of distribution of Arabic verbal roots in the vocabulary is shown to exist. Presented results suggest that similar regular distributions of semantic units in other languages may be found with the help of word formation rules and vocabulary statistics. Possible applications in approaching the problem of "true" multiple meaning in MT are being discussed.

The notion of "semantic unit" may be formulated in several ways /1/ so that the application of this term makes its explicit definition indispensable. It seems that difficulties in defining it arise from the fact that like most general terms it should be related to some definite theory. At present we do not possess any sufficiently strong and general theory of the semantics of natural languages, though important preliminary steps in this direction have already been made /2/. For this reason most semantic investigations of natural languages still preserve the "artisanlike" character stressed by M.Coyoud and all definitions of the semantic notions remain rather tentative - as well as all the more general conclusions drawn from such investigations. This, too, holds true for the present contribution, in which an empirical fact is described and some remarks on its possible applications to the problem of the "true" multiple meaning have been made.

For this paper it seems advisable to hold apart two notions: that of the "concept" and that of the "semantic unit". Given a generative descriptional device G /grammar/ and a projective system of the type proposed by Katz and Fodor S /semantics/ we can describe a semantic concept in a language L as a set of n-tuples of symbols from G and S, ordered or partially ordered by the relations which define the formal rules of these systems, and having a common derivation in S. This broad frame allows us to regard as a concept every dictionary entry - except for the "grammatical words" which do not possess any derivations in S - and leaves us a wide margin of freedeom in constructing arbitraty "concept -systems" with a priori established features.

In a similar way we may describe a semantic unit as a set of n-tuples of G-symbols, G-rules of word formation and S-symbols, ordered or partially ordered by means of relations which define the formal rules of these systems, and having a common derivation in G from some G-symbol uniquely related to some S-symbol. This allows us to relate with the notion of a semantic unit the linguistic notions of morpheme /or more strictly: semanteme/ and of "word family", defined in terms of grammatical derivations.

The thesauric approach to the problem of meaning in MT /s.e.g.3/ pays tribut to the idea of ordering the symbols within the concepts, but at the same time it brings to light the problem of multiple meaning. This problem has been much discussed already /s.e.g.4/, but it is still far from being solved in all its aspects. Generally speaking the main difficul-

ty arises from the fact that the "concept-systems" of languages are not isomorphic and even if we manage to bring them closer together there remains some amount of "looseness" within the concepts themselves, giving rise to the problem of "true" multiple meaning. The "contextual" multiple meaning may be resolved - in principle, at least - by extending the notion of concepts both in the source and in the target languages to whole sentences or even larger utterances; this is allowed by our " broad" treatment of this notion, not specifying the maximal size of the n-tuples of symbols. By this extension the inner structure of concepts makes the relations defining the isomorphism of the "concept -systems" more apparent; thus even such cases as the adequate translation of the Russian *изменение* as the English "changing /the order of integration/"and "varying /argument/" are theoretically resolvable. Yet there exist instances where the extension of concept would have to go beyond limits and to involve the whole language: these are cases of "stylistic" difference in which there are not apparent reasons for choosing one of the possible synonyms instead of the other but where the difference is distinctly felt by competent bilingual speakers. The problem is important for the translation of literary pieces, especially poetry; by the present stand of MT it is still an "academic" problem, of course, but it exists after all. It may be best illustrated by the question whether there are "better" and "worse" translations of nonsensical expressions, such as the famous "furiously sleeping ideas". Negative answer would mean that every translation is equally good, which in turn would mean that only "meaningful" sentences are translatable; in that case

the MT problems would be "enriched" with the whole load of philosophical questions - an embarassing development, certainly.

Vaguely felt differences between the intrinsic "semantic values" of different elements of language have given rise to the notions of "size" or "content" of semantic elements /5/ and several attempts - both to define these notions and to furnish models of the underlying mechanism have been made /5,6/. The main assumption - based on observations of Willis - was that there existed a "natural hierarchy" of concepts in natural languages, forming a tree or at least a lattice with some definite statistical properties.

The present paper gives some results of an investigation undertaken in order to test this hypotheses. Because of the marvelous clarity of the grammatical structure Arabic has been chosen as a "laboratory example". About 90% of Arabic semantemes are verbal roots, with very few exceptions consisting of three consonants C_1-C_2-C_3; the usual dictionary form is the 3^d pers. sg. masc. perf. of the form $C_1aC_2aC_3a$, s.g. kasara "to break" /lit. "he has broken"/. There are more than ten different verbal stem-patterns i.e. word formation rules, modyfying the basic meaning of the root in a specific way; thus the stem-pattern II: $C_1aC_2C_2aC_3a$ adds to the basic meaning the shade of intensity, e.g. kasara "to break" kassara "to smash"; the stem-pattern III is conative, the IV - causative, etc.

All the triliteral verbal roots in the Arabic vocabulary have been divided into separate classes according to their ability to form s = 1,2,..., n different stems. As only

the number of stem-patterns was considered and further applicable word formation rules /substantivisations, adjectivisations etc./ were disregarded this classification is a very rough approximation to the hypothetical underlying hierarchy. It has been assumed that the number of stem-patterns defining a given class may be approximately viewed as an exponent of the "content" or "semantic value" of the semantic units belonging to this class and that - if the hypothetical hierarchy was really based on this principle - the number of roots with greater s should be smaller than that with smaller s. Baranov's Arabic-Russian Dictionary /7/ has been used for counting the roots and it has been found that the relation between s /the number of stem--patterns characterizing the given class/ and r /the number of roots belonging to this class/ was not only inversly proportional but also nearly functional and that the distribution of roots in the Arabic vocabulary may be described as a simple function $r/s/ = N/As^2 + Bs + C/$, where N is the sum-total of roots and A, B and C are specific constants. The goodness of fit has been tested by the chi-square distribution and it has been found that the differences between the empirical data and the theoretical distribution - except for one value - do not exceed 0.3 significance level.

In order to estimate the possible differences between particular dictionaries - which could arise from differences between the materials used for their compilation - two samples of ca. 700 items each have been taken from two different dictionaries /7,8/ and the distribution of roots in them compared with each other and with the over-all distribution.

All the distributions show a striking similarity, rendering nearly identical chi-square values.x/
This result is a strong argument for the general validity of the discussed distribution in Arabic - and this fact in its turn speaks in favour of the existence of "natural hierarchies" of the semantic units in general.

x/ The figures are as follows:

s / r	1	2	3	4	5	6	7	8	9	N
Baranov's Dictionary	988	714	586	411	254	154	74	18	10	3209
theoretical distribution	974	754	561	398	262	155	76	26	4	
sample /Baranov/	213	163	131	99	55	25	18	3	1	708
sample /Wehr/	229	163	117	95	50	26	13	3	1	697

The constants for Baranov's Dictionary are:

A = 0.004419 , B = 0.082 , C = 0.3812

It seems very probable that similar regular distributions might be found in other languages, too - perhaps the ensemble of the "semantic parameters" would have to be much wider and the "trial and error" investigations would require more time but the whole work can be easily mechanised. The idea of interconnections between the syntactic and semantic structures of language is not new in structural linguistics /s.9 and 10/ and investigations along these lines have already been led in the domain of computational linguistics under direction of P.Garvin /11/. My suggestions go towards discovering such regular

distributions which would facilitate the task of finding more strict correlations between the synonyms within particular concepts on computational basis. The underlying assumption is that the "universes of discours" in various languages are of about the same "size" /whatever it would mean - but such an assumption is tacitely made in every translation/, and that the semantic units underlying the components of concepts are ordered according to their "content", so that the problem of "true" multiple meaning in certain cases may be solved by means of matching the components of concepts of the source and target languages on the basis of their "semantic value".

As an illustration let us consider a few equivalent English verbs in two different translations /A. -12, N. -13/ of the Koranic Sura 84, being translations of Arabic verbs derived from roots all belonging to the same class /5 stem -patterns/, i.e. according to our assumption having about the same "semantic value". The "value" of corresponding English verbs has been tentatively estimated by the number of different sub-entries in Chambers's XXth Century Dictionary /numbers in brackets/:

Arabic	English	
infatara	/A./to split	/16/
	/N./to severe	/3/
garra	to deceive	/5/
	to beguile	/4/
sawiya	to shape	/18/
	to fashion	/11/
sala	to roast	/9/
	to burn	/30/

The applied "method" being unsystematic and ad hoc the example allows no generalisations but it may illustrate our argument that the problem of "true" multiple meaning arises in cases of "expressive language" from the fact that even when the concepts of source and target languages agree there is no correlation between their respective components except for differences between their "value", based on differences on the paradigmatic level. Thus e.g. for the concept "applying heat on something" two different semantic units could have been arbitrarily chosen by the two interpreters, as they regarded th subsets of synonyms within the concepts as unordered. My suggestion is that these subsets might be at least partially ordered by means of the intrinsic value of the semantic units underlying them and that correlations between them might be established in more objective terms of numeric measures of their content.

References

/1/ Coyaud M. - Quelques problèmes de construction d'un "langag formalisé sémantique". La Traduction Automatique 1963 fasc.2

/2/ Katz J.J., Fodor J.A. - The structure of a semantic theory. Language 39/2/,1963

/3/ Sparck-Jones K. - Mechanised semantic classification. 1961 International Conference on Mech.Transl. and Applied Language Analysis. London 1962. Vol.II

/4/ Janiotis A., Josselson H.H. - Multiple Meaning in Machine Translation. ibid.

/5/ Herdan G. - Type-Token Mathematics, Mouton et Co. The Hague 1960

/6/ Mandelbrot B. - On the Language of Taxonomy: an Outline of a "Thermostatistical" Theory of Systems of Categories with Willis /Natural/ Structure. Information Theory, ed. G.Cherry,London 1956

/7/ Baranov X.K. - Arabsko-russkij slovar, /2d ed./, Moskwa 1958

/8/ Wehr H. - Arabisches Wörterbuch für die Schriftsprache der Gegenwart. O.Harrasowitz, Leipzig 1952

/9/ Kuryłowicz J. - Dérivation lexicale et dérivation syntaxique. /Contribution à la théorie des parties du discours/. Bull. de la Soc. de Linguistique de Paris, Vol.XXXVII, 1936

/10/ Kuryłowicz J. - Zametki o značenii slova. Voprosy Jazykoznanija, 1955, No 3.

/11/ Swanson D.R. - The Nature of Multiple Meaning. Proceedings of the National Symposium on MT /Los Angeles 1960/, ed. H.P.Edmundson

/12/ Arberry A.J. - The Koran Interpreted. Oxford Unv. Press 1964

/13/ Nicholson R.A. - A Literary History of the Arabs. The Cambridge Unv. Press 1907