N. A. Stokolova
Philadelphia, Pa.

Syntactic Tools and Semantic Power of Information Languages (Pt. II of 'Elements of a Semantic Theory of Information Retrieval')

Stokolova, N. A.: **Syntactic tools and semantic power of information languages.** Pt. II of 'Elements of a semantic theory of information retrieval'. In: Intern. Classificat. 3 (1976) No. 2, p. 75–81
Different kinds of syntactic tools of information languages (IL) in use, considered as meaning-distinguishing tools, are described as simplified forms of some initial IL grammar tools called 'standard phrases' which are n-place relational predicates of a special kind. A quantitative evaluation is attempted of the effects which the idiosyncracies of the syntactic tools of IL's have on their semantic power. (Author)

1. Introduction

The primary aim of this study is to suggest a formal definition ("explication") of *"relevance relationship"* between texts, including the explication of the concept of *"degree of relevance"*. In Part I* a set T of natural language texts of documents and requests dealt with in an IRS and an orientated bigraph representing the *"strict relevance"* relations (corresponding to the semantic inference relationships) on this set T were considered.

The proposed explication of degree of relevance makes possible the algorithmic completion of this bigraph by relevance relations of different degrees; a formula for calculating the values of *"coefficient of relevance"* was presented.

This coefficient was introduced as a quantitative measure of probable relevance of one elementary text (t_p) to another one (t_s) and was defined as the ratio of the number of all texts from T which are strictly relevant to both t_p and t_s to the sum of that number and the number of all texts from T which are strictly relevant only to t_p but not to t_s.

The concepts of information language (IL), its vocabulary and syntax and the notion of the *"semantic power"* of an IL were defined. The latter concept was defined as the number of non-synonymous natural language expressions which this IL can express and distinguish; a natural language expression being expressed by IL if it has nonempty translation into the IL; any two expressions being distinguished by IL if they have two different translations into this IL.

The above-mentioned bigraph of relevance relations was considered as a model of an ideally functioning IRS; the function of a real IRS, dealing with a given set T, was seen to be the algorithmic reproduction of the bigraph of relevance relations on T by processing the indexes of documents and requests. Two different kinds of deviations from the ideal bigraph, possible in a real IRS, were considered.

The further purpose is to study the role of the different semantic components of an IRS (indexing rules, information language, paradigmatic tools) in the algorithmic reproduction of the bigraph of relevance and particularly in producing and avoiding the different kinds of deviations from the ideal model. In this Part II the main attention is paid to the role of the syntax of the IL; the role of the paradigmatic tools will be considered in Part III.

The final result of the study (presented in Parts II and III) is a procedure proposed for the choice of the semantic components of an IRS suitable for the achievement of some predetermined level of its performance (corresponding to the predetermined level of deviation from the ideal bigraph of relevance relations).

2. Grammar Tools of Information Languages, their Kinds and Functions

In Part I we mentioned only the simplest syntactic tool of a post-coordinate type IL, the constructing of indexes by simple coordination of desriptors, i. e. listing (in an arbitrary order) all the descriptors corresponding to the keywords of the indexed natural language text.

But, as experience confirms, in some subject fields it proves to be insufficient to use such simple syntax to meet the fundamental requirements of an IL: there are such texts in T, the set of natural language texts for which IL is constructed, which are not mutually relevant but nevertheless are represented by identical descriptor sets which contradict the fundamental requirements to be met by the IL. This kind of deviation from the ideal IL we called (in Part I) *"cohesion"*. In order to avoid such cohesion, it is necessary to use in ILs more complex syntactical tools, which we will call *"IL grammar tools"*.

As was noted in Part I the syntax of a post-coordinate type IL is the set of rules for constructing the expressions of the language from its lexical units, which are the descriptors.

The role of IL grammar tools is analogous to the role of natural language grammar, by which natural language expressions – sentences – are built up from meaningful words. Using natural language grammar, different sentences – with different meanings – can be built up from the same set of meaningful words. Similarly using IL grammar tools, different IL expressions can be constructed from the same set of descriptors.

Two ILs with the same keyword sets and vocabulary but with different syntax would be capable to express the same texts of a file: if some text had non-empty translation into the first IL this text has to contain at least one keyword u_i of this IL thesaurus; then this text would have non-empty translation into the second IL also as this IL's thesaurus contains u_i also. Differences in syntactic tools of these two ILs would have influence only upon the different cohesion levels of this file; so the syntactic tools prove to be the meaning distinguishing tools of the ILs.

* Not as yet published. Full reference will be given in Part III, forthcoming in this journal.

The author's experience of IL development in such subject fields as organic chemistry, biology, geology (1–6) has indicated that for detailed descriptions of the meanings of texts in these fields the IL grammar tools called *"standard phrases"* which are n-place relational predicates of a special kind are most appropriate.

Analysis of pertinent literature has shown that IL grammar tools widely in use may be described as different kinds of simplification of some ILs of the "standard phrase" type. Such ILs in their turn are simplified variants of some basic *"language of meaning"*, explicitly displaying the semantics of natural language. The study and construction of such "semantic languages" is one of the principal problems of contemporary computational linguistics (7–11).

According to some of these studies (12–16) and to the accumulated experience of mathematical logic, the language of predicate calculus is the most obvious model for a language of meaning.

By widely used grammar tools of ILs we mean *"links"* (17–21), *"roles"* (17–22), binary predicates (23, 24) and special descriptors of predicative nature which were called in a previous study (25, 26) – *"aspect descriptors"*.

In order to clarify the nature of these different grammar tools, we will consider examples of standard phrases (used in an IL for chemistry) and will show what other kinds of grammar tools may be obtained by the simplification of these standard phrases.

We shall consider two examples of standard phrases (cited here in simplified versions) which enable us to describe such extralinguistic situations typical for chemistry as those of performing a chemical reaction (1) or the purification of a substance (2).

"Substance x, *the agent of chemical reaction of type* y, *reacts with substance* z, *to yield the main products* u *and* v *of this reaction and the by-product* r *with the substance* w *used as catalyst and substance* q *as solvent."* (1)

"The purification of substance x *from impurity* z *is accomplished by treatment with solvent* u*."* (2).

As one can see from these examples, standard phrases are semantically standardized sentence schemes containing some variables (denoted by x, y, z, u etc.); substitution of all (or some) of these variables by descriptor yields meaningful expressions of IL, whose evident interpretations are sentences in natural language. (When some variables are not replaced by descriptor it is meant that they are bound by existential quantifiers.)

In standard phrase ILs the following supplementary semantic tools are used:

1) the anaphoric connections are indicated for revealing the identity of two or more objects, denoted in different sentences by different or even the same generic names (descriptors);
2) using sentences (i. e. already substituted predicates) as possible values of variables (besides descriptors) in other predicates (This technique corresponds to the linguistic process of *"insertion"* (10)).

The more important "roles" (role indicators to be assigned to descriptors), recommended in the "Thesaurus of Engineering Terms", are equivalent to the indication of those variables which these descriptors would substitude for in the two standard phrases above.

As examples, consider the following "roles":
the "role" 1 – raw material ((1) x, (1) z, (2) x)

(We indicated here and below the corresponding variables in the foregoing phrases by means of the number of the phrase – (1) or (2) – and the variable letters in this phrase – x, y, z etc.).

the "role" 2 – product, by-product ((1) u, (1) v, (1) r, (2 x))
the "role" 5 – enviromnent, solvent ((1) w, (1) g, (2) u)
the "role" 3 – impurity ((2) z)

As can be seen from these examples these "roles" don't make any difference between some different variables of these standard phrases.

The facet formula suggested by Vickery (27) contains in part the category names, i. e. the names of broad classes to which the terms always belong, due to their inherent properties. Such, for example, are the following variables in this formula: *"P – substance, product, organism"* (here *"substance" and "organism"* are category names unlike *"product"* which is the denomination of a syntactical role indicator, insofar as it is the name of a class to which terms belong depending upon context: in some contexts a substance may be a product, in another, a raw material), *"Q – property and measure", "E – action, operation, process, behaviour".*

Other places in this facet formula are just context dependent role indicators, which correspond to some variables in the above-mentioned standard phrases. Such, for example, are the following items of variables:

"C – constituent"
"R – object of action, raw material"
"A – agent, tool"

"MEDLARS" Index Language Subheadings (28, p. 129) such as *"Analysis", "Chemical Synthesis", "Chemically Induced", Occurrence", "Prevention and Control" "Utilization"* correspond to some items in standard phrases for chemistry, other subheadings may be considered as the names of items in other standard phrases, corresponding to significant extralinguistic situations in the subject fields covered by "MEDLARS". Insofar as these subheadings are used in pairs with descriptors, they are equivalent to such syntactic terms as "roles".

The functions of keywords used in multi-word combinations, are like the functions of some IL grammar tools. For example in the case of the absence in an IL of such a predicate (or of the corresponding "role") as *"Material* x *has the property* y" (this "role" is used particularly in the Semantic Code Language (29, 30)) the corresponding meanings are described by keywords assembled into multi-word combinations, *"some material with such and such property"*. For example, the following word combinations: *"elastic materials", "electroconductive plastics", "antispasmolitic substances"*, or *"refractory (building materials)"* and alike.

In spite of the usefulness of grammar tools in avoiding the cohesion of meanings, such tools increase IRS operational costs, due to complications in the search algorithm and indexing procedures. Besides that, the usage of com-

mon and rather ambiguous grammar tools inevitably causes ambiguous indexing and hence the decrease of the recall ratio.

Aitchison and Gilchrist (31) noted that "links" and "roles" are "precision devices which, except in certain subject areas, are likely to be detrimental to recall. The reasons for this are clear:

1. It is difficult for indexers to apply the roles consistently.
2. It is even more difficult for the searcher to match the use of role by the indexer. . .
3. But it is not only the ambiguity of the roles which complicates searching, it is also the fact that the searcher is in ignorance of the interrelationship of terms in the index, when roles may be affected by the existence of unknown terms not featured in the terms of the question."

Van Oot et al. (32) investigated the influence of "roles" upon IRS performance. They found that the absence of mutual exclusiveness of "roles" causes indexing ambiguity.

In several studies it was shown that the usefulness of "roles" and "links" changes substantially with the subject field. Montague (33) and Van Oot et al. (32) noted that it is useful to apply "roles" for describing chemicals in reactions and processes. On the other hand, for describing documents at the Air Force Material Laboratory, Sinnet (34) notes that it is more useful to apply "links" than "roles".

Montague (33) asserts that "roles" may be effective in certain fields but not in others. She confirms that for giving real effectiveness "roles" should be capable of precise and unambiguous application. So, in her experiments with chemical requests, recall dropped – due to "roles" application – by only 4 percent, while for non-chemical questions recall dropped by 52 percent.

A special analysis (35) has shown that the usage of such grammar tools as the n-place predicates remarkably increases the semantic power of the IL; nevertheless it is often possible to achieve acceptable precision ratios by using some simplified version of these syntactic tools.

It proves to be useful to develop firstly an n-place predicate syntax for a given representative file T; afterwards each simplified syntactic tool can be interpreted (and so precisely described) by means of this n-place predicate syntax. (As was noted in different already cited studies, precise description of grammar tools are very important for their effectiveness). Moreover, in this case it proves to be possible to choose the appropriate simplified tools on the basis of the investigation of n-place predicates, occuring in the semantically powerful IL_T and of the representations of texts from T by this IL_T. This method of syntax construction will be presented later.

3. Semantic Power of ILs with Different Grammar Tools and a Method of Syntax Construction

It is seen from the foregoing discussion and from opinions of some investigators (presented in paragraph 2) that IL's grammar tools prove to be useful in some cases (particularly in some subject fields), but sometimes their influence on IRS performance is not significant; sometimes their unambiguous application is possible, but in other cases this is difficult, and results in recall ratio decrease.

So it is desirable to have some objective, and, if possible, quantitative criteria concerning the usefulness of IL grammar tools and the measure of this usefulness depending upon the characteristics of a given subject field, the idiosyncracies of texts, etc. Such criteria will allow one to choose suitable grammar tools depending on the one hand, upon these characteristics and on the other hand upon the requirements of the precision ratio desired.

One can see that, in so far as grammar tools are meaning distinguishing tools, the ideal IL grammar for a given file T is such a grammar that makes possible to distinguish all different meanings expressed in texts of T and so to avoid the "cohesion" of texts with different meanings.

To meet this requirement, as experience shows, it is possible to construct IL with an n-place predicate grammar (such as standard phrase grammar) – IL_1 To construct the simplest IL_1 for T, meeting this requirement it is necessary to base this IL upon file T: only such descriptors and predicates are to be included in this IL for which corresponding keywords and semantic relations are contained in texts from T (such an IL, will be denoted by $IL_{1,T}$).

The expressions of $IL_{1,T}$ are unordered sets of statements, each of them being built up by means of one predicate, in which in one or more places the corresponding descriptors occur – descriptors of such categories as are domains of these variables. The number of expressions of such an IL is larger than the number of texts in any document file. So it is obvious that such a semantically powerful IL, as $IL_{1,T}$, would be capable of expressing and distinguishing a variety of meanings, which are absent in file T-*"nonactual meanings"*.

At the same time such a complicated IL, as $IL_{1,T}$, would complicate and make more expensive the corresponding IRS. Besides, the high precision ratio, achievable by using $IL_{1,T}$ may not be required. Hence the optimal IL for T is an IL with the same vocabularly as vocabularly of $IL_{1,T}$ (in order to express all *"actual meanings"* – meanings of texts from T) but with grammar tools which are simpler than the n-place predicate grammar of $IL_{1,T}$ (cohesion of different meanings, including actual meanings is inevitable in this case). So the grammar of an optimal IL should be the simplest one which makes possible the achieving of the required value of the precision ratio.

Hence, the criterion for the choice of grammar tools of an IL for T proves to be the cohesion level of *"actual meanings"* corresponding to a given level of the precision ratio. At the same time the simplified grammar tools of such an optimal IL for T have to be described as simplifications of the n-place predicate grammar of $IL_{1,T}$, because such a precise description of these simplified grammar tools would allow one to avoid their ambiguous application.

In order to obtain the above-mentioned quantitative criterion it is desirable to calculate the semantic power values of ILs with different grammar tools and an identical vocabulary. The value of the semantic power of any IL with a simplified grammar, compared with the semantic power of the corresponding $IL_{1,T}$, characterises the average cohesion level taking place when using this IL. But such average cohesion levels don't reflect precisely enough the "quality" of these grammar tools for a certain

file T. Therefore it is desirable to estimate which grammar tools are more "suitable" for such and such characteristics of a given file T (and by these grammar tools the actual meanings for T must cohere to a lesser degree than the nonactual meanings) and moreover to be able to estimate what are the simplest grammar tools, which would give the highest acceptable cohesion level for actual meanings.

Unlike the above-mentioned average values of the "cohesion" level, the values of "cohesion" level of actual meanings cannot be directly calculated. A better way of estimating the "cohesion" level of actual meanings by using different grammar tools is to investigate the cohesion mechanism in order to estimate what characteristics of texts influence the cohesion level in the cases of different grammar tools.

For all these purposes a special investigation was carried out (35) in which some ILs with different grammar tools and an identical vocabulary were precisely described and compared. For this investigation the following kinds of grammar tools were chosen: the simplest kind of syntax (this is the case of absence of grammar tools); "aspect" descriptors; "roles", "links", binary predicates; n-place predicates (used without the technique of "insertion" and without indicating objects' identity).

The following seven IL types were considered (these ILs were described by means of a generative grammar for $IL_{1,T}$ and by algorithms of translation from $IL_{1,T}$ into each of the six other different ILs.) The $IL_{1,1}$ language using n-place predicate grammar tools was described above.

An example of $IL_{1,T}$ expression consisting of three statements will be:

$P_1^{2,3}(d_{2,4}^1, d_{3,5}^2), P_2^{3,3,3}(d_{3,5}^1, d_{3,12}^2, d_{3,7}^3), P_2^{3,3,3}(d_{3,6}^1, d_{3,10}^3)$ (A), where

$P_1^{2,3}$ – the predicate number one, on the first place of which a descriptor of second category may occur, on the second place – a descriptor of the third category may occur; it is seen that this predicate is two-place one;

$P_2^{3,3,3}$ – the predicate number two in the three places of which the descriptors of the third category, may occur;

$d_{2,4}^1$ – the descriptor which occurs on the first place of the corresponding predicate. It belongs to the second descriptor category and has the number four within this category;

$d_{3,10}^3$ – the descriptor which occurs on the third place of the corresponding predicate, it belongs to the third descriptor category and has the number ten within this category.

The IL simultaneously using "roles" and "links" – $IL_{2,T}$ is constructed in such a way that for each place of each predicate of $IL_{1,T}$ there is a corresponding "role" and for each statement of $IL_{1,T}$, formed by a single predicate, there is one "link" in the corresponding expression of $IL_{2,T}$.

Each "link" of $IL_{2,T}$ consists of an unordered sequence of descriptors, each provided with a "role" indicator. This "role" indicates the place which the descriptor occupies in the corresponding predicate of $IL_{1,T}$.

An example of $IL_{2,T}$ expression, which is the translation of the foregoing example of $IL_{1,T}$ expression, will be:

$(p_{1,1}d_{2,4}, p_{1,2}d_{3,5}), (p_{2,1}d_{3,5}, p_{2,2}d_{3,12}, p_{2,3}d_{3,7}), (p_{2,1}d_{3,6}, p_{2,3}d_{3,10})$ (B), where

$p_{1,1}$ – a "role" which corresponds to the first place of predicate number one (in $IL_{1,T}$);

$p_{1,2}$ – a "role" which corresponds to the second place of predicate number one (in $IL_{1,T}$);

$d_{2,4}$ – the descriptor which belongs to the second descriptor category and has number four within this category;

$p_{1,1}d_{2,4}$ – a "roleterm" which corresponds to the predicate number one in $IL_{1,T}$ on the first place of which the descriptor $d_{2,4}$ occurs.

Each bracket in this expression (a "link") corresponds to one statement in the foregoing expression of $IL_{1,T}$.

Another IL – $IL_{3,T}$ – simultaneously using "aspect" descriptor and "links" is constructed in such a way, that for each "role" of $IL_{2,T}$, there is a corresponding "aspect" descriptor and for each statement of $IL_{1,T}$ engendered by a single predicate, there is one "link" in the corresponding expression of $IL_{3,T}$; each "link" of $IL_{3,T}$ consists of an unordered sequence of "aspect" descriptors and ordinary descriptors (which we will call "object" descriptors).

An example of $IL_{3,T}$ expression, which is the translation of the foregoing example of $IL_{1,T}$ expression, will be:

$(p_{1,1}, p_{1,2}, d_{2,4}, d_{3,5}), (p_{2,1}, p_{2,2}, p_{2,3}, d_{3,5}, d_{3,7}, d_{3,12}), (p_{2,1}, p_{2,3}, d_{3,6}, d_{3,10})$ (C), where

$p_{1,1}, p_{1,2}, \ldots, p_{2,3}$ – aspect descriptors, deriving from the corresponding "roles" of $IL_{2,T}$;

$d_{2,4}, d_{3,5}, \ldots, d_{3,12}$ – object descriptors, deriving from the descriptors of $IL_{1,T}$ and $IL_{2,T}$.

The IL, using binary predicates – $IL_{4,T}$ – is constructed in such a way, that for each of two places of each predicate of $IL_{1,T}$ in $IL_{4,T}$ a corresponding binary predicate is chosen; the expressions of $IL_{4,T}$ are unordered sequences of assertions, formed of these binary predicates (corresponding to "syntagms").

An example of $IL_{4,T}$ expression, which is the translation of the foregoing example of $IL_{1,T}$ expression will be:

$p_1^{1,2}(d_{2,4}, d_{3,5}), p_2^{1,2}(d_{3,5}, d_{3,12}), p_2^{1,3}(d_{3,5}, d_{3,7}), p_2^{1,3}(d_{3,6}, d_{3,10}), p_2^{2,3}(d_{3,12}, d_{3,7})$ (D), where

$p_1^{1,2}$ – a binary predicate, corresponding to the first and second places of the first predicate of $IL_{1,T}$;

$p_2^{1,2}$ – a binary predicate, corresponding to the first and the second places of the second predicate of $IL_{1,T}$;

$d_{2,4}, d_{3,5}$ – descriptors.

The IL using only "roles" – $IL_{5,T}$ – is constructed in such a way that its "roles" coincide with the "roles" of $IL_{2,T}$, and the expressions of $IL_{5,T}$ are unordered sequences of descriptor – "role" pairs *("roleterms")*.

An example of $IL_{5,T}$ expression, which is the translation of the foregoing example of $IL_{1,T}$ expression will be:

$p_{1,1}d_{2,4}, p_{1,2}d_{3,5}, p_{2,1}d_{3,5}, p_{2,1}d_{3,6}, p_{2,2}d_{3,12}, p_{2,3}d_{3,7}, p_{2,3}d_{3,10}$ (E), where

$p_{1,1}, p_{1,2}, \ldots, p_{2,3}$ – "roles", coinciding with the "roles" of $IL_{2,T}$;

$d_{2,4}, d_{3,5}, \ldots, d_{3,12}$ – descriptors.

The IL using "aspect" descriptors – $IL_{6,T}$ – is constructed in such a way that its aspect descriptors coincide with the "aspect" descriptors of $IL_{3,T}$. The expressions of $IL_{6,T}$ are unordered sequences of "aspect" and "object" descriptors.

An example of $IL_{6,T}$ expression, which is the translation of the foregoing example of $IL_{1,T}$ expression will be:

$p_{1,1}, p_{1,2}, p_{2,1}, p_{2,2}, p_{2,3}, d_{2,4}, d_{3,5}, d_{3,6}, d_{3,7}, d_{3,10}, d_{3,12}$ (F), where

$p_{1,1}, p_{1,2}, \ldots, p_{2,3}$ – "aspect" descriptors.

$d_{2,4}, d_{3,5} \ldots, d_{3,10}$ – "object" descriptors.

The $IL_{7,T}$ is the IL without grammar; its expressions are unordered sequences of "object" descriptors.

An example of $IL_{7,T}$ expression, which is the translation of the foregoing example of $IL_{1,T}$ expression will be:

$d_{2,4}, d_{3,5}, d_{3,6}, d_{3,7}, d_{3,10}, d_{3,12}$. (G)

Let us give examples of interpretation for some of the above-mentioned formal expressions. Let $P_1^{2,3}$ denote the predicate: " A chemical reaction of type x is accomplished to yeald the main product y" and let $P_2^{3,3,3}$ denote the predicate: "The purification of substance x from impurity y is accomplished by treatment with solvent z."[1]

Let $d_{2,4}$ denote "reduction reaction"
$d_{3,5}$ – "Aniline"
$d_{3,12}$ – "Nitrobenzene"
$d_{3,7}$ – "Hydrochloric acid"
$d_{3,6}$ – "Toluidine"
$d_{3,10}$ – "Sulphuric acid"

Then the above-mentioned formal expression of $IL_{1,T}$ (A) has the following interpretation (I): "Aniline is produced by a reduction reaction; aniline is purified from nitrobenzene by treatment with hydrochloric acid; toluidine is purified from some impurity by treatment with sulphuric acid."

Then "roles" used in $IL_{5,T}$ are the following:

$p_{1,1}$ – x is the chemical reaction"
$p_{1,2}$ – x is the main product of chemical reaction"
$p_{2,1}$ – x is the substance which is purified"
$p_{2,2}$ – x is the impurity which another substance is purified from".
$p_{2,3}$ – x is the solvent by which a purification procedure is accomplished".

Then the above-mentioned formal expression of $IL_{5,T}$ (E) may have beyond the interpretation (I) the following interpretation (II): "Aniline is produced by reduction reaction; toluidine is purified from nitrobenzene by treatment with hydrochloric acid; aniline is purified from some impurity by treatment with sulphuric acid."

It is seen that there are many other different interpretations of this expression of $IL_{5,T}$ which do not coincide with interpretation (I). All these natural language expressions are cohered when using $IL_{5,T}$. If some of these different interpretations are included in T ("actual meanings") this cohesion will inevitably cause decrease of the precision ratio.

1) $P_2^{3,3,3}$ corresponds to the standard phrase (2), see page 76, and $P_1^{2,3}$ is a simplified version of the standard phrase (1), see page 76.

The above-mentioned formal expression of $IL_{7,T}$(G) may have – beyond all interpretations of expression of $IL_{5,T}$ – a lot of other different interpretations, for example the following one (III): "Nitrobenzene is produced by reduction reaction; toluidine is purified from some impurity by treatment with hydrochloric acid; aniline is purified from toluidine by treatment with sulphuric acid." This increase of cohesion level will cause decrease of precision ratio.

It is possible to consider that each expression of such a semantically powerful IL as $IL_{1,T}$ is meaningful. Each expression of some IL of the group $IL_{2,T} - IL_{7,T}$ is obtained from an expression of $IL_{1,T}$ (or from several expressions of $IL_{1,T}$, which cohere by the translation into this IL of the group) and therefore the interpretation of each such expression is the interpretation of this original expression of $IL_{1,T}$ (or is the set of interpretations of these several original expressions of $IL_{1,T}$) and so each such expression is meaningful. Therefore the total number of expressions of any IL of this group as well as of $IL_{1,T}$, is equal to its semantic power. These values were calculated. Formulae were obtained expressing these numbers as functions of different characteristics of $IL_{1,T}$, such as the total number of predicates and descriptors in $IL_{1,T}$, the number of descriptor categories, the number of descriptors in these categories, the number of different variables in $IL_{1,T}$, predicates, whose domain is the same descriptor category and others. We shall give an example of calculation of semantic power of one IL from this group – $IL_{5,T}$.

Let us denote the number of different categories of descriptors in $IL_{1,T}$ – N, the category i containing m_i descriptors, category j – m_j descriptors etc. In predicates of $IL_{1,T}$ there are v_i variables, whose domain is category i, v_j variables, whose domain is category j ,. . . , v_n variables whose domain is category n, (respectively in $IL_{2,T}$ there are $v_i, v_j, \ldots, v_n$ roles which can be "combined" with descriptors of the categories $i, j, \ldots, n$).

Then the number of well formed expressions of $IL_{5,T}$ can be calculated in the following way. The total number of the role-terms (the pairs consisting of concrete descriptors of category i – $p_i = m_i \cdot v_i$. The total number of expressions of $IL_{5,T}$ containing only the descriptors of category i – $C_i = 2\,p_i - 1$. The total number of expressions containing only the descriptors of one category

$$= \sum_{i=1}^{N} C_i.$$

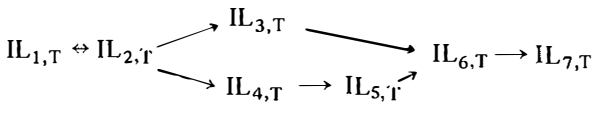
The total number of expressions of $IL_{5,T}$ containing the descriptors of two categories

$$= \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} C_i \cdot C_j.$$

The total number of expressions of $IL_{5,T}$ =

$$\sum_{i=1}^{N} C_i + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} C_i \cdot C_j + \sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \sum_{m=j+1}^{N} C_i \cdot C_j \cdot C_m + \ldots + (C_1 \cdot C_2 \cdot \ldots \cdot C_N)$$

The results of the calculations of semantic power of languages $IL_{1,T} - IL_{7,T}$ are outlined by the following scheme:



where "$\leftrightarrow$" denotes equality of semantic powers and "$\longrightarrow$" denotes decrease of semantic power.

The order of decreasing of semantic power, indicated by this scheme has to correspond to the order of decreasing precision ratios, which the use of these ILs will yield. (By stating this we assume that the average level of cohesion, reflected in the values of the semantic power of the IL is at the same time the average level of cohesion of actual meanings).

For the estimation of the "quality" of these different ILs for a given file T (their capability to distinguish the actual meanings) the mechanism of cohesion while translating from one IL of this group $IL_{q,T}$ into another $IL_{p,T}$ (with less semantic power) was investigated and the dependence was established of the measure of this cohesion upon characteristics of the expressions of $IL_{p,T}$. These peculiarities are expressed in their turn as the characteristics of such expressions of $IL_{1,T}$, from which coherent expressions are derived. The results of quantitative investigation of the cohesion mechanism were expressed by algebraic formulae. From these formulae it is possible to indicate such numerical characteristics of an $IL_{p,T}$ expression, which determines the number of different $IL_{q,T}$ expressions, cohering into this $IL_{p,T}$ expression by translation from $IL_{q,T}$ into $IL_{p,T}$. These formulae allow to appreciate quantitatively the level of this cohering if values of characteristics mentioned above are known.

We shall give an example of deduction of one such formula.

Each expression of $IL_{3,T}$ can correspond to several expressions of $IL_{2,T}$: the average cohesion level for $IL_{3,T}$, as seen from the foregoing scheme, is higher than for $IL_{2,T}$.

For each bracket i n $IL_{3,T}$ (for example the m^{th}), containing r groups of aspect descriptors, each of them (for example the i^{th}) contains L_i such aspect descriptors, which can be combined with the descriptors of the same category (the number of descriptors of this category in this bracket is K_i), the number of different possible sets of role-terms in group i of bracket m $- S_{m,1}$ - (if $L_i > 1$ and $K_i > 1$) is equal to the number of different partitions of L_i different objects into K_i "non-empty" groups i.e.(36).

$$S_{m,i} = K_i - C^1_{K_i}(K_i - 1) + C^2(K_i - 2) + \ldots + (-1)^{K_i - 1} \cdot C^{K_i - 1}_{K_i} \cdot 1.$$

If in bracket m there are r groups of aspect descriptors (as group i) then the number of brackets in $IL_{2,T}$, generated by bracket m in $IL_{3,T}$ will be $S_m = S_{m,1} \cdot S_{m,2} \cdot \ldots \cdot S_{m,r}$ where
$S_{m,1}, S_{m,2} \ldots, S_{m,r}$ are calculated as $S_{m,i}$.

If in this expression in $IL_{1,3}$ there are q brackets (such as bracket m), the number of expressions $IL_{2,T}$ generated by such expression of $IL_{3,T}$ will be $S = S_1 \cdot S_2 \cdot S_q$, where $S_1, S_2, \ldots, S_q$ are calculated as S_m. So the number of expressions in $IL_{2,T}$, generated by one expression in $IL_{3,T}$ depends upon the quantity of L,K in each group of the brackets in expression in $IL_{3,T}$, upon the quantity of r in each bracket of this expression and upon the number of brackets q. The corresponding correlation is expressed by the foreging formulae.

So by approximately estimating the average values of these characteristics, estimated for texts of T (or, what is easier, for their translations into $IL_{1,T}$) it is possible to predict the real average cohesion level produced by translation of these texts from $IL_{q,T}$ into $IL_{p,T}$ and to predict in such a way the ratio of precision values by using $IL_{q,T}$ and $IL_{p,T}$ for a given representative file T.

The above-mentioned formulae are obtained for the following pairs of IL, belonging to the IL group described above: $IL_{2,T} - IL_{3,T}$, $IL_{3,T} - IL_{4,T}$, $IL_{2,T} - IL_{4,T}$, $IL_{4,T}$, $- IL_{5,T}$, $IL_{5,T} - IL_{6,T}$, $IL_{6,T} - IL_{7,T}$ (35).

The procedure of selecting the optimal simplified grammar tools (in the above-mentioned sense) for a given file T can be drawn from the results of this investigation. The $IL_{1,T}$ with its tools and the representations of texts from T proves to be in this case the precise reflection of the semantic idiosyncracies of texts from T. Such an IL will appear as optimal for the so-called *"fact retrieval"* or *"question answering"* systems (37) i. e. systems that provide a direct answer to a question rather than retrieving a piece of relevant texts. In such systems it is necessary to use IL with high semantic power for the precise descriptions of corresponding facts.

References

(1) Stokolova, N. A.: Meaning – Expressing tools of Information Languages of "Standard Phrases" (In Russian). In: Naučno – techničeskaja Informacija, ser. 2 (1970), No. 6, p. 15–17.

(2) Stokolova, N. A., Tonijan, A. V.: Application of 'Standard Phrases' in Developing a Language for Explanatory Phrases in Chemical Indexes. (In Russian) In: Naučno – Techničeskaja Informacija, ser. 2 (1970), No. 3, pp. 16–25 (A. J. 70.10.84.)

(3) Tkach, S. M.: A New Version of an Information Retrieval Language for Solid State Physics. (In Russian) In: Naučno – Techničeskaja Informacija, ser. 2 (1973), No. 11, p. 19.

(4) Tonijan, A. V., Stokolova, N. A., Ershov, B. B.: On Grammar Tools of Information Retrieval Language for Texts on Chemical Machine-Building. (In Russian) In: Naučnyj Simposium "Semioticheskie problemy Jazykov Nauki, Terminologii, Informatiki" part 2, Izdatelstvo Moskovskogo Universiteta, Moskow 1971. p. 23.

(5) Vasilèva, I. I., Stokolova, N. A.: Development of Information Retrieval Language for Insects Morphology with Application of Technique of "Standard Phrases". (In Russian) In: Naučno – Techničeskaja Informacija, ser. 2 (1970) No. 7, p. 18–27.

(6) Vleduts, G. E., Stokolova, N. A.: About a Method of Constructing Information Languages Having Grammar. (Translated from the Russian by Joe Lineweaver) Bangalore, India; Doc. Res. & Training Center 1974. 29 pp, FID/CR Report No. 13.

(7) Chomsky, N.: Aspects of the theory of syntax. Cambridge, Mass.: MIT Press, Massachusetts Institute of Technology, 1965.

(8) Mel'čuk, I. A.: Experience of the Theory of Linguistical Models "Meaning $\leftrightarrow$ Text". (In Russian) Moscow, Nauka: 1974.

(9) Katz, J. J., Postal, P. M.: An integrated theory of linguistic descriptions. Cambridge, Mass. 1964.

(10) Weinreich, V., On the semantic structure of language. In: Universals of Language. Cambridge, Mass., 1963.

(11) Wierzbicka, A.: Semantic primitives. Frankfurt, 1972.

(12) Bellert, J.: On the use of Linguistic Quantifying Operators in the Logico – Semantic Structure Representation of Utterances. In: Proceedings of the Internation Conference of Computational Linguistics, Sanga – Saby, Sweden 1–4 September 1969.

(13) Kay, M., Su, S.Y.W.: The Mind System: The Structure of the Semantic File. Santa Monica, Calif. RM – 6265/3 – PR 1970. Rand Corporation.

(14) Padučeva, E. V.: The Language of Mathematical Logic as a Semantic Model for Natural Language. In: Social Science Information 7 (1968) pp. 27–39.
(15) Padučeva, E. V.: Semantic Analysis of Natural Language During Translation into the Language of Mathematical Logic. (In Russian) In: Vsesojuznaja Konferencija po Informacionno – Poiskovym Sistemam i Avtomatizirovannoj Obrabotke Naučno – Techniceskoj Informacii. Moscow 1967. p. 156.
(16) Shapiro, S. C., Woodmansee, G. H.: A Net Structure Based on Relational Question Answerer. In: Norton, Walker (Eds.): Proceedings of the International Joint Conference on Artificial Intelligence, MITRE Corporation, Bedford, Mass., 1969. pp. 325–346.
(17) Blagden, J. F.: How Much Noise in a Role-Free and Link-Free Coordinate Indexing System? In: J. of Doc. 22 (1966) p. 203–209.
(18) Cleverdon, C., et al.: Factors Determining the Performance of Indexing Systems. Cranfield, England: College of Aeronautics, A A L I B, Cranfield Research Project, 1966.
(19) Costello, J. C.: Storage and retrieval of chemical research and patent information by links and roles in Du-Pont. In: Amer. Doc. 12 (1961) No. 2 p. 11.
(20) Perry J. W., et al.: Machine Literature Searching. New York: Interscience Publ. 1956.
(21) Costello, J. C., Wall, E.: Recent Improvements in Techniques for Storing and Retrieving Information. Wilmington, Del.: E. J. Du Pont de Nemours & Co. 1959.
(22) Thesaurus of Engineering Terms. New York: Engineers Joint Council 1964.
(23) Gardin, J. C.: Syntol. New Brunswick, N. Y.: Rutgers, the State University 1960.
(24) Skorochodko, E. F.: The Information Retrieval System at the Institute of Cybernetics, Academy of Sciences of the Ukrainian S. S. R. (In Russian). In: International Forum on Informatics, Moscow: VINITI 1969. Vol. II, p. 68–78 (A. J. 70.3.160).
(25) Otradisnkij, V. V., Stokolova, N. A.: On one Methodology of Construction of Information Retrieval Languages without Grammar. (In Russian). In: Naučno-Techničeskaja Informaeija, scr. 2. (1968) No. 6, p. 14–18.
(26) Preisler, W.: Thesaurusarten und Probleme ihrer Strukturierung. In: Informatik 20 (1973) No. 4, pp. 9–16.
(27) Vickery, B. C.: Faceted Classification. London. A S L I B. 1960.
(28) Lancaster, F. W.: Vocabulary Control for Information Retrieval. Washington. Information Resources Press 1972. 233 pp.
(29) Kent, A., Perry, J. W.: Searching Metallurgical Literature. In: Casey R. S., et al.: Punched Cards: Their Application to Science and Industry. 2nd. ed., New York: Reinhold 1958, p. 248–260.
(30) Perry J. W., Kent, A.: Tools for Machine Literature Searching Searching. New York: Interscience Publishers, Inc. 1958.
(31) Aitchison, J., Gilchrist, A.: Thesaurus Construction. A practical manual. London: ASLIB 1972. 95 p.
(32) Van Oot, J. G., et al.: Links and Roles in Coordinate Indexing and Searching: An Economic Study of their Use, and an Evaluation of their Effect on Relevance and Recall. In: J. of Chem. Doc. 6 (1966) p. 95–101.
(33) Montague, B. A.: Testing, Comparison and Evaluation of Recall, Relevance and Cost of Coordinate Indexing with Links and Roles. In: Pro. Amer. Doc. Inst. 1 (1964) p. 357–367.
(34) Sinnett, J. D.: An Evaluation of Links and Roles Used in Information Retrieval. Dayton, Ohio: Air Force Materials Laboratory, Wright-Patterson Air Force Base, 1964. A D 432 198.
(35) Stokolova, N. A.: On a Technique of Evaluation of the Semantical Power of Information Languages with Grammar. (In Russian). Moskow. Naučnyj Sovet po Kompleksnoj probleme "Kibernetika", Vsesojuznyj Seminar po Informacionnym Jazykam, 1971. vyp 3.
(36) Vilenkin, N. Ya. Combinatorics. (In Russian). Moscow: Nauka 1969.
(37) Simmons, R. F.: Natural Language Question-Answering Systems. In: Comm. ACM 13 (1970) No. 1, p. 15–29.

W.-W. Höpker
Pathologisches Institut der Universität, Münster i. W.

Struktur und Kompatibilität des Thesaurus der Medizin (Structure and compatibility of the Thesaurus of Medicine)

Höpker, W.-W.: **Struktur und Kompatibilität des Thesaurus der Medizin.** (Structure and compatibility of the Thesaurus of Medicine) (In German). In: Intern. Classificat. 3 (1976) No. 2, p. 81–84.
The medical thesaurus described is an international, compatible thesaurus in the German language the structure of which is based upon both a hierarchical classification and a variable classification of facets. Both classification principles are shown by three digit letters and figures. The thesaurus is compatible with the Clinical Key of Diagnosis (KDS), the International Classification of Diseases (E) (ICD/E), and the Systematized Nomenclature of Pathology (SNOP). Its volume comprises 22,000 different terms. (Author)

1. Problemstellung

Die differenzierten Ansprüche, welche Benutzer an ein Dokumentationssystem stellen, finden zunächst ihren Niederschlag in einem entsprechenden Schlagwortverzeichnis, Sachkatalog oder – Thesaurus. Wir verstehen unter Thesaurus *ein Klassifikationssystem mit klartextlichem Einstieg* und legen – dem gegenwärtigen Stand automatisierter Dokumentationsvorhaben Rechnung tragend – das Schwergewicht auf Klassifikationssystem. Es zeichnet sich ab, daß sogenannte „Totallösungen" der Dokumentationsfrage in der Medizin immer weiter in die Ferne rücken. Um so mehr muß angestrebt werden, mit den bisher verfügbaren Methoden und mit vertretbaren finanziellen Mitteln pragmatische Teillösungen zu realisieren, die dann als Ausgangspunkt für weitere Entwicklungen angesehen werden können. Wir sind Gegner von perfektionistischen Systemen, die „zu gut" sind, als daß sie funktionieren können. Diese Systeme haben meist noch die Angewohnheit, niemals „fertig und funktionstüchtig" zu werden. Aus diesen Gründen haben wir wenig Scheu, einen (in diesem Sinne nur halb-fertigen) Thesaurus vorzustellen, der noch nicht vollständig ausgetestet ist und noch sicherlich zahlreiche inhaltliche Fehler aufweist. Gesamtkonzept und Struktur haben sich jedoch bereits bewährt.

2. Voraussetzungen

Unsere Arbeitsgruppe ist von folgenden *Voraussetzungen* zur Realisierung eines international kompatiblen Thesaurus ausgegangen: