János S. Petöfi
Universität Bielefeld

Some Aspects of a Multi-Purpose Thesaurus

Petöfi, J. S.: **Some aspects of a multi-purpose thesaurus.**
In: Intern. Classificat. 1 (1974) No. 2, p. 69–76

Linguistic and documentation theoretical researches aim at developing a lexicon that contains a maximum amount of verbal and an optimum amount of encyclopaedic information. The study discusses the following questions concerning a lexicon that constitutes a component of a text theory: (a) the structure of the lexicon in general, (b) the structure of the elementary units of the lexicon, (c) the structure of the definitions in the lexicon, (d) the relations among the units of the lexicon, and (e) the employment of the lexicon in text processing. Since in the formal and content structure of this lexicon the various requirements of its applications have also been taken into consideration, this lexicon is capable of fulfilling all functions that a thesaurus must fulfil and thus it can be considered as a multi-purpose thesaurus. (Author)

0. Introduction

The direction taken by the study of the thesaurus-structure in the theory of documentation and by the study of the lexicon-structure in linguistics in recent years makes it desirable that a closer cooperation between the researchers in these two fields (or at least a more effective exchange of information) be established. Such a cooperation would surely prove to be advantageous for both fields. The present paper, in which some aspects of the thesaurus/lexicon structure will be analysed more closely, is meant to be a contribution to this cooperation.

In the analysis I start from the following considerations:

a) The different aspects of text processing (the different forms of text interpretation, and the different forms of information storage and retrieval) impose different but not mutually exclusive demands on a theory that is to underlie textprocessing; thus, it is theoretically possible to conceive of one single theory fulfilling all demands of textprocessing.

b) Linguistics (and those related branches of knowledge, necessarily involved in linguistic research as e. g. the different logics) are capable of reaching a stage where a theory meeting the requirements indicated in a) can actually be developed.

c) In textprocessing as well as in a theory fulfilling the demands of textprocessing the basically important component is what can be called thesaurus or lexicon.

Being a linguist and regarding the 'multipurpose thesaurus' referred to in the title of this paper as one of the components of a *text theory* also aiming at fulfilling the claims of different applications I shall turn first to some problems of application and then to some aspects of the construction of a text theory.

1. Textprocessing and a theory of text

1.1 By the term '*textprocessing*' I wish to indicate all operations which can be performed on a text. The two main classes of textprocessing are text interpretation and information storage and retrieval.

The term '*text interpretation*' will be used as a term referring to a complex of operations. The interpretation of a text means performing the following operations: the *grammatical description* (assignment of the possible intensional-semantic representations of the text), and the *extensional-semantic description* (assignment of the possible extensional-semantic representations to all intensional-semantic representations of the text). There is, in addition, a third (subsidiary) operation: the *commenting description* (explanation and/or evaluation of the single extensional-semantic representations from some viewpoint). Since here theoretical operations are concerned, a basic requirement is that the descriptions be explicit and as comprehensive as possible. It is easy to see even without any detailed analysis that both the depth of the intensional-semantic description and the appropriate performability of the extensionalization depend to a great extent on the lexicon/thesaurus component of the theory by means of which the interpretation is carried out.

By the term '*information storage and retrieval*' I refer to a complex of operations to which, among others, the following operations belong: indexing, abstracting, extracting, condensing, establishment of data banks, elaboration of question-answer-systems. If we wish to automate these (in the first line intensional-semantic) operations, all operations which are involved in the automatic text analysis and synthesis must be ranked with here, too. It is obvious that effective completion of these operations depends to a great extent on the structure of the thesaurus/lexicon.

1.2 It is not absolutely necessary for a theory describing the structure of the units of a given language to make allowance for the extralinguistic applicability of the theory; however, the above enumerated textprocessing operations obviously require an applicable theory. Since it is theoretically not impossible to develop a theory which meets both the linguistics-internal and the application requirements at the same time, it is expedient to aim at developing such a theory. The present progress in linguistics, the formal syntactic, semantic, pragmatic and text-theoretical research, offers favourable conditions for it.

If we use the term '*theory of text*' to indicate a theory which aims at analysing and describing all aspects of texts, a theory which is capable of performing those

https://doi.org/10.5771/0943-7444-1974-2-69 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

intensional and extensional-semantic operations which have been referred to in connection with the interpretation of texts and the information storage and retrieval can be called a *partial theory of text*.

The so-called 'text structure world-structure theory' (abbreviated: 'TeSWeST' after the original German term) is conceived to be such a partial theory of text.

The TeSWeST is an empirically motivated logic-oriented theory. To characterize its components briefly, it is expedient to start from the semiotic triangle (form-intension-extension). Since it is obvious that the theory never operates with the intensions and extensions themselves, but with their representations, the semiotic triangle must be assigned a triangle containing the representations (F-- R_i -- R_e) By way of this assignment we obtain a double-triangle (cf. Figure 1).

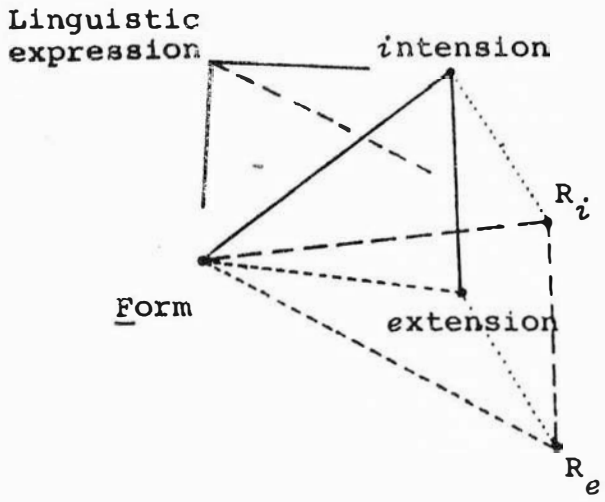

Figure 1

If we aim at the explicit representation of the intensions (and this aim is not only natural but also necessary), the R_i-s mean in all cases an approximation fixed in the given theory with respect to the intensions. This approximation is sufficient or insufficient, depending on the depth of the given R_i-s and/or the relevance of the elements of which the given R_i-s are built up. (Of relevance can, of course, only be spoken with reference to an aim).

The TeSWeST consists of two main components: The first one, the so-called *text grammar* component is concerned with the relation F--R_i, while the second one, the so-called *extensional-semantic* (or world-semantic) component is concerned with the relation R_i--R_e.

(One can, of course, also consider a relation F--R_e, which plays an especially important role in learning a language and in the various psycholinguistic investigations; this relation, however, will not be dealt with here.)

The interpretation of the R_e and the R_i of a text within the TeSWeST can briefly be given as follows:

An R_e is always the descirption of a world-complex, namely of that particular world-complex which can be assigned to a given intensional representation as an extensional interpretation (as a model). The description of this world-complex has to contain the following elements:

a) the list of the *objects existing* in the different subworlds of the world-complex,
b) the list of the *facts/events/* ... which are *true/valid* ... in the different subworlds of the world-complex,
c) the diagram of the *temporal* and/or *logical connections* of these facts/events/ ...,

and finally

d) the diagram of the *accessibility relations* among the different subworlds.

(These lists can, of course, only comprise those objects and facts/events/ ..., which have also been represented in the intensional representation.)

I began this short characterization deliberately with the R_e, because, as a matter of fact, this is the most important unit from the point of view of almost all applications of a text theory. Literary theoreticians, jurists, theologians, and documentation theoreticians are (more precisely: they even must be), interested, first of all in what kind of extensional interpretation can be assigned to the text analysed by them, and an R_e is the representation of such an interpretation. (From the point of view of the applications this is, of course, only a minimal interpretation that can or must be commented in several respects.)

On the basis of the above brief characterization of the structure of an R_e one can easily imagine what an R_i should contain. Since the task of the R_i-s is to permit the construction of R_e-s, an R_i is the description of the ordered complex of facts/events ... manifested in a given text. This description must contain the following elements:

a) the description of the objects manifested in the text,
b) the descriptions of the facts/events/ ... manifested in the text,

and finally

c) the description of the temporal and/or logical relations of these facts/events/ ....

The descriptions a) and b) are implicit or explicit. The implicit or explicit character depends on how the lexicon of the TeSWeST is used when constructing the R_i.

Since the "R_i" is the mediator between the "form" (written or spoken verbal expressions) and the "R_e" (the denotata that can be assigned to the written or spoken verbal signs), and the set-up of the R_i-s depends, in the last analysis, on the representation of the intension of the words, it goes without saying that the lexicon, thesaurus component plays the main role also within the TeSWeST.

So far the terms 'lexicon' and 'thesaurus' have been used as equivalent terms. The reason why I used them in that fashion is that in the theory of documentation the component in question is preferably indicated by the term 'thesaurus', whereas the corresponding term in linguistic is 'lexicon'; the denotatum of both these terms can, however, already be considered as being practically the same. In the following, for simplicity's sake, I will always speak of 'lexicon'.

2. On the structure of the lexicon

In connection with the structure of the lexicon the following question should be analysed briefly:

https://doi.org/10.5771/0943-7444-1974-2-69 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

1. the structure of the lexicon in general
2. the structure of the elementary units of the lexicon
3. the structure of the definitions in the lexicon
4. the relations among the units of the lexicon.

2.1 The structure of the lexicon in general

The lexicon of a natural language contains the representations of the intensions that can be assigned to the words, more precisely: to the single readings of the words, of the language, that is, it does not contain words but theoretical constructs.

The two main sectors of the lexicon are the sector of the definitions and that of the relations.

The sector of the definitions is based on the hypothesis that within the set of the representations of the intensions it is possible to separate a proper subset which can be declared to be the set of the *elementary semantic representations* (ESeR). This means, in other words, that a proper subset can be found the elements of which function as undefined (atomic) units in the lexicon (in the grammar) of the given language; by means of these atomic units each unit belonging to the complementary set, the so-called *lexical representations* (LeR), can be defined.

A definition in the lexicon is an LeR (definiendum) and an SeR (definiens) connected by the definition sign (=D, where the symbol "D" is always on that side of the equality sign where the definiens stands).

The symbol "SeR" is the abbreviation of "semantic representation". An SeR is either an ESeR, or a set of ESeR-s, or such a set consisting of SeR-s and ESeR-s which can be traced back in a finite number of steps to a set consisting solely of ESeR-s.

The *sector of relations* contains all relations among the units of the sector of definitions which cannot be derived from the definitions.

Remarks:

The two sectors of the lexicon that were dealt with above contain the so-called primary lexicon units. Without entering into details of this question it should be mentioned that either a distinct sector of the lexicon or a subcomponent of the transformation component must contain the list of the so-called 'secondary lexicon units'.

The term 'primary lexicon unit' refers to those lexicon units which can occur in an intensional-semantic representation. The term 'secondary lexicon units', on the other hand, refers to those units (morphemes and lexical units) which are introduced by means of transformation rules in the case of the synthesis, and are eliminated by inverse transformations in the case of the analysis.

The lexicon-entry of these units must be able to specify the conditions of their introduction and elimination, respectively, as well as the instructions concerning their introduction and elimination, respectively.

2.2 The structure of the elementary units of the lexicon

The elementary units of the lexicon are the elementary semantic representations (ESeR) and the lexical representations (LeR).

Both the ESeR-s and the LeR-s are interpreted and represented as predicate functions. These predicate functions have a well-defined internal structure. (In forming this structure the following investigations play a major part: manysorted logical investigations, investigations concerning the various classification systems, and valence- and case-grammatical investigations.)

To display the internal structure of the predicate fuctions let us examine an example:

The most simple predicate function that can be assigned to one of the readings of the verb *tell* (x tells y to z) is:

(1) *tells* (x, y, z)

However, the choice of this simple notation entails that, on the one hand, restrictions concerning the intensional-semantic well-formedness will not be represented and, on the other hand, the functions that "x", "y" and "z" fulfill in this predicate function cannot be represented in an explicit way.

A more complex and explicit way of notation can be achieved by making use of a many-sorted logical representation. Such a representation is e. g.

(2) *tells* (x^{s_i}, x^{s_j}, x^{s_k})

where, say, x^{s_i} and x^{s_k} indicate living creatures satisfying certain conditions, and x^{s_j} indicates an object satisfying certain conditions. (This means that x^{s_i} is, a matter of fact, an abbreviation: it indicates an x to which the predicates $f_{s_i1}(x)$, $f_{s_i2}(x)$, ... $f_{s_in}(x)$ can be related; f_{s_i1}, f_{s_i2}, ... f_{s_in} characterize together the set from which x must be chosen in order that the predicate gained from the predicate function *tells* (x, y, z) be well-formed. -- x^{s_j} and x^{s_k} can be interpreted similarly. (A predicate can be gained from a predicate function by providing the variables occurring in it with values.)

This representation, however, does not indicate the function fulfilled by x^{s_i}, x^{s_j}, and x^{s_k}. (The knowledge of the function performed by an element in some predicate function amounts to the knowledge of whether the element in question is an 'agent', an 'experiencer', a 'patient', etc. in the 'action' indicated by the predicate function.) However, even this function can be expressed with the aid of an even more complex many-sorted logical notation.

We can operate for example with the following representation:

(3) *tells* (x^{S_L}, x^{S_M}, x^{S_N}).

Since the number of the different functions is presumably finite, we can designate them with $a_1, a_2, \ldots a_n$. In this case the above notation can be interpreted by saying that "x^{S_L}" indicates an element with the sort-specification "s_i" performing the function "a_m" (e. g. a living creature satisfying certain conditions, having the function 'agent'); the other two symbols of the representation are to be interpreted similarly.

However, valence- and case-grammatical investigations seem to prove that the functions and the object-classes are in some respect independent from one another, and presumably any object from any object-class can perform any function. From this it follows that the specification

https://doi.org/10.5771/0943-7444-1974-2-69 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

of the function and the specification of the class of elements can be treated as characteristics that are dependent on but separable from one another. The above notation can be modified as follows:

(4) *tells* $(a_m : x^{s}i, a_r : x^{s}j, a_z : x^{s}k)$

With this notation the arguments-parts "$a_m : x^s$", etc. are to be handled as single argument-symbols, and the formula (4) is to be interpreted as follows: $x^{s}i$; tells $x^{s}j$ to $x^{s}k$. The fact that in (4) it is $x^{s}i$ that tells something is known because we know that $x^{s}i$ fulfills the function a_m (this function is usually called "agent-function"); the fact that it is $x^{s}j$ that is told is known because we know that $x^{s}j$ performs the function a_r (this function can be called "patient" or "object" function) and, finally, the fact that it is $x^{s}k$ to whom $x^{s}j$ is told through $x^{s}i$ is known because we know that x^s performs the function a_z (this function can be called "experiencer").

It appears from this interpretation, too, that the functions are not identical with the grammatical subject, object, etc.

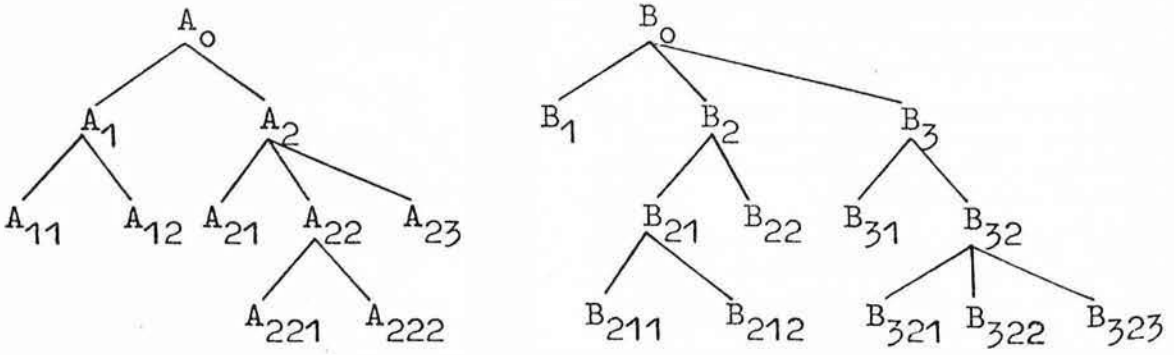

Figure 2

Remarks:

I cannot enter here into a detailed discussion of the questions concerning the structure of the elementary units of the lexicon, so I must content myself with some remarks only:

a) In the above analysis I have dealt only with one type of the elementary units, I have demonstrated the structure of the representation of a verb. The first step of setting up a lexicon is to define the types of the elementary units. The precondition of this is to have such a grammar at our disposal which is capable of defining a canonical form for the representation of a so-called extended simple sentence. Only on the basis of this canonical form is it possible to define the possible functor-argument-relations as the theoretical basis for the distinction of the types.

b) With respect to the single types it is necessary to specify the possible (obligatory and optional) argument-functions and those atomic or complex units which can fulfill these functions. (Also the specification of the possible complex units presupposes the grammar mentioned in a) as a basic condition.)

c) The markers of the argument-functions can be considered as elementary or non-elementary semantic representations. If they are considered as non-elementary semantic representations it is necessary to give their definitions, and in this case the symbols "a_i" are to be considered as abbreviations which stand for the respective function-definitions.

d) The sort-specification can generally be carried out bot in the case of the atomic and in that of the complex units by means of multi-hierarchical classification-systems. The sort-specification can be marked by a complex symbol or by a non-specified variable plus the hierarchy-specific predicates related to it. Let us assume for example that we have two classification hierarchies at our disposal (cf. Figure 2). In this case a certain predicate function can be noted either in the way.

(5) $f(a_i : x^{22,321})$

(declaring the convention that the first number-superscript refers to the system "A", and the second number-superscript to the system "B"); or in the way

(6) $f(a_i : x) \,\&\, A_{22}(x) \,\&\, B_{321}(x)$

where both "a_i" and the other symbols are abbreviations (on the basis of a defined redundancy-convention): "a_i" refers to the definitions of an argument-function (if the markers of the functions are not to be considered as elementary semantic representations), and the specificatio related to "x" in (5) and (6) indicate that with respect to "x" the predicates $A_0(x)$, $A_2(x)$, $A_{22}(x)$, $B_0(x)$, B_3 $B_{32}(x)$ and $B_{321}(x)$ are valid.

2.3 The structure of the definitions in the lexicon

When treating the structure of the definitions in the lexicon, the formal and the content structure of these definitions must be treated separately.

Let us consider first of all a definition (Cf. Figure 3; "l and "p" are symbols which stand for predicates belonging to different categories; the symbol "a" stands for function 'agent', "e" for 'experiencer', "h" for 'haben "o" for 'object', "s" for the 'unspecified source', "lg" 'local goal' and "T" for the 'global temporal function' the common language description in the column on th right hand side is meant to demonstrate how a definit is to be read.)

I want to emphasize that this definition is only to exemplify the structure of definitions; it does not lay claim any respect to completeness and definitiveness.

Concerning the *formal* structure of the definitions th following can be said:

https://doi.org/10.5771/0943-7444-1974-2-69 https://www.inlibra.com/de/page/agb - Open Access -

p^o book	$o:_1x2$	$=_D$		*iff* something$_1$ /=$_1$ x2/ is a book, *then*
		OBJECT22	$o:_1x2$	something$_1$ is an OBJECT22
	$_1p$	collection-of	$o:_1x2$ $s:_2x2$	something$_1$ is a collection of a certain number /=qu/ of something$_2$ /=$_2$ x2/
	$_2p$	PURPOSE-OF	$h:_1x2$ $o:_1P$ $_3P$	something$_1$ has the purpose that
	$_3p$	read	$o:_1x2$ $a:_1x1$	somebody$_1$ reads something$_1$
		OBJECT1	$o:_1x1$	somebody$_1$ is an OBJECT1
	$_4p$	have	$e:_1x1$ $o:_4x2$	somebody$_1$ has something$_4$
		OBJECT1	$o:_2x1$	somebody$_2$ is an OBJECT1
	$_5p$	adult	$o:_2x1$	somebody$_2$ is adult
	$_1p^o$	sheet	$o:_2x2$	something$_2$ is a sheet
	$_6p$	print	$lg:_2x2$ $a:_2x1$ $o:_3x2$	somebody$_2$ prints something$_3$ on something$_2$
	$_2p^o$	text	$o:_3x2$	something$_3$ is a text
	$_3p^o$	ability-of-reading	$o:_4x2$	something$_4$ is the ability-of-reading
	$_6p^T$	*takes-place*	$o:_6p$ T:PAST	$_6p$ takes place in the past

Figure 3

A definiens is, as regards its form, a special TextSeR. Its special character lies in the fact that it does not contain a number of elements in a explicit form, which ought to be contained by a real TextSeR. The reason for this is that with respect to the definientes general redundancy-conventions can be formulated which are valid within the whole lexicon. (One of the redundancy-conventions is that the "and"-symbol connecting the predicates is not explicitly indicated.)

A definiens consists globally of the following parts:

a) the main sector of the definiens (in Figure 3 the first definiens-part reaching to the first broken line) contains the 'sort-predicate(s)' (in Figure 3 this is "OBJECT22 $o:_1x2$") and the 'primary defining predicates' (in Figure 3 they are $_1p$, $_2p$ and $_3p$); "OBJECT22" refers to a classification system (cf. Figure 4, where also the interpretation of "x1" and "x2", the so-called 'global sort-variables', can be found);

b) the sector of the information relating to the arguments (in Figure 3 the second part of the definiens, reaching to the second broken line; the predicates are arranged here according to the order of the first argument-variables (cf. the subscripts on their left hand side));

c) the sector of the information which cannot be derived automatically from the redundancy-conventions (in Figure 3 it is $_6p^T$).

On the basis of the formal structure of the definientes two classes of definitions can be distinguished: the class of the *open* definitions and that of the *closed* definitions. A definition is formally open if one of the predicates of the definiens contains in one of the argument-places a so-called 'con-textual parameter'. Such open definitions are e. g. the definitions of most of the adjectives (and, consequently, all definitions in which these adjectives occur). For example, if something is said to be "big", this means that it is qualified as big on the basis of either an individual measure or a socially agreed convention. The knowledge of this measure or convention is a condition of the extensionalisation. If a definition is formally not open then it is formally closed.

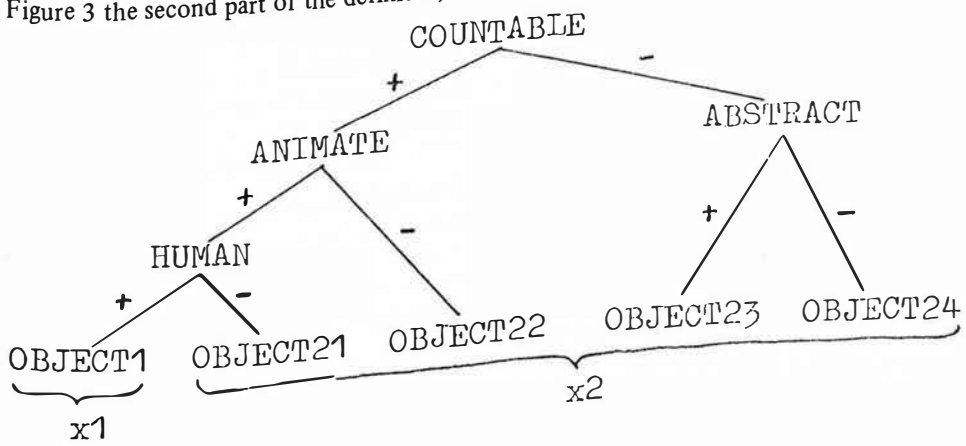

Figure 4

https://doi.org/10.5771/0943-7444-1974-2-69 https://www.inlibra.com/de/page/agb - Open Access -

The most important aspects of the *content* structure of the definitions are the aspect of the technical language specificity of the definitions and the aspect of the depth of the definitions.

The *technical language specificity* of the definitions is for several reasons a complex question. To solve the content-questions in connection with the lexicon it is necessary to clarify first of all the relation of the various technical languages to one another and the relation of all technical languages to ordinary language. The clarification of these relations is closely correlated with the questions of the classification of the sciences and humanities and with the questions of the linguistic text typology. Only after these questions have been clarified is it possible to discuss the technical language specificity of the definitions at all. (I only wanted to indicate these problems here, a detailed analysis should follow on another occasion.)

The *depth* of the definientes can be defined on the basis of various criteria. Two of the possible criteria should be mentioned here:

a) The definientes should be so deep (should contain so many defining predicates) as to guarantee that a definiens defines only one definiendum within the given lexicon. This criterion specifies a necessary and sufficient formal condition.

b) The definientes should be so deep as to permit the derivation of all semantic implications from a complex structure set up of definientes (from the intensional semantic representation of a sentence or a text) which meet a given expectation. Such expectations can be determined in the case of ordinary language on the basis of the verbal and factual knowledge of the average speaker/listener, and in the case of a technical language on the basis of the respective sciences or humanities. This criterion specifies a necessary and sufficient content condition.

It is obvious that while a) can be fulfilled in a generally acceptable way, several problems arise in connection b).

It is, however, also obvious that the fulfillment of b) is much more important with respect to the different applications than the fulfillment of a).

Remark:

From among the various questions concerning the definitions the following should still be mentioned: the question of the 'flexible depth' which concerns both the formal and the content aspect of the definientes. By a secondary structuring of the definientes one can make it possible for definiens-segments of different depths to be assigned to a definiendum. This is a lexicon-specificity which is useful for all kinds of text processing.

2.4 Relations among lexicon units

It is obvious that the definientes permit the immediate derivation of several relations among elementary lexicon units: thus the sort-predicates (especially in the case of multi-hierarchical sort-specification) permit the derivation of various hyperonym – hyponym (broader term – narrower term) relations, the defining predicates permit the derivation of the definitory relations (inclusive of the different field-relations). However, in the course of text processing it is necessary to quickly recognize/apply also relations which either cannot be derived at all or not immediately from one of the definitions.

The relations which cannot be derived from definitions at all include those between the elementary semantic units (ESeR). Two types of them should be mentioned here: the explicative and the hierarchical relations.

a) The *explicative* (or to use another term: convertibility) relations are, as a matter of fact, pseudo-definitory-relations. Since the ESeR-s are constructs the functors of which are marked by "words", which usually have several meanings, it must be ensured that the ESeR-s are desambiguated in the lexicon. This can be achieved by assigning to the ESeR-s, an explication (pseudo-definition) likewise consisting of ESeR-s. These explications are pseudo-definitions, because the circularity cannot be eliminated from them. The reason why the explicative relations are also called convertibility relations is that these pseudo-definitions permit the ESeR-s to be substituted for by their explications (by the pseudo-definientes), i. e. the definientes can be converted, whereby the paraphrasing capacity of the lexicon (the grammar) increases.

b) The *hierarchical* relations are the so-called elementary classificatory relations, upon which the sort-specification applied in the lexicon is based. They can generally be represented in the form of different tree-structures.

Among those relations which cannot immediately be derived from definitions one can rank the *synonymy*, *antonymy* and *converse* relations. They must be given in the sector of relations in the form of lists.

Remark:

A special question concerning the relations is the question of the so-called broadly interpreted antonymy relations, i. e. the question of the relation of a definiens and its negated form. Since the structure of a definiens is a complex structure, it is necessary to specify unambiguously which element (elements) of the definiens is (are) to be negated if the definiendum is negated. These are relational pieces of information which must be built into the definientes themselves.

3. Some remarks concerning structural relations among lexica

Though in this study the main emphasis has been on the analysis of the internal structure of the lexicon, some aspects of the structural relations among lexica also must be briefly touched upon. Concerning these relations a distinction must be made between the relations of the sublanguage-lexica of a natural language and the relations of the lexica of different natural languages.

3.1 The main question concerning the sublanguage-lexica is how the single sublanguage-lexica must be constructed in order that a lexicon-complex resulting from two or more sublanguage-lexica fulfills the same requirements as the single lexica (axiomatic and circularity-free construction, etc.). This question can generally be formulated as follows: is it possible to construct the single sublanguage-lexica in such a way that they can be considered as

https://doi.org/10.5771/0943-7444-1974-2-69 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

parts of one single general lexicon? A satisfying solution of this question presupposes of necessity the clarification of such already mentioned questions as the relation of technical languages to one another, the relation of technical languages and ordinary language, the depth of the definiens that can be altered according to the given parameters, etc.

The basic questions concerning a *general lexicon* consist in deciding in which sublexicon which elements are to be considered as ESeR-s, in which other sublexicon (or sublexica) the single sublexicon-specific ESeR-s are defined, and which are the real ESeR-s of the general lexicon. (It is not necessary that all sublexicon-specific ESeR-s be ESeR-s in the general lexicon if the latter is considered as one single lexicon.)

3.2 The main question concerning the structural relations among lexica of natural languages is to decide whether it is possible to postulate a *universal ESeR-set*. Giving a definitive answer to this question (if a definitive answer can be given at all) and building a universal ESeR-set surely cannot be done at once. One must initially be content with so-called restricted universals, i. e. if one succeeds in finding universals with respect to a language-family or some class of languages.

The ESeR-s are within a language 'per definitionem ESeR-s'. Their postulation as ESeR-s is justified by the fact that the elements of the complementary set can be defined by their help in the way required by the given theoretical framework. ESeR-sets can surely be chosen in many different ways, and the choices can have different motivations. The bringing about of a restricted universal ESeR-set depends obviously on whether one succeeds in finding a universal motivation which directs the choice within the single languages of a language-family or of some class of languages.

A short remark concerning bilingual lexica: It is obvious that an axiomatically constructed bilingual lexicon would be able to mirror the relations among the lexical units of the two languages much more explicitly than the bilingual lexica available so far. – From the viewpoint of text processing it is very important to investigate the question whether it is possible to elaborate, within a class of lexica to which the ESeR-set is common such an algorithm as would make it possible to arrive from a given definiens of a given lexicon at a definiendum that can be assigned to this definiens by means of another given lexicon.

4. The lexicon in text processing

After having discussed some aspects of the structure of the lexicon, let us now consider the role in text processing of an axiomatically built up lexicon containing explicit intensional-semantic representations.

The lexicon outlined in point 2 is applied in text processing as a component of the TeSWeST. (The lexicon is at the same time – and this is a very important fact with regard to the applications – not only a component of the TeSWeST but also one of its linguistic-internal applications in that the structure of the definientes is an intensional-semantic text representation meeting the formal requirements of the TeSWeST.)

From among the application-aspects of the discussed lexicon-structure I deem it necessary to point out the following:

a) The main sector of the definientes (cf. 2.3 (a)) is capable of fulfilling all functions which a documentation thesaurus must fulfill. (This sector on the one hand, presupposes the existence of different kinds of classificatory systems and, on the other hand, it furthers their development.)

b) Since the structure of the definientes is compatible with the structure of the intensional-semantic representation of a text, this lexicon (together with the TeSWeST) is a suitable means for bringing about data banks. (Cf. the short description of the sector of the arguments (2.3 (b)) and the structure of the R_i-s (1.2).)

c) From b) it follows that this lexicon (together with the TeSWeST) is a suitable means for elaborating question-answer-systems.

d) Owing to the explicit and formal structure of the definientes and the flexible depth, this lexicon is a suitable means for automatic text analysis and synthesis.

e) From b), c) and d) it follows that this lexicon (together with the TeSWeST) is suitable for both the theoretical and the empirical investigation of all questions relating to the "machine intelligence"-research.

f) From the aspects treated in point 3 the following can be concluded:

(f1) an axiomatically constructed lexicon is a suitable means for analysing the relation between technical language and ordinary language in an explicit way, and thereby it contributes to the investigation of the structure and the classification of the sciences and the humanities;

(f2) an axiomatically constructed bilingual lexicon can be a suitable means for machine translations.

Considering these points, I think that it is justified to speak of a convergence of lexicon and thesaurus research and that it is also justified to regard a text theoretical lexicon as a multi-purpose thesaurus.

5. Concluding remarks

In the Introduction I pointed out the advantages and the necessity of closer cooperation between linguists and documentation theoreticians, a point I want to emphasize again in conclusion.

In recent years linguistic literature has increased immensely (due to a great extent to the intensified interrelationships of linguistic and logical research), so that keeping track of it or even merely overseeing it is becoming increasingly difficult. The same seems to be true of the literature dealing with the questions of text processing, too. Thus, the solution of a complex task can only be expected from teamwork.

The questions related to the lexicon are the most urgent ones in both these fields of research, and their solution will depend on efficient teamwork between linguists and documentation theoreticians.

https://doi.org/10.5771/0943-7444-1974-2-69 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Bibliographical notes:

The lexicon discussed in the present study and the theoretical framework (the 'text-structure world-structure theory' /TeSWeST/), of which it is a component are analysed from different points of view in the following studies:

van Dijk, T. A., Ihwe, J., Petöfi, J. S., Rieser, H.: *Zur Bestimmung narrativer Strukturen auf der Grundlage von Textgrammatiken*. (= Papiere zur Textlinguistik, Band 1 und 1A). Hamburg: Buske 1972. (2. Auflage mit einem Nachwort von H. Rieser, 1974).

Petöfi, J. S.: *Towards an empirically motivated grammatical theory of verbal texts*. In: Petöfi, J. S., H. Rieser (Eds.): Studies in Text Grammar. Dordrecht: Reidel 1973.

Petöfi, J. S.: *Zum Aufbau eines „Lexikons"*. In: Rave, D., Brinckmann, H., Grimmer, K. (Eds.): Syntax und Semantik juristischer Texte. Darmstadt 1972.

Petöfi, J. S.: *Modalität und topic-comment in einer logisch fundierten Textgrammatik*. In: Dahl, O. (Ed.): Topic and comment, contextual boundness and focus. (= Papiere zur Textlinguistik, Band 6). Hamburg: Buske 1974.

Petöfi, J. S., Rieser, H.: *Präsuppositionen und Folgerungen in der Textgrammatik*. In: Petöfi, J. S., Franck, D. (Eds.): Präsuppositionen in Philosophie und Linguistik/ Presuppositions in Philosophy and Linguistics. Frankfurt: Athenäum 1973.

Petöfi, J. S., Rieser, H.: *Probleme der modelltheoretischen Interpretation von Texten*. (= Papiere zur Textlinguistik, Band 7) Hamburg: Buske 1974.

It should be noted that information gained from documentation theory and thesaurus research has influenced the construction of the TeSWeST. Cf.:

Petöfi, J. S.: *A tezaurusz-kérdés jelenlegi helyzete*. [The present state of the thesaurus problem]. Budapest: OMKDK 1969.

Petöfi, J. S.: *On the problems of co-textual analysis of texts* COLING (International Conference on Computational Linguistics in Stockholm), Preprint No. 50. 1969. In German translation in: Ihwe, J. (Ed.): Literaturwissenschaft und Linguistik I–III. Frankfurt: Athenäum 1972.

Regarding the present state of the empirical work aiming at the construction of a multi-purpose thesaurus/lexicon cf. the description of the research-project "Aufbau eines axiomatischen Kern-Lexikons der deutschen Sprache" in this issue (p. 99).

Robert Fugmann
Hoechst AG, Frankfurt/M.-Höchst

The Glamour and the Misery of the Thesaurus Approach

Treatise IV on Information Retrieval Theory[1]

Fugmann, R.: **The Glamour and the Misery of the Thesaurus Approach.**
In: Intern. Classificat. 1 (1974) No. 2, p. 76–86

If any important natural-language term which a documentalist encounters in storing literature and in phrasing enquiries is admitted as an addition to a thesaurus, then the thesaurus will soon exceed the limits of its operancy and will increasingly fail to serve the purpose of an efficient device for reliable terminological control in the input and retrieval stage. This continous decline can effectively be counteracted by conceptual analysis of candidate terms and by resynthesis of the terms of their conceptual constituents. This suggests a balanced combination of the thesaurus and the analytico-synthetic classification approach, particularly in large information retrieval systems. The representation of certain, predominantly syntactical relations, however, exceeds the capabilities of both approaches. These relations can be managed by two different devices described, namely by a clearly defined set of relation indicators and by an optionally additional graphical representation of extended concept relations. (Author)

1. Introduction

Any mechanized literature search aims at retrieving documents from a file that are relevant to the special topic of the inquirer. In order to enable the search mechanism to perform this task the inquirer will have to *define* the special goal of his literature search. In such a search request it must be laid down *in advance*, i. e. without any previous knowledge about relevant documents contained in the file, which particular features should be possessed by the desired documents and are to be considered as an indication of their relevance to the special topic of the inquirer (cf. 1, postulate of definability, p. 134). This is at least true of a test search directed to a sample of the entire file, on the basis of which the request can be modified and then directed to the entire file. In particular, it must be laid down in advance in the

1 Extended version of a paper presented at the Third International Conference on Classification Research, Bombay, January 1975 First treatise: Ref. 1; Second treatise: Ref. 3; Third treatise: Ref. 15

https://doi.org/10.5771/0943-7444-1974-2-69 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY