[From: A.Zampolli, N.Calzolari (eds.) *Computational and mathematical linguistics*. Proceedings of the International Conference on Computational Linguistics, Pisa, 1973, vol.II (Firenze: Olschki, 1980)]

ANNELY ROTHKEGEL

IDIOMS IN AUTOMATIC LANGUAGE PROCESSING

The automatic identification of multi-word idioms in texts is part of the project "Automatic Lemmatization",[1] the main problem of which is to recognize lexical units of the stem dictionary and their grammatical characterization in a text which – as a rule – contains inflected forms, ambiguities, and idioms as multi-word expressions.

In this paper some examples shall be given to demonstrate the treatment of the latter ones by means of special dictionary entries and a special algorithm.

The problem is that groups of several words are to be identified as semantic and lexical units. The linguistic point of such a classification is based on sign theory. In this connection the individual word cannot generally be regarded as a linguistic unit.

In addition to the difficulties of identification there are problems of the dictionary. Most of the idioms are ambiguous; they have an idiomatic meaning as well as a literal one. These are expressions like *White House* ("government of the USA"), *green light* ("licence"), *draw red herrings* ("devert"), *kick the bucket* ("die") or in German *rechte Hand* ("assistant"), *blinder passagier* ("deadhead"), *grüne Finger haben* ("be a skillful gardener"), *in der Tat* ("indeed"), etc.[2]

It is an important point that there are semantic relations between the idiomatic and the literal meaning.[3] Therefore the use of idioms in sentences is often restricted, e.g. *kick the bucket* must be connected with a subject marked as "person" or at least as "animate", which is not the same for *die*. Relations of this kind – as far as they are known – should be made explicit in the lexicon.

---

[1] This is one of the projects at the Sonderforschungsbereich "Elektronische Sprachforschung", University of Saarbrücken, Germany.

[2] It should be mentioned that the problem of idioms is a general one and not dependent on a specific language. In this paper most of the examples are given in English. This doesn't mean that the system shown in the following is restricted to English idioms.

[3] Cp. R. W. LANGACKER (1968, p. 80) "standardized metaphors".

One way to do this is to integrate idioms into the normal dictionary by referring them to the one-word units which are likewise elements of the idioms. Thus it is possible to compare the semantic features of both and to mark the internal metaphorical relations still being under investigation.

The description of the automatic analysis can be divided into three parts:[4]

1) encoding of special morphologic, syntactic, and semantic markers, which are important for the identification;

2) integration of the multi-word units together with the grammatical information into the stem lexicon;

3) algorithm for the discovery of such units within the sentence analysis.

1. These expressions are divided into classes corresponding to their fixed syntactic patterns. Such a pattern can be described as part of a tree embedded in a phrase structure tree. The dominating node represents the part of speech of the whole expression. The categories verb (*V*), noun (*N*), adverb (*A*) are substituted by *V'*, *N'*, and *A'* each of them dominating an idiomatic subsystem. These systems are generated by small phrase structure grammars: for each pattern several productions are fixed in form and order to represent the whole derivation of the idiom.

Further we have to consider the fact, that the same structural components and their combinations can be found in various syntactic patterns which are dominated by different nodes, e.g. coordination connected with preposition:

beer and skittles	dominated	by *N'*
by leaps and bounds	»	by *A'*
rain cats and dogs	»	by *V'*

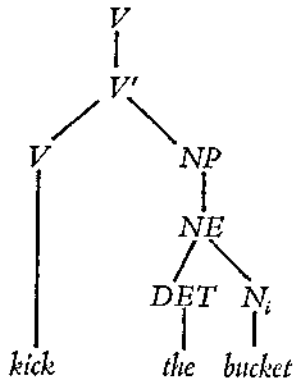
With regard to this it is necessary to use the auxiliary symbols needed to produce different patterns the structure of which is partly the same.

---

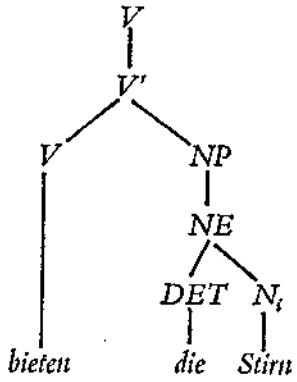
[4] An exhaustive description of German idioms is given in A. ROTHKEGEL (1973).

Examples:

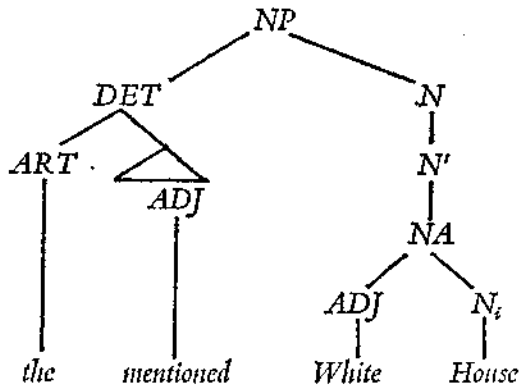
a) *kick the bucket*



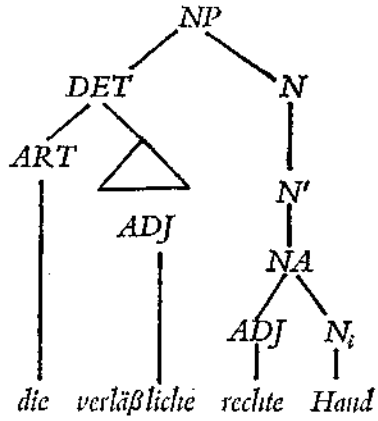
a') *die Stirn bieten*



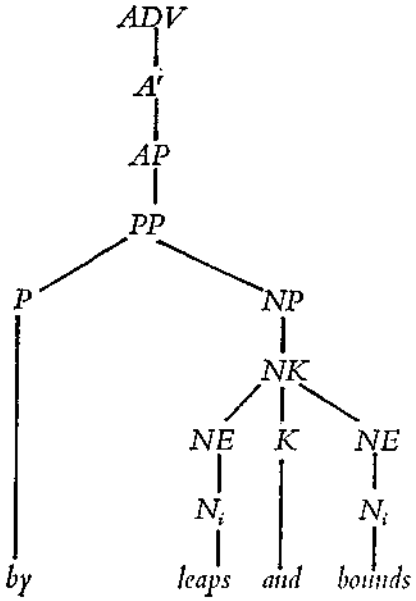
b) White House in *the mentioned White House*



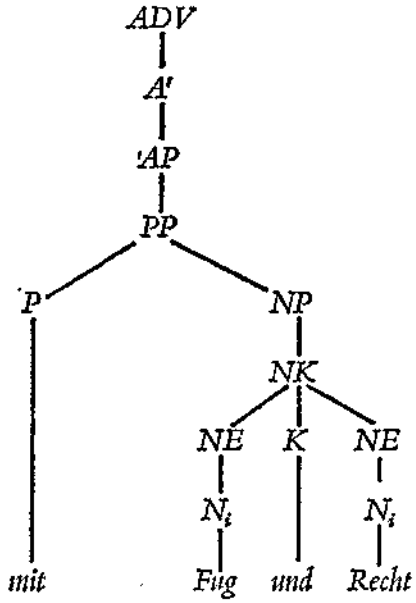
b') rechte Hand in *die verläßliche rechte Hand*



c) by leaps and bounds



c') mit Fug und Recht



The class an idiom belongs to is symbolized by an eight-figure number. It is formed systematically, so that the productions necessary for the syntactic structure can be derived from it. The number of the first place (on the left) determines the part of speech of the whole expression, the number of the last place refers to the part of speech of the nominal basic lexeme like noun$_i$ (noun in idiom), adjective, adverb, verb. The other numbers characterize the syntactical surface structure of the idiom (cp. the lists of the basic rules). By such a procedure new classes can be added without changing anything in the system.

List of basic rules:

Number of the rule	rule	marked in class number as
(1)	$Z \rightarrow N_i$	1 (8. place)
(2)	$Z \rightarrow ADJ$	2 »
(3)	$Z \rightarrow ADV$	3 »
(100)	$N' \rightarrow NP$	4 (1. place)
(200)	$A' \rightarrow AP$	2 »
(210)	$AP \rightarrow NP$	if 1 → 9 at 2. place
(220)	$AP \rightarrow PP$	» *A* or *B* »
(230)	$AP \rightarrow PP\ PP$	» *C* »
(240)	$AP \rightarrow PP\ P$	» *D* »
(400)	$V' \rightarrow VP$	4 (1. place)
(410)	$VP \rightarrow V\ NP$	if 1 → 9 at 2. place
(420)	$VP \rightarrow V\ PP$	» *A* or *B* »
(430)	$VP \rightarrow V\ PP\ PP$	» *C* »
(10)	$NP \rightarrow NE$	1 (2. place)
(20)	$NP \rightarrow NA$	2 »
(30)	$NP \rightarrow NK$	3 »
(40)	$NP \rightarrow NE\ NE$	4 »
(50)	$NP \rightarrow NE\ NA$	5 »
(60)	$NP \rightarrow NE\ PP$	6 »
(60), (99)	$NP \rightarrow NE\ PP,\ P \rightarrow P_{en}$	7 »
(70)	$NP \rightarrow NA\ PP$	8 »
(70), (99)	$NP \rightarrow NA\ PP,\ P \rightarrow P_{en}$	9 »
(80)	$PP \rightarrow P\ \ NP$	*A* »
(80), (99)	$PP \rightarrow P\ \ NP,\ P \rightarrow P_{en}$	*B* »
(90)	$PP \rightarrow P\ \ NP\ P$	*D* »
(95)	$PP \rightarrow PP\ PP$	*C* »
(11)	$NE \rightarrow Z$	0 (from 3. place)
(12)	$NE \rightarrow DET\ Z$	1 »
(13)	$NE \rightarrow (DET)\ Z$	2 »

(21)	$NA \rightarrow ADJ\ Z$	0	»
(22)	$NA \rightarrow DET\ ADJ\ Z$	1	»
(23)	$NA \rightarrow (DET)\ ADJ\ Z$	2	»
(24)	$NA \rightarrow (ADJ)\ Z$	3	»
(25)	$NA \rightarrow DET\ (ADJ)\ Z$	4	»
(26)	$NA \rightarrow Z\ ADJ$	5	»
(31)	$NK \rightarrow NE\ K\ NE$	0	»
(32)	$NK \rightarrow NA\ K\ NA$	1	»
(33)	$NK \rightarrow NE\ K\ NA$	2	»
(34)	$NK \rightarrow NA\ K\ NE$	3	»
(41)	$PP\ PP \rightarrow P\ NP\ P\ NP$	0	»
(42)	$PP\ PP \rightarrow P_{en}NP\ P_{en}NP$	1	»
(43)	$PP\ PP \rightarrow P_{en}NP\ P\ NP$	2	»
(44)	$PP\ PP \rightarrow P\ NP\ P_{en}NP$	3	»
(99)	$P \rightarrow P_{en}$		

Example: *by leaps and bounds* or *mit Fug und Recht*

class number: 2A3000-1
1. place: 2 rule (200) : $A' \rightarrow AP$
2. » : A dependent on the 1. place additional rule (220) : $AP \rightarrow PP$ for A (80) : $PP \rightarrow P\ NP$
3. » : 3 rule (30) : $NP \rightarrow NK$
4. » : 0 rule (31) : $NK \rightarrow NE_1\ K\ NE_2$
5. » : 0 rule (11) : $NE_1 \rightarrow Z$
6. » : 0 rule (11) : $NE_2 \rightarrow Z$
8. » : 1 rule (1) : $Z \rightarrow N_i$

A generation of the productions is important for the analysis, because 1) the inventary and its order must be compared with the text and 2) the internal syntactic structure of the idiom can be represented as part of the sentence structure.

In addition to the productions there are morphological and syntactical restrictions. In a lot of cases they help to distinguish the idiomatic meaning from the literal one:

1) no discontinuity is permitted: *White House*, not *the white big house* or *the house is white; schwarzer Peter*, not *der schwarze kleine Peter* or *Peter ist schwarz*;

2) attribution is allowed: *the (mentioned) White House, eine (grosse) rolle Spielen*;

3) attribution is not allowed: *kick the bucket, die Stirn bieten*;

4) the noun must be in singular: *White House, rule the roost, kalte Küche, Fuss fassen*;

5) the noun must be in plural: *get cold feet, burn one's fingers kalte Füsse bekommen, Wurzeln schlagen*;

6) noun with article: *have the face, die Stirn bieten*;

7) noun without article: *from pillar to post, play second fiddle, Fuss fassen, Hand anlegen*;

8) only in negation: *cut no ice, nicht von Pappe sein*;

9) no passive: *kick the bucket, den Kopf verlieren.*

Generally the encoding refers to the whole expression. The list of all marked idioms is the basis to get the lexical entries automatically by attaching the grammatical information to the corresponding basic lexeme of the idioms described in the following chapter.

2. The simplest way to connect the expression and the grammatical information would be to build up a separate lexicon of idioms as is practised in general. In this case the dictionary only has the function of a simple word-store and data carrier for the algorithm. From a linguistic view this is not satisfactory for reasons I mentioned above. Beyond that there is the fact that idioms belong to the field of word-formation. The basic lexemes of the idiom are already elements of the vocabulary represented in the dictionary. Therefore it is adequate to use the one-word entries of the normal dictionary as addresses of idiom entries.

With regard to the lexicon we distinguish two types of idioms:

I) the meaning of only one lexeme is determined by the partner lexeme as in *blinder Passagier, four-letter word.* Only *blind* respectively *four-letter* has an idiomatic meaning.

II) The meaning of the whole expression strictly depends on the fixed connection of several lexemes. It is not possible to divide the meaning corresponding to the single constituents, e.g. *White House, kick the bucket, bring home the bacon, Staub aufwirbeln, den Mond anbellen.*

Concerning I there are no problems in the dictionary: the single lexemes (only basic lexemes) are marked as parts of an idiom by I_t representing the class number, and by *R*, representing the information about the grammatical restrictions. The lexeme with the idiomatic meaning is characterized additionally by *S*, the component of the special semantic features. The corresponding lexical entries of the examples in I are as follows:

blind	:	I_t, *R*, *S*	*four-letter*	:	I_t, *R*, *S*
Passagier	:	I_t, *R*	*word*	:	I_t, *R*

With regard to II there is no semantic congruence between the constituents of the idiom. Instead, the partner lexeme itself has the function of a marker. The component I_{II} represents the class number of this kind of connection, *R* and *S* referring to the whole expression are added only to one of the constituents. The following entries concern the idioms *White House, rechte Hand, kick the bucket, in die Hand nehmen.*

White	:	I_{II}	*House*	:	I_{II}, *R*, *S*, *white*
recht	:	I_{II}	*Hand*	:	I_{II}, *R*, *S*, *recht*
bucket	:	I_{II}	*kick*	:	I_{II}, *R*, *S*, *bucket*
Hand	:	I_{II}	*nehmen*	:	I_{II}, *R*, *S*, *Hand*

The component *S* also contains the information as to whether the idiom as a whole is ambiguous, i.e. whether there could be a literal meaning in the sentence. This point is important for the result of idiom identification.

3. The identification of idioms is relevant in that part of the sentence analysis in which the parts of speech are determined. The ambiguity of the idioms can only be solved on a higher level of the syntactic analysis. But here it is possible to decide whether there is an idiom in a sentence or not.

During the dictionary look-up the idiom marker *I* gets importance. As only basic lexemes are marked and not function words (e.g. only *Tat* of *in der Tat*, only *pink* of *in the pink*) there can be an idiom in a sentence even if only one *I*-marked word exists. On the other hand there can be several *I*-marked words which do not necessarily belong together and form an idiom.

The program of identification is divided into the following steps:

a) check whether there are the same or different class numbers of the *I*-marked words (elements of one idiom have the same class number.). Form groups of *I*-marked words which have the same class number.

b) Compare whether the quantity of basic lexemes of every group corresponds to the requirement of the class.

c) Generate the sequence of the parts of speech dependent on the class number and compare with the sequence of *I*-marked words and their context in the text.

d) Check the grammatical restrictions of the potential idiom and compare with the text.

e) Compare the partner lexemes of the text with the dates of the lexicon.

If one of the conditions from *b* to *e* is not fulfilled, the analysis stops with a negative result.

The following examples show the complexity of the analysis, when there are parts of idioms which potentially belong to different expressions. First a list of such idioms is given. It also contains a paraphrase of their idiomatic meaning, the class number, the quantity of required basic lexemes corresponding to this class, and the sequence of the parts of speech, which can be generated by means of the class number.

Examples:

idiom	class number	quantity of basic lexemes	sequence of parts of speech
green light (traffic light, licence)	120----1	2	*ADJ*, N_i
red light (traffic light, danger)	120----1	2	»
red tape (bureaucratism)	120----1	2	»
a red rag to a bull (object which enrages s.o.)	18*A*211-1	3	*DET*, *ADJ*, N_i, *P*, *DET*, N_i
in the light (of) (in the view of)	2*A*11---1	1	*P*, *DET*, N_i
on the rocks (in financial difficulties)	2*A*11---1	1	»
in the red (in debt)	2*A*11---1	1	»
see red (become enraged)	410----2	2	*V*, *ADJ*
see the light (be born)	411----1	2	*V*, *DET*, N_i
see the red light (see danger ahead)	421----1	3	*V*, *DET*, *ADJ*, N_i
draw red herrings (devert)	420----1	3	*V*, *ADJ*, N_i

The corresponding lexicon entries of the basic lexemes are as follows. Here only the idiom information is listed which is to be added to the usual grammatical markers.

bull	*I*(18A211-1)
draw	*I*(420 ----1), *R,S,* *red, herring*
green	*I*(120 ----1)
herring	*I*(420 ----1)
light	1.*I*(120 ----1),*R*,S_1, *green* 2.*I*(120 ----1),*R*,S_2, *green* 3.*I*(120 ----1),*R*,S_1, *red* 4.*I*(120 ----1),*R*,S_2, *red* 5.*I*(2A11 ---1) *R, S,* 6.*I*(411 ----1) 7.*I*(421 ----1)
rag	*I*(18A211-1), *R,S,* *red, bull*
red	1.*I*(120 ----1) 2.*I*(18A211-1) 3.*I*(410 ----2) 4.*I*(421 ----1) 5.*I*(420 ----1) 6.*I*(2A11---1)
rock	*I*(2A11---1)
see	1.*I*(410 ----2),*R,S,* *red* 2.*I*(411 ----1),*R,S,* *light* 3.*I*(421 ----1),*R,S,* *red, light*
tape	*I*(120 ----1),*R,S,* *red*

Analysis of idioms in sentences:

(1) *He cannot understand that the company is on the rocks.*

I-marked lexeme	class number
rock	2*A*11 - - - 1

The program then follows the several steps of the identification scheme given above:

a) There is only one group of *I*-marked words.

b) Only one basic lexeme is required. There is agreement with the text.

c) As sequence of parts of speech *P DET N* is generated, which corresponds to the context of *rocks* in the sentence.

d) Grammatical restrictions: no discontinuity, definite article, noun in plural.

These conditions are fulfilled.

e) As the class 2*A*11---1 requires only one basic lexeme, a partner lexeme can not be compared.

Result: *on the rocks* can be an idiom. It is ambiguous.[5]

(2) *But I saw the red light some time before.*

I-marked lexemes	class number	
see	410----2	
	411----1	
	421----1	
red	120----1	421----1
	18*A*211 -1	420----1
	410----2	2*A*11---1
light	120----1	411----1
	2*A*11---1	421----1

Scheme of analysis:

a) a group must have elements of the same class number or one element with a single class number. The following groups are formed:

		required lexemes
1) *see*	410----2	2
red	410----2	
2) *see*	411----1	2
light	411----1	
3) *see*	421----1	3

[5] This information is given by the *S*-component of the lexicon entry.

	red	421 - - - - 1	
	light	421 - - - - 1	
4)	*red*	120 - - - - 1	2
	light	120 - - - - 1	
5)	*red*	18*A*211 - 1	3
6)	*red*	420 - - - - 1	3
7)	*red*	2*A*11 - - - 1	1
	light	2*A*11 - - - 1	

b) After comparing the required lexemes with the given ones groups 5 and 6 are omitted. Group 7 must be substituted by two different groups, because the class 2*A*11 - - - 1 requires only one basic lexeme.

8)	*red*	2*A*11 - - - 1	1
9)	*light*	2*A*11 - - - 1	1

The idiom identification program continues with a group containing the most elements. In this case it is group 3.

c) Sequence of part of speech corresponding to class 421 - - - - 1: *V DET ADJ N*. It is the same sequence as in the text. The text also corresponds to the

d) grammatical restrictions as "noun in singular", "definite article", "no further attribution".

e) The partner lexemes of *see* are *red* and *light*. There also is an agreement with the text.

Result: *see the red light* is a potential idiom. It is ambiguous.

In this case it is not necessary to check the other groups.

(3) *She saw a light on the green rocks.*

I-marked words	class number	
see	410 - - - - 2	
	411 - - - - 1	
	421 - - - - 1	
light	120 - - - - 1	411 - - - - 1
	2*A*11 - - - 1	421 - - - - 1
green	120 - - - - 1	
rock	2*A*11 - - - 1	

Scheme of analysis:

a)	groups	class number	required lexemes
	1) *see*	410 - - - - 1	2
	2) *see*	411 - - - - 1	2
	light	411 - - - - 1	
	3) *see*	421 - - - - 1	3
	light	421 - - - - 1	
	4) *light*	120 - - - - 1	2
	green	120 - - - - 1	
	5) *light*	2*A*11 - - - 1	1
	rock	2*A*11 - - - 1	

b) Group 1 and 3 are omitted because there is no agreement of the quantity of basic lexemes. Group 5 is to be substituted by two groups:

6) *light*	2*A*11 - - - 1	1
7) *rock*	2*A*11 - - - 1	1

c) The program continues with one of the biggest groups. The first is group 2. Sequence of parts of speech corresponding to class 411 - - - - 1: *V DET N.* This is the same as in the text.

d) The grammatical restrictions are violated: the idiom must contain the definite article, but the sentence contains the indefinite article.

Result: group 2 does not represent any idiom. The next is group 4; the program restarts with

c) Sequence of parts of speech corresponding to class 120 - - - - 1: *ADJ N.* The immediate context of *light* is not an adjective. Therefore the result is negative: group 4 also does not represent any idiom.

The next is group 6; restart at

c) Sequence of parts of speech for 2*A*11 - - - 1; *P DET N.* There is no agreement with the sequence in the text which is *V DET N* (*saw a light*).

The last one is group 7:

c) Sequence of parts of speech: *P DET N.* The context of *rocks*

differs from the required sequence of the parts of speech. It is *DET ADJ N* (*the green rocks*).

Result: this sentence does not contain any idiom.

The next two examples show that a syntagm can be an idiom or a part of another syntagm.

(4) *After having bought a car he is i n t h e r e d*
(5) *She lives i n t h e r e d house.*

In this case the part of speech of the basic lexeme given by the class number is relevant. If it is an adjective, the following context must be checked.

Finally, besides the practical purpose which is evident in computational language analysis automatic idiom identification is characterized by the general question, how to integrate linguistic aspects into computational linguistics. I think, there is an opportunity to get automatically new bases for language research. With regard to the lexicon there is another point: a computer lexicon does not need to be readable by a human user. Therefore it is possible to make explicit principles of productivity like word-formation and metaphor which would be a more adequate description of the vocabulary.

Symbols:

A'	subsystem of adverb
ADJ	adjective (part of speech)
ADV	adverb (part of speech)
AP	adverbal phrase
DET	determiner (*ART* article)
N'	subsystem of noun
N_i	noun in idiom (part of speech)
NA	nominal phrase attributed by adjective
NE	simple nominal phrase
NK	nominal phrase with coordination
NP	nominal phrase (not specified)
P	preposition (part of speech)
P_{am}	preposition with article (e.g. *zum, im*)
PP	prepositional phrase
V'	subsystem of verb
V	verb (part of speech)
VP	verbal phrase
Z	dummy symbol for noun, adjective, adverb

REFERENCES

E. AGRICOLA (ed.), *Wörter und Wendungen. Wörterbuch zum deutschen Sprachgebrauch*, Leipzig 1963².

J. BAR-HILLEL, *Idioms*, in W. N. LOCKE, A. D. BOOTH (eds.), *Machine translation of languages*, New York 1957² (first ed. 1955), pp. 183-193.

R. PH. BOTHA, *The function of the lexicon in transformational generative grammar*, The Hague 1968.

W. L. CHAFE, *Idiomaticity as an anomaly in the Chomsky paradigm*, in « Foundations of Language », IV (1968) pp. 109-127.

W. FLEISCHER, *Wortbildung der deutschen Gegenwartssprache*, Leipzig-Tübingen 1971² (first ed. 1969).

B. FRASER, *Idioms within a transformational grammar*, in « Foundations of language », VI (1970), pp. 22-42.

W. FRIEDERICH, *Moderne deutsche Idiomatik*, München 1966.

J. J. KATZ, P. M. POSTAL, *Semantic interpretation of idioms and sentences containing them*, Quarterly Progress Report 70, Research Laboratory of Electronics, M.I.T., Cambridge (Mass.) 1963, pp. 275-282.

R. KLAPPENBACH, *Feste Verbindungen in der deutschen Gegenwartssprache*, PBB(H) 82, 1961, pp. 443-457.

R. W. LANGACKER, *Language and its structure: Some fundamental linguistic concepts*, New York 1968.

I. A. MEL'ČUK, *O terminax 'ustojcivòst' i 'idiomaticnost'*, in « Voprosy Jazykoznanija », IV (1960), pp. 73-80.

A. ROTHKEGEL, *Feste Syntagmen. Grundlagen, Strukturbeschreibung und automatische Analyse*, Tübingen 1973.

U. WEINREICH, *Problems in the analysis of idioms*, in J. PUHVEL, *Substance and structure of language* (ed.), Berkeley-Los Angeles 1969, pp. 23-81.

H. WISSEMANN, *Das Wortgruppenlexem und seine lexikographische Erfassung*, in Indogermanische Forschungen », LXVI (1961), pp. 225-258.

F. T. WOOD, *English Colloquial Idioms*, London 1969.