Ricarda DORMEYER, Ingrid FISCHER, Martina KEIL, IMMDII, University of Erlangen

A Database for Verbal Idioms

Abstract

Phrasal idioms are of major interest for traditional as well as for computational linguists. *Phraseo-Lex* is a database for German verbal idioms that was designed with both groups of researchers in mind. A database system is used to classify idioms from a syntactic, semantic, and pragmatic point of view. Interfaces to this database have been specified to insert, delete, change, and search the information it contains. On top of these interfaces a graphical user interface is developed to make their use easy. But these interfaces can also be used to connect *Phraseo-Lex* directly to any other system the idiomatic knowledge can be useful for.

Keywords: verbal idioms, computational dictionary, database

1. Introduction

Several traditional dictionaries for German phrasal idioms exist, for example Duden. Most of the information they contain has been assembled for traditional linguists or learners of German as a foreign language. It is difficult to use them in a computational environment. On the other hand, there are many computer based lexicons that are hardly readable for a human and not applicable in other areas. With our phraseological database *Phraseo-Lex* we position a dictionary for German verbal idioms in this gap.

Important advantages of computational dictionaries compared with traditional ones are the higher amount of volume capacity and the diversity of access to the stored information. What is more, computer based dictionaries are no more limited to a linear form of representation, for example an alphabetical order, but allow the user to determine the degree of detailed information themselves. For making use of these advantages it is necessary to develop a general computational description of phraseological knowledge. Such a kind of phraseological description has a further advantage: a large part of the information about a specific idiom can be made available in a way that enables other computer programs, for example parsers, to use it.

These thoughts led us to the following architecture for *Phraseo-Lex*: The core is formed by a database containing all information about an idiom. To collect, extract, change, and search this information, interfaces have been defined in a formal way. As *Phraseo-Lex* is also intended to be used by non-computer specialists, we developed a sophisticated graphical user interface on top of these interfaces, making it easy to use the database. But it is not necessary to use the graphical interface, it is also possible to attach other programs directly to the interfaces to allow the programs to extract the information they need. To enable humans as well as computers to profit from the information contained in the *Phraseo-Lex* database, we have chosen a semi-formal way to represent the data. This means that for parts of the information, for example an idiom's syntactic description, a formal representation is used, which is visualized in the graphical user interface. For other attributes, like examples of an idiom's use, no restrictions are given.

At the moment, we work on a second version of *Phraseo-Lex*. In the first version the underlying database was implemented in Scheme resp. Prolog, as this offered a good chance to connect it to a variety of other programs. The graphical user interface was written in C and Motif. This version lacked good possibilities to change and search the underlying data structures.

The new version is implemented in Java as this offers platform independence, a good connection between the graphical user interface and the underlying database, and the possibility of an interface to the World Wide Web. As database we have chosen a relational SQL-based one connected via the *Java Database Connectivity* Interface which is included in the normal Java Package.

In the following, the lexical entry in *Phraseo-Lex* for one idiom is described in detail and is illustrated by screen shots taken from the graphical user interface.

2. The *Phraseo-Lex* Lexical Entry

The phraseological information about an idiomatic entry, as it is implemented in *Phraseo-Lex*, is structured in a tree-like fashion. Starting with the idiom's lemma and base lexemes on top, it is then possible to specify the idiom's properties on the syntactic, semantic, and pragmatic levels of description. In the following, we introduce the complete set of attributes of which a *Phraseo-Lex* dictionary entry consists. For each of these attributes different functions are defined to retrieve, to change and to search for the information they contain. These functions form the set of interfaces the database can be accessed with.

2.1. Lemma and Base Lexemes

In *Phraseo-Lex*, like in conventional dictionaries, a lexical entry is headed by the lemma, which should be represented in a general or canonical form. Sternkopf (1992:222) criticizes the traditional citation form for verbal idioms, because for a large number of idioms it lacks information about the subject position (see example (1)). The lacking information may be of a semantic nature, for example if the subject position is [+ human] or [- human]. The only exception are fixed predicative phrases like (2), i.e. idioms with a fixed internal subject.

(1)	jdm. einen Bären aufbinden sb. a bear tie-on to tell a tall tale to sb.	(2)	der Kopf raucht jdm. the head smokes sb. sb.'s head is spinning

We call verbal idioms with a variable subject VPL1[1], and fixed predicative phrases VPL2. For VPL2 idioms, we have decided to keep the traditional citation form, whereas for VPL1 idioms, to complete it with a marker *jemand* (somebody), *etwas* (something) or *jemand/etwas* (somebody or something) for the subject position. Example (3) shows the idioms cited above in *Phraseo-Lex* notation.

(3)	(jmd.) jdm. einen Bären aufbinden VPL1 (Bär, aufbinden)	jdm. raucht der Kopf VPL2 (Kopf, rauchen)

To allow access to the idiomatic entries from a lexical level, the lemma is indexed with a list of base lexemes. Base lexemes are the words occurring in the idiom that belong to an open word class, i.e. the idiom's nouns, adjectives, adverbs, and full verbs. They can be used in at least three ways, firstly by a parser to retrieve all idiomatic entries for a given word, secondly by a program that builds lexical entries for a conventional dictionary and uses *Phraseo-Lex* as a source for idiomatic knowledge, and thirdly by the *Phraseo-Lex* user interface to sort idioms according to their base lexemes.

The lemma as well as the base lexemes act as keywords for an idiomatic entry, with the lemma representing the entry's unique key.

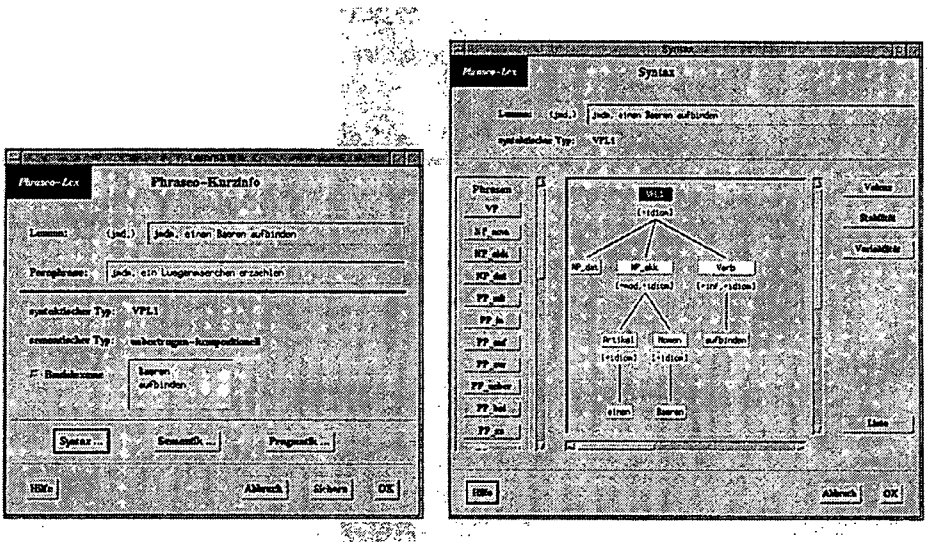

Figure 1: Central Attributes and the Main Window of the Syntactic Description

Together with three of the attributes to be explained below, namely syntactic type, semantic type, and main paraphrase, an idiom's lemma, and its base lexemes are regarded as central to our idiomatic description, and must be registered for every entry. The *Phraseo-Lex* graphical user interface provides a main window for each database entry in which these attributes are displayed (see Fig. 1, left side).

2.2. Syntactic Level

The syntactic level of description consists of the following parts:

We call the classification into VPL1 and VPL2 given above an idiom's syntactic type. It determines the idiom's basic syntactic structure: An idiom of type VPL1 is a complex verb phrase, an idiom of type VPL2 is a complete sentence.

The internal syntactic structure of a phrasal idiom is described by means of a phrase structure tree, which contains the idiom's constituent structure and is therefore regarded as the central element of the syntactic description.

For the construction of an idiomatic phrase structure tree, a phrase structure grammar is necessary. In *Phraseo-Lex*, it is implicitly given by an interactive tree building facility that is part of the graphical user interface. The right side of Fig. 1 shows the main window of the syntactic level, in which the tree is constructed. After the user has specified an idiom's syntactic type, the window displays the root of the tree: a VPL1 or a VPL2 node. The user can add nodes to the tree, using the syntactic category buttons on the window's left side. To keep *Phraseo-Lex* independent from all specific grammatical frameworks, we allow a wide range of syntactic categories as child nodes for any given node.

For further syntactic information, the user can define a set of syntactic features of their own choice and add them to each tree node. For example the phrase structure tree in Fig. 1 contains the features [+ idiom] for the nodes that are part of the idiom, [+ mod] for a constituent that may be modified, and [+ inf] to mark the verb as being in the infinitive form. These are the most common features in our current idiomatic database.

Just like simple verbs, verbal idioms require a certain number of complements to build a complete sentence. These are called the idiom's external valencies. Due to the phrasal structure, the internal structure can also be described in terms of valency theory, namely as the verb's complements or internal valencies (Wotjak 1992:54-56). The verbal idiom *(jmd.) jdm. einen Bären aufbinden* has one internal valency, *einen Bären*, and two external valencies, a nominative and a dative noun phrase.

If the convention mentioned above is followed, an idiom's internal and external valencies can be generated automatically from the syntactic structure. The syntactic type provides the information about the nominative argument (the subject); the other complements are derived from the phrase structure tree by finding out which parts are specified at word level, and which ones exist as phrase nodes only.

Optional complements are recorded by specifying them in the syntactic structure in the same way as obligatory complements. To mark them as optional, a user defined feature in the phrase structure tree might be defined.

In contrast to the traditional citation form for verbal idioms, our approach allows to distinguish between variable pronouns, i.e. external valencies, and pronouns that are fixed parts of an idiom. We mark this distinction in the phrase structure tree using the following convention: A variable pronoun is specified only as a phrase node, whereas a fixed part of the idiom is specified up to word level. For an example of a variable pronoun, see the NP_dat (noun phrase in the dative case) node in Fig. 1.

For each lexical entry, conventional variants can be listed. We distinguish variants from idiomatic synonyms and from modifications of the idiom by requiring that a variant must have the same connotations as the idiom's lemma. Therefore, a variant can be seen as an additional lemma for which most of the information in the lexical entry is also valid.

We distinguish between lexical variants like (4) and structural variants like (5). For a lexical variant, the entire entry except for the base lexemes is valid, whereas for a structural variant the phrase structure tree is slightly incorrect.

(4) jdm. kein Haar/Härchen krümmen
sb. no hair/hair-DIMINUTIVE bend
not to harm a hair of sb.'s head

(5) jdn. aufs/auf das Abstellgleis schieben
sb. on+the/on the railway siding shove
to push sb. aside

In the first implementation of *Phraseo-Lex*, variants are recorded as simple texts, but we are currently planning to develop a way to represent the changes to the base lexemes and the phrase structure tree explicitly.

An important characteristic of idioms is their resistance to syntactic manipulations: Some idioms cannot undergo certain syntactic transformations without losing their idiomatic reading; furthermore, so-called syntactic anomalies and unique components may occur in an idiom.

Syntactic transformations that can be marked as possible or impossible in *Phraseo-Lex* are for example passivization, relativization, negation, wh-question, and quantification. Transformations that apply to the idiom as a whole can be marked as possible, impossible or undecidable; for those that apply to a constituent of the idiom, typically to a noun phrase, a list of base lexemes that allow the transformation can be given.

A syntactic anomaly is a construction that would be considered syntactically incorrect outside the idiom it occurs in. Examples for this are missing determiners, additional pronouns that do not serve any function, and deviations of verb valency. In *Phraseo-Lex*, there is a list of possible syntactic anomalies, from which those that are found in the current idiom can be selected.

A unique component is a word that does not exist outside the idiom, i.e. a kind of lexical or morphological anomaly. An example for this is the word *Kerbholz* in the idiom *etwas auf dem Kerbholz haben* (to have done s.th. bad, to have quite a record). In *Phraseo-Lex*, each idiom's unique component(s) can be listed explicitly.

Idioms with a syntactic anomaly or a unique component do not have a non-idiomatic reading.

2.3. Semantic Level

The central notion for an idiom's semantic description in *Phraseo-Lex* is its compositionality. Most of the semantic level of description is grouped around this idea.

We believe with Wasow et al. (1983:102-115) that there exists a class of idioms, usually called *compositional* or *decomposable* idioms, for which parts of the idiom "have identifiable meanings which combine to produce the meaning of the whole" (Wasow et al. 1983:109).

We distinguish three semantic classes of idioms in order to characterize the different degrees of semantic compositionality: noncompositional idioms like *to kick the bucket* do not have an internal semantic structure, partially compositional idioms like *to rain cats and dogs* contain at least one element that keeps its literal meaning, though the rest of the idiom may be noncompositional, and in figurative compositional idioms like *to spill the beans* the meaning of the idiom can be distributed among its parts, where the resulting meanings "are not the literal meanings of the parts" (Wasow et al. 1983:109). We call this classification the idiom's semantic type.

Furthermore, we classify idioms with regard to their ambiguity, i.e. with regard to the question whether they have a nonidiomatic reading, or not, and with regard to their degree of motivation.

The meaning of an idiom is often explained in terms of another idiom. Korhonen (1988:202) notes that this kind of explanation can be found even in dictionaries. This way of explaining a meaning is useful neither for a non-native speaker who is unfamiliar with the idiom used in the explanation, nor for a language processing system in computational linguistics.

We therefore propose to describe the meaning of an idiom by means of one or more literal, non-idiomatic paraphrases, one of which is selected as the main paraphrase. The main paraphrase is required to reflect the idiom's semantic type, i.e. it must take the identifiable parts of meaning into account.

This poses no restriction on noncompositional idioms. For partially compositional idioms, it means that the idiom's literal element should become part of the main paraphrase as well. For figurative compositional idioms, the main paraphrase should have the same syntactic structure as the idiom, in such a way that the meanings of the idiom parts correspond to the meanings of appropriate paraphrase parts. A suitable paraphrase for the figurative compositional idiom *einen Bock schießen*, which translates literally as *to shoot a buck*, would be *einen Fehler machen* (to make a mistake), where *Bock* corresponds to *Fehler*, and *schießen* corresponds to *machen*, rather than simply *to err*. We take this idiom to be compositional because the *Bock* part can not only be modified, as shown in example (6), but even referred to in one of the following sentences in a discourse.

(6) Sie hat einen großen Bock geschossen.
She has a big buck shot.
She has made a big mistake.

The purpose of the semantic structure is to make the mapping between idiom parts and paraphrase parts explicit. Furthermore, each internal or external valency is assigned a semantic role (agent, patient, etc.), with those internal valencies that do not carry a meaning of their own being marked as having no role. We can proceed like this because meaningful parts of an idiom can be considered figurative arguments, whereas parts of a noncompositional idiom are not arguments, or quasi-arguments. The choice of an appropriate semantic role is based on the part's figurative meaning, not on its literal meaning.

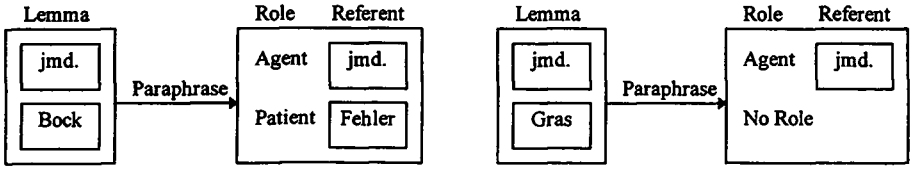

Figure 2: Semantic Structure of a Figurative Compositional and a Noncompositional Idiom

Fig. 2 shows the semantic structure for the figurative compositional idiom *einen Bock schießen* and for the noncompositional idiom *ins Gras beißen* (see example (7)).

(7) ins Gras beißen
in+the grass bite
to die

(8) nach dem Mond gehen
after the moon go
to be way out

For language processing purposes, the meaning of the idiom is coded as a logical formula. Possible logical formulas for the idioms in Fig. 2 are *make(X,Y)*, *mistake(Y)* and *die(X)*, respectively.

Fig. 3 shows the window of the graphical user interface containing this information for the idiom *jdm. einen Bären aufbinden.*

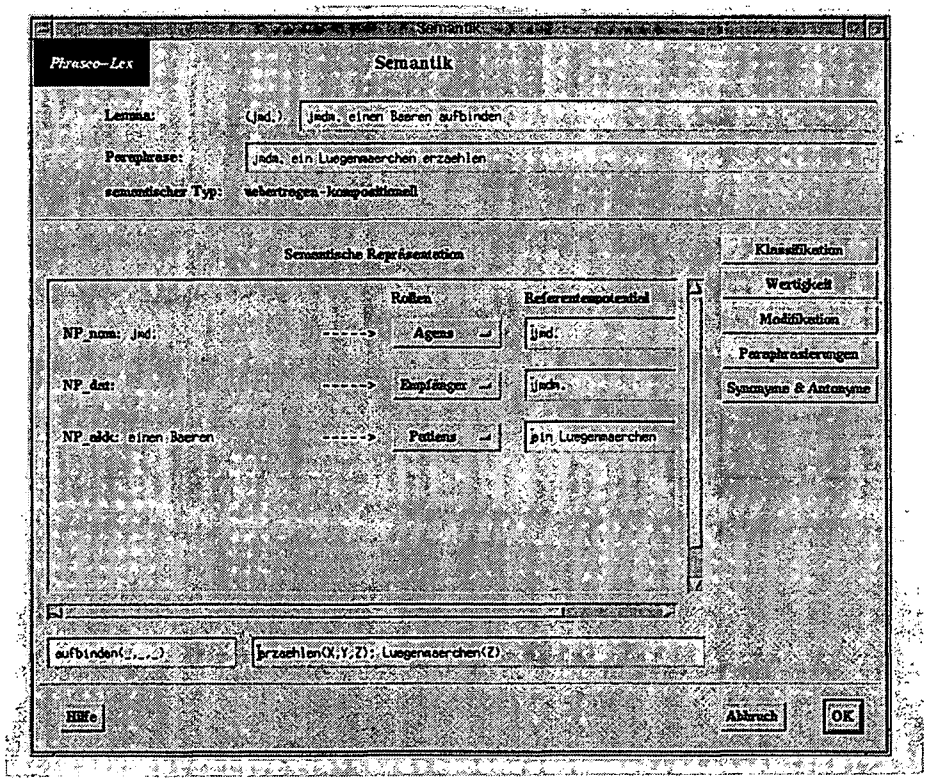

Figure 3: Window Showing Semantic Information

By using the terms *somebody* and *something* as place-markers for an idiom's external valencies, we place a semantic restriction on the lexical items filling these positions. Additionaly, we enable the user to specify further restrictions on both internal and external valencies by means of user defined semantic features, for example [+ human], [+ abstract], [+ institution].

Some idioms have valencies with extremely severe restrictions: In example (8), the idiom's subject position may be filled by clocks only. In *Phraseo-Lex*, this could be noted by using a

feature [+ clock]. A language processing system may make use of this information if it contains a detailed ontology component.

Modification of parts of an idiom, called internal modification, occurs frequently in idiom usage. We distinguish between word play, which is outside our field of interest, and systematic modification, i.e. linguistically unmarked additions that often take the form of adjective or genitive attributes. Wasow et al. (1983:108) give the example *leave no legal stone unturned*, meaning that all legal methods are used. A German example is given in (6).

Because of the wide range of possible modifications, it is not possible to keep a complete list for each part of each idiom. In addition to this, Dimitrij Dobrovol'skij[2] established that speakers differ greatly in their judgment on the grammaticality of a modification. Nevertheless, we believe such a collection of modification elements to be useful for analyzing the suspected connection between the compositionality of an idiom and the internal modifications it allows.

In addition to the literal paraphrases, we keep a list of idiomatic synonyms and antonyms for each idiom, i.e. a list of cross-references to idioms with the same or the opposite meaning. Because of our restricted definition of lexical variants, we classify idioms differing only in one lexical item as synonyms even if they have the same basic meaning, as long as they differ in their connotations. Therefore, the idioms in example (9) are regarded as synonyms, not as variants.

(9) Halt den Mund/die Schnauze.
Hold the mouth/the muzzle.
Hold your tongue/Shut your gob.

2.4. Pragmatic Level

An important characteristic of idioms, distinguishing them from their literal paraphrases, is their high degree of expressiveness. To take this into account, *Phraseo-Lex* completes the syntactic and semantic levels with a pragmatic level of idiom description.

When paraphrasing an idiom, it often becomes apparent that a non-idiomatic synonym lacks certain connotative aspects like for example a pejorative, emphatic or ironical connotation. Keil (1997:182) believes that this explains why people tend to explain an idiom in terms of another idiom. In *Phraseo-Lex*, emotional and general connotations can be stated explicitly for each idiom by means of a user defined list of values.

Many idioms are restricted in their usage with regard to region, style, or social groups using them. *Phraseo-Lex* maintains three separate categories storing diasystematic information, thus enabling the user to mark an idiom for example as specific to a regional dialect, as colloquial speech, or as belonging to a professional terminology.

Furthermore, some idioms are typically used in certain contexts, for example political discussion, card game, or sport. *Phraseo-Lex* allows to specify the typical situations of usage for each idiom.

Actual examples of idiom usage in texts can serve as an appropriate basis for further investigation about syntactic and semantic regularities concerning idioms. Therefore, *Phraseo-Lex* provides the possibility to collect a large number of realistic examples taken from written or spoken language for each idiomatic entry.

3. Searching the Database

Fig. 4 shows the graphical user interface for searching the *Phraseo-Lex* database. It is possible to search for all the parts of an idiom's lexical entry that have been introduced above. Additionally, the queries for an entry's different attributes can be connected with the logical operators *and*, *or*, and *not*. The system returns a list of all the idioms that fulfill the query. From this list their idiomatic entries can be viewed, changed, or deleted.

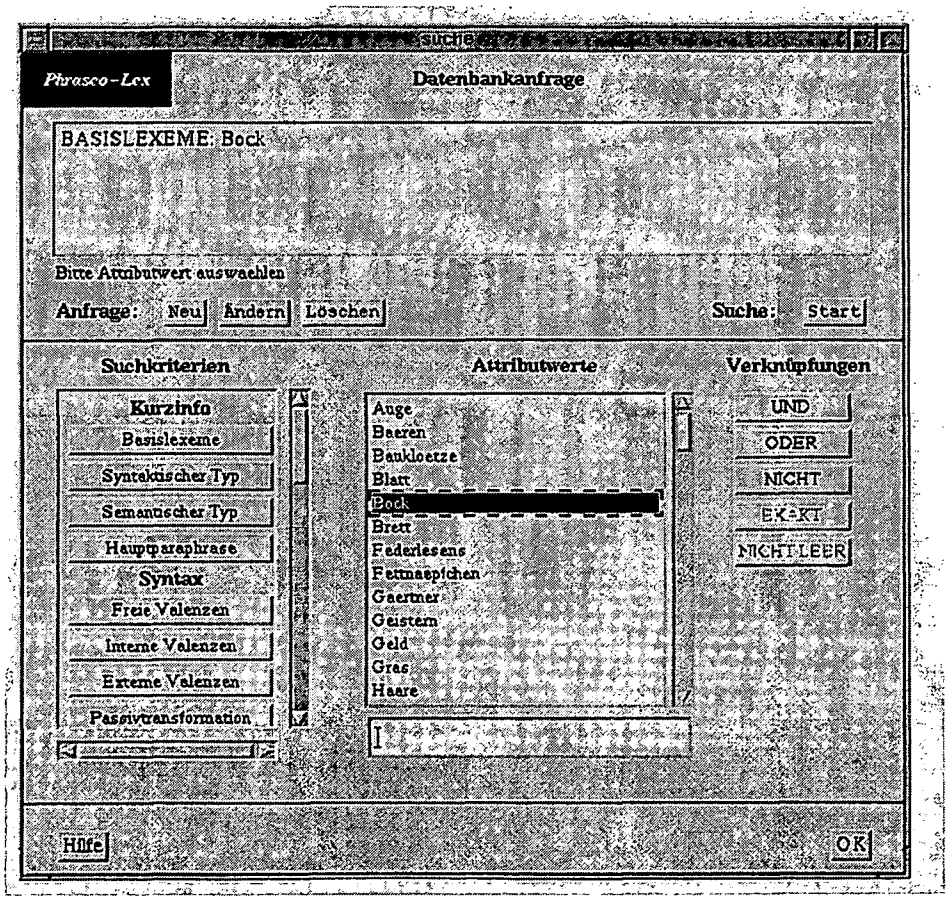

Figure 4: Window for Searching *Phraseo-Lex*

4. Conclusions

In this paper we presented a phraseological database for German verbal idioms. We developed a computer oriented description of idiomatic knowledge. In *Phraseo-Lex*, phrasal idioms can be classified from a syntactic, semantic, and pragmatic point of view. On the one hand *Phraseo-Lex* can be considered as a computational tool for collecting and representing idiomatic knowledge for lexicographical or phraseographical purposes. On the other hand, *Phraseo-Lex* serves as a computational lexicon in the field of computational linguistics.

The development of *Phraseo-Lex* is still in progress, and we are planning on dealing with the following problems: Firstly, the number of idioms in the database must be extended. With this extension the layout of the window system can be tested. After that, we will examine whether *Phraseo-Lex* can be used as a lexicon generating tool. If a specification of the lexicon structure is given, it should be possible to generate an arbitrary lexicon from *Phraseo-Lex* containing all necessary information about idioms in for example a HPSG-like structure. Finally, an interface to the World Wide Web will be implemented to make the information available and hopefully to collect new data.

5. Notes

[1] VPL is a short form for *Verbaler PhraseoLogismus*, meaning verbal idiom.

[2] Personal communication.

6. References

Keil, Martina (1997) *Wort für Wort - Repräsentation und Verarbeitung verbaler Phraseologismen*. Niemeyer, Tübingen.

Korhonen, Jarmo (1988) Zur (Un-) Verständlichkeit der lexikographischen Darstellung von Phraseologismen, in T. Magay and J. Zigány (eds.), *Euralex '88 Proceedings*. Budapest, pp. 197-206.

Sternkopf, Jochen (1992) Valenz in der Phraseologie? Ein Diskussionsbeitrag, in *Deutsch als Fremdsprache 4/92*. Leipzig, pp. 221-224.

Wasow, T., I. Sag, and G. Nunberg (1983) Idioms: An Interim Report, in S. Hattori and K. Inoue (eds.), *Proceedings of the XIIth International Congress of Linguists*. CIPL, Tokyo.

Wotjak, Barbara (1992) *Phraseolexeme in System und Text*. Niemeyer, Tübingen.