Published in: Steffens, Petra (Ed.): Machine Translation and the Lexicon.
Proceedings of the Third International EAMT Workshop, Heidelberg, Germany,
April 26-28, 1993. - Berlin [et al.]: Springer, 1995. Pp. 35-50.
(Lecture Notes in Computer Science 898)

Description and Acquisition of Multiword Lexemes

Angelika Storrer[1] and Ulrike Schwall[2]

[1] University of Tübingen (Germany)
[2] Sietec, München (Germany)

Abstract. This paper deals with multiword lexemes (MWLs), focussing on two types of verbal MWLs: verbal idioms and support verb constructions. We discuss the characteristic properties of MWLs, namely non-standard compositionality, restricted substitutability of components, and restricted morpho-syntactic flexibility, and we show how these properties may cause serious problems during the analysis, generation, and transfer steps of machine translation systems. In order to cope with these problems, MT lexicons need to provide detailed descriptions of MWL properties. We list the types of information which we consider the necessary minimum for a successful processing of MWLs, and report on some feasibility studies aimed at the automatic extraction of German verbal multiword lexemes from text corpora and machine-readable dictionaries.

1 Introduction

The treatment of multiword lexemes (MWLs) has always been a challenge for natural language processing (NLP) in general and for machine translation (MT) in particular. Most problems are caused by the fact that MWLs differ considerably from analogous free s yntagmatic constructions with respect to semantic compositionality, substitutability of components, and morpho-syntactic flexibility.

In this paper, we first give examples of the morpho-syntactic and semantic peculiarities of MWLs and we explain how these may pose problems for the analysis, generation, and transfer steps of machine translation systems. We then compile a list of those t ypes of information regarding MWL properties which have to be included in lexicons for NLP applications to enable a successful processing of the various MWL types. Finally, we report on some feasibility studies carried out as part of the ELWIS project at the University of Tübingen[3] with the aim to extract German verbal MWLs from text corpora and machine-readable dictionaries. The studies show that statistical methods which have proved to be successful in regard to English text corpora cannot simply be applied to German, and that a combination of corpus and dicti onary-based methods is a more favourable approach towards the automatic acquisition of MWLs provided that machine-readable dictionaries of high quality are available.

[3] ELWIS is a project on corpus-based development of Lexical Knowledge Bases carried out at the university of Tübingen (cf. [18]); the project is funded by the ministry of Science and Research of Baden-Württemberg.

2 Types and Properties of Multiword Lexemes

We use the following working definition for the term *multiword lexeme*:

> Multiword lexemes are units of a language's lexical system (lexemes) composed of several words.

This definition sets out two important properties of MWLs. The first property is lexeme status, which implies that MWLs, in contrast to free syntagmatic constructions, are stored and retrieved as complex units of the (mental) lexicon. The second property, to be composed of several words, distinguishes MWLs from simplex words.[4]

Given this definition, the term *multiword lexeme* covers quite a heterogeneous group of lexical units such as idiomatic expressions, lexicalized support verb constructions, lexicalized multiword compounds, phrasal verbs, and polylexical technical terms (cf. [5]). In the following section, we shall focus on verbal multiword lexemes, because this is the most interesting group characterized by a wide range of morpho-syntactic and semantic peculiarities. We use the term *verbal multiword lexeme (VMWL)* as a generic term, encompassing both verbal idioms and support verb constructions:

- Verbal idioms, such as (1) and (2),

  (1) **to kick the bucket**
  (2) **to spill the beans**

  have the structure of a verb phrase. However, the meaning of verbal idioms is not a compositional result of the idiom-external meaning of its constituents. Compared to free VP constructions, the idiomatic construction is subject to various morpho-syntactic constraints. Like other idioms, verbal idioms belong to informal or colloquial registers and express an affective evaluation of the things they denote.
  Prototypical idioms often involve metaphors or another form of figuration. As a consequence, sentences such as (3) and (4),

  (3) **John kicked the bucket.**
  (4) **John spilled the beans.**

  have, aside from their idiomatic reading, an alternative non-idiomatic reading, which may, to a more or lesser degree, be plausible in a given context.

- Support verb constructions (SVC), as (5) and (6),

  (5) **to take into consideration**
  (6) **to raise an objection**

  consist of a support verb (SV) and a predicative noun (Npred), which is, typically, a nominalization of an abstract verb or adjective.[5] The support

[4] From an NLP point of view we consider words as strings between blank spaces, although we are aware that this is a very simplistic conception from a linguistic point of view.

[5] We have adopted the term *predicative noun* from French research, cf. [4] and [7].

verb mainly contributes grammatical features such as person, tense, and mode, but influences the denotative meaning of the SVC only to a small degree. Compared to the basic simplex verb, in our examples *to take* and *to raise*, it has lost most of its semantic content. It is rather the function of the Npred to determine the denotative meaning and the argument structure of the construction as a whole. The arguments are normally inherited from the basis of the nominalization: the argument of the SVC in (6), for instance, realized as a prepositional phrase with the preposition *against*, is inherited from the argument of the basic verb *to object against*.
SVCs expand the range of expression of a verbal system: they can be used to make the process expressed by the Npred passive (*to receive praise*) or causative (*to set in motion*) and they can alter the aspectual dimension (*to get/to be/to keep in touch*).[6]

The borderline between verbal idioms and SVCs seems to be quite clear-cut for prototypical cases. In other cases, however, the distinction is difficult to draw and different possibilities of classification exist, depending on the criteria used.[7] What verbal idioms and SVCs have in common is that they differ from free syntagmatic constructions with regard to semantic compositionality, substitutability of components, and morpho-syntactic flexibility. In the following sections, we will describe these properties in more detail.

2.1 Non-standard Compositionality

The principle of semantic compositionality implies that the meaning of an expression is a function of the meaning of its parts and the syntactic rules by which they are combined.

Although it is still controversial whether idioms can be processed compositionally (cf. [19], [14]), there is general agreement that they cannot be analyzed in the same way as their non-idiomatic counterparts. Instead, different strategies have to be provided for three distinct cases of non-standard compositionality:

1. Verbal idioms like *to kick the bucket* have one single undecomposable idiomatic reading, assigned by convention to the expression as a whole. Approaches to treat these idioms compositionally lead to analyses which do not conform to linguistic intuition.

2. Verbal idioms like *to cast pearls before swine* can be broken up into their individual components, the meaning of which is motivated by conventionalized metaphors (*pearls* = something of value; *swine* = unworthy person). Based on the respective metaphorical links, idioms of this type may be compositionally analyzed, although the meaning of the idiomatic construction as a whole is still a matter of convention.[8]

[6] Cf. [16] for a contrastive account of the semantic-functional contribution of SVCs.
[7] An overview of possible criteria is given in [12] and [3].
[8] The relationship of metaphoric contents and compositionality is discussed in [19].

3. Some verbal idioms contain components which retain their normal MWI-external meaning: the noun *Streit* in the example (7)

   (7) einen Streit vom Zaun brechen
   Lit.: a quarrel from fence break
   Engl.: to suddenly start a quarrel

   is a lexically fixed component, i.e., it cannot be substituted by a semantically close lexeme. Nevertheless, the idiom-internal meaning of *Streit* is identical to its idiom-external meaning.

SVCs may be compositionally analyzed to a certain extent, because the SVC-internal meaning of the Npred is identical to its SVC-external meaning. The meaning of the SV, however, does not correspond to the meaning of the respective simplex verb, but is reduced to grammatical features such as aspect, passive, and causative.

2.2 Non-standard Substitutability

Synonyms, i.e., words denoting the same type of objects, may replace one another in a complex expression without changing the semantic value of the expression as a whole. In verbal idioms like *to kick the bucket*, however, the noun *bucket* cannot be replaced by the noun *pail*, even though the two nouns are synonymous in non-idiomatic contexts. Non-standard substitutability, like non-standard compositionality, is a consequence of the fact that many idiom chunks do not refer to objects in the usual way, so that meaning postulates, which are usually valid, cannot be applied.

Some verbal idioms, however, have a component which can be lexicalized by several different lexemes, e.g., *to take a bow/curtain*. The choice of one lexeme or the other will not affect the denotative meaning of the idiom, even though style and frequency of usage may be different. These cases have to be distinguished from cases such as *to be on good/bad terms with*, where the antonyms *good* and *bad* contribute in their regular way to the meaning of the idiom, thus leading to two antonymous idiom variants.

In comparison to idiom components, SVC components are substitutable to a certain extent. A particular Npred, however, cannot be combined with an arbitrary SV which has the required grammatical characteristics. For example, the verbs *bringen (zur Verzweiflung bringen*[9]*)* and *setzen (in Bewegung setzen*[10]*)* both have the features causative/inchoative. The Npred *Brand*, nevertheless, can be combined only with *setzen (in Brand setzen*[11]*)*, and the Npred *Anwendung* only with *bringen (zur Anwendung bringen*[12]*)*. It is obviously the Npred, which selects the appropriate support verb; a phenomenon, which causes severe problems for natural language generation.

[9] Engl.: to drive sb to despair.
[10] Engl.: to set in motion.
[11] Engl.: to start a fire.
[12] Engl.: to apply.

2.3 Non-standard Morpho-syntactic Properties

Grammars for NLP systems generally reduce the number of permissible syntactic structures to a limited amount of basic patterns on which a number of syntactic operations are defined. MWLs, specifically VMWLs, provide critical data for NLP grammars, because their morpho-syntactic properties differ from those of compositional expressions in various ways: a certain amount of MWLs represents morpho-syntactic irregularities, because their constructions do not conform to regular syntactic patterns. In (8),

(8) Sie ist nicht ohne.
Lit.: She is not without.
Engl.: She is quite something.

the preposition *ohne* occurs in isolation, i.e., without a dependent NP. In (9) and (10),

(9) Sie bewahrt ruhig Blut.
Lit.: She retains quiet blood.
Engl.: She remains calm.

(10) Sie ist gut Freund mit ihm.
Lit.: She is good friend with him.
Engl.: She is good friends with him.

there is no morphological adjective-noun agreement within the noun phrases *ruhig Blut* and *gut Freund*. In these cases of morpho-syntactic irregularities, the conventional grammatical patterns for compositional expressions are too restrictive to accomodate the MWLs.

In most cases, however, the grammatical patterns specified for compositional expressions are too permissive, because MWLs are subject to various types of morpho-syntactic constraints:[13]

- The number and determiner of noun phrases may be morpho-syntactically fixed. In (11),

  (11) She has thrown in the towel.

  any number or determiner variation would cause the phrase to lose its idiomatic meaning, as in (11').

  (11') She has thrown in a towel/towels.

- The possibilities of modifying an NP within an VMWL are also quite restricted. Any modification of the NP *the towel* by an adjective or a genitive NP causes the sentence (11) to lose its idiomatic meaning, as can be demostrated by (11").

[13] A detailed investigation of various types of constraints is given in [5].

(11'') He has thrown in a blue towel / Peter's towel.

- There are various constraints on the syntactic operations which can be applied to MWL constituents. Due to their complex internal structure, VMWLs are the most interesting group in this regard. In German, the constraints affect operations such as passivization, topicalization, clefting, wh-questioning, scrambling, and coordination (cf. [9]).

However, not every VMWL is subject to all types of restriction. There is a large number of VMWLs which behave more or less identical to their corresponding free syntagmatic constructions. Which syntactic operations and which modifications are possible depends on the referential properties of the MWL constituents involved (cf. [19]).

The relationship between the internal semantic structure of MWLs and their morpho-syntactic properties is still requiring further research, which should be based on examining text corpora.[14] A particular problem for corpus-based MWL research is caused by the fact that, especially in literary and newspaper texts, verbal idioms may be de-idiomatized in various ways in order to intend a pun. Along with the process of de-idiomatisation, all types of the above-described constraints can be intentionally ignored.[15] It is, however, important to note that these constraints are an essential prerequisite for the rhetorical effect which is intended by these puns.

3 Multiword Lexemes as a Problem for Machine Translation

The morpho-syntactic and semantic peculiarities of MWLs discussed in Section 2 pose various types of problems for the analysis, generation and transfer steps of machine translation systems. The following collection of problems is not claiming to be complete. It rather has the function of motivating our proposal for describing MWL properties in NLP lexicons (cf. Section 4).

3.1 Multiword Lexemes in Analysis

The main difficulty in analyzing MWLs is to recognize MWLs as such. Verbal idioms are more difficult to identify if an idiomatic and a literal reading are both possible, as in (12):

(12) John kicked the bucket.

Knowledge of morpho-syntactic constraints on verbal idioms can be used to resolve such ambiguities. For example, a detailed description of the morpho-syntactic properties of the verbal idiom *to kick the bucket* would enable us to determine that (13) and (14)

[14] Empirical research on these questions is discussed in, e.g., [13], [15], and [9].
[15] Cf. [21] and [20].

(13) The bucket was kicked by John.
(14) John kicked the empty bucket.

can only be analyzed in a literal sense. In addition, preference rules can be used in cases where one of the readings is more plausible or more frequent (like the idiomatic reading in *He will bite the dust.*). However, there are cases in which ambiguity can only be resolved by means of discourse, as in the German sentence (15):

(15) Er nahm das Kind auf den Arm.
Lit.: He took the child on the arm.

Depending on the context, (15) has either an idiomatic reading (*He pulled the child's leg.*) or a literal reading (*He picked up the child.*). A complex problem even for human translators are puns based on the ambiguity between a literal and an idiomatic reading of a construction (cf. [21] for examples). Such cases are likely to remain beyond the capacity of machine translation for some time.

3.2 Multiword Lexemes in Transfer

The complexity of transfer difficulties caused by MWLs depends on the degree of structural and lexical correspondence between a MWL and its equivalent in the target language. The following cases have to be distinguished:[16]

- There are cases of total lexical and structural correspondence as in (16):

  (16) die Katze aus dem Sack lassen <->
  to let the cat out of the bag

  in which, from an NLP point of view, the verbal idiom does not need to be identified as such, as long as selectional restrictions are not violated.

- Other VMWL equivalents display an analogous internal structure but lexical differences, which are usually motivated by different figurations:

  (17) einen Frosch im Hals haben <-> avoir un chat dans la gorge
  (18) zwei Fliegen mit einer Klappe schlagen <->
  to hit two birds with one stone

  In (17) and (18), the verbal idioms must be identified as such but can be translated in a straightforward way, provided that the lexical correspondences between idiom chunks of source and target language are defined in the transfer lexicon. The transfer lexicon entries may then contain surprising entries, particularly when the equivalents cast a different perspective on the same event, as in (19):

  (19) sich das Leben nehmen <-> se donner la mort

[16] Cf. [5] for a detailed typology of MWL translation equivalence.

in which the German words *Leben (life)* und *nehmen (take)* are translated into the French words *mort (death)* and *donner (give)* respectively.

- Problems increase if MWL equivalents differ with respect to both internal structure and lexical components, as in (20) and (21):

  (20) jdm. einen Baeren aufbinden <-> to take sb. for a ride
  (21) unter dem Pantoffel stehen <-> to be hen-pecked

  When MWLs of this type are modified, the target MWL must often be restructured in a complex way. Sentences such as (22),

  (22) Er stand unter dem Pantoffel seiner Chefin.

  have to be paraphrased rather than translated, because the English equivalent offers no possibility of reproducing the modifier *seiner Chefin (his boss)*. Complex restructuring processes may also be necessary if an MWL in the source language corresponds with a simplex verb in the target language, as in (23):

  (23) Nutzen ziehen aus <-> to benefit from

  If the nominal component of such a construction is modified by an adjective, as in (23'), the adjectival modifier has to be transformed into an adverb:

  (23') grossen Nutzen ziehen aus <-> to benefit largely from

- Similar problems arise if MWL equivalents are modified in a different way: the nominal part of the English SVC *to take into consideration*, for instance, can be modified by an adjective. The corresponding German construction *in Betracht ziehen*, in contrast, only allows for adverbial modification of the whole construction:

  (24) He took his objections into careful consideration.
  (24') Er zog seine Bedenken sorgfaeltig in Betracht.

  Therefore, the adjective modifier of the Npred in the English sentence (24) must be realized as an adverbial modifier in the German translation (24').

3.3 Multiword Lexemes in Generation

In the generation step of MT the structure of the target multiword lexeme has to be generated and correctly embedded in the target sentence. To fulfil this purpose the system needs detailed specifications of the morpho-syntactic constraints on MWL constructions. The handling of syntactic irregularities implies particular problems: sentences such as (8), (9), and (10) have to be generated, although they do not conform to regular syntactic patterns. In the analysis step, such irregular constructions may be handled using robust analysis techniques. In the generation step, however, mechanisms must be provided to handle irregular structures which occur only in particular MWLs. Highly inflected languages are greatly affected by this problem, since such irregularities often have to do with inflection features.

4 Description of Multiword Lexemes in NLP Lexicons

In this section, we list those types of information which should be included in NLP lexicons as the neccessary minimum.[17] In Section 5, we will use this list to evaluate different methods of lexical acquisition by checking it against the set of information types which can automatically be extracted. This will help us to identify those types of informatio n which must be specified manually in the course of NLP lexicon development.
NLP lexicons should include the following types of information on SVCs:

1. Morpho-syntactic formation of the Npred.
2. Morpho-syntactic formation of SVC arguments.
3. Adverbial vs. attributive modifiability.
4. Semantic contribution of the SV.

NLP lexicons should include the following types of information on idioms:

1. Part of speech of the idiom.
2. Internal syntactic structure of the idiom.
3. Lexical components of the idiom.
4. Idiom-external arguments.
5. Internal semantic structure of the idiom.
6. Lexical variability of the idiom chunks.
7. Morpho-syntactic variability of the idiom chunks.
8. Modifiability of the idiom chunks.
9. Restrictions on syntactic operations applicable to the idiom.

With regard to the complexity of the information, the use of intelligent lexicon formalisms with deduction components which support default inheritance would be of great benefit. This way, only information which cannot be derived from general principles must be specified directly. Sensible default assumptions are, for instance, that verbal idioms are lexically and morpho-syntactically fixed, that noun components should not be modified, and that only those syntactic operations may be applied to idiom chunks, which may also be applied to other types of non-referential expressions (e.g., expletive *it*; cf. [15], [9]). The general principles governing the passivization of verbal idioms still need to be investigated in greater depth, as does the influence of metaphoric contents on the morpho-syntactic flexibility of idiom chunks.

5 Acquisition of Multiword Lexemes from Text Corpora and from Machine-Readable Dictionaries

In the previous section, we saw that NLP lexicons must contain complex information on MWLs and their properties. Since the development of complex MWL

[17] A more detailed list, which takes different types of translation equivalence into account, can be found in [5].

descriptions is a time-consuming and laborious enterprise, it is sensible to check first whether existing machine-readable lexical resources can facilitate the task. In the following section, we shall report on several feasibility studies carried out at the University of Tübingen with the aim of automatically extracting information on German MWLs from text corpora and machine-readable dictionaries.

5.1 Corpus-based Acquisition of Support Verb Constructions

One feasibility study, carried out by E. Breidt (cf. [1]), uses text corpora and statistical methods for the extraction of German SVCs and noun-verb collocations.[18] In this study, the Mannheimer Korpus 1 (MK1)[19] was used as text corpus together with tools which calculate two statistical measurements on the basis of bigram tables:[20]

- **Mutual information (MI):** Compares the joint probability p(x,y) that two words, x and y, occur together in the corpus within a predefined distance with the independent probabilities, p(x) and p(y), that x and y occur independently.
  The probability p(x) of x is calculated by dividing the total number of occurrences f(x) of the word x by the total number of word tokens occurring in the corpus. Mutual information values can give an estimate of the degree of association existing between pairs of words.
- **T-score:** A significance measure, which estimates the significance of the word associations relative to the corpus being used.

In the study, the infinitive forms of 16 German support verbs were taken as key words. MI and t-score measurements were taken in a six word window to the left of the respective key-words for bigrams with a frequency of at least three occurrences. Precision was calculated for the set of all word pairs with a t-score greater than 0.6. The precision varied from 57 to 91 percent, the average precision being 72.7 percent.[21] These results are not as convincing as the results achieved in studies which used similar methods to extract English collocations. This is due to characteristic properties of the German language which make the task of extracting interesting collocations from text corpora more difficult than for English:

[18] Noun-verb collocations, such as *to pay attention to*, are habitual associations of verbs and nouns which cannot be predicted by the rules of the language system. Both the verb and noun have lexeme status and are therefore not MWLs according to our definition.

[19] The MK1 was made available to the University of Tübingen by the "Institut für deutsche Sprache" in Mannheim. It is a mixed corpus and contains about 2.7 million word tokens deriving from fiction, scientific texts, newpapers, and magazines.

[20] A detailed specification of the two measurements is given in [6].

[21] Precision was defined as the percentage of word pairs which are noun-verb-collocations or SVCs relative to the set of all word pairs detected with this method. The figures refer to the study reported in [1]; additional studies simulating lemmatized and part-of-speech-tagged corpora are described in [2]; the best ave rage precision in these studies was 87.6 percent.

- **Inflection**: German is a highly inflected language: the inflection paradigm of a strong verb with stem alternation like *kommen* includes three stem forms and in total 24 different word forms. The paradigm of a noun like *Haus* includes two stem forms and in total five word forms. Thus, in contrast to less inflected languages like English. one German word form only covers a small part of the complete inflectional paradigm of the searched lemma. To compensate for this drawback, the availability of a lemmati zed corpus would be of great benefit for the exploitation of German corpora.

- **Word Order**: In German sentences, the Npred and the SV of an SVC can appear in different positions and may be separated from each other by an unpredictable number of words:

  (25) **Sie kommen mit keinen Menschen in Beruehrung.**
  (26) **In Beruehrung mit Menschen kommen sie nicht.**
  (27) **Sie kommen mit Menschen, die ...., nicht in Beruehrung.**
  (28) **In Beruehrung mit Menschen, die ...., kommen sie nicht.**
  (29) **weil sie mit Menschen nie in Beruehrung kommen.**
  (30) **weil sie nie in Beruehrung mit Menschen kommen.**
  (31) **weil sie in Beruehrung mit Menschen, die ..., kommen.**

  The constituent order SV–Npred, as in (25), is unmarked in main clauses; but the order Npred–SV is also possible, as illustrated in (26). Examples (27) and (28) demonstrate that the distance between the SV and the Npred may be unpredictably long if the argument of the SVC (in our examples the PP *mit Menschen*) is further modified, e.g., by a relative clause. It is therefore obvious that rules such as *Semantic agent is used before the verb; semantic object after.* (cf. [17]), which were successfully used for English corpora cannot simply be applied to German.
  German subordinate clauses, however, demand the fixed constituent order Npred–SV, as illustrated in the examples (29)-(31). This explains why search windows to the left of the verb yield much better results than search windows to the right. But even in subordinate clauses, Npred and SV may be separated from each other by an SVC argument, as in (30), which may be extended by argument modifiers, as in (31).
  To overcome these problems caused by German word order, the availability of parsed corpora is a prerequisite.

In addition to these language-specific problems, corpus-based statistical methods have general limitations: they may detect SVCs, but they do not reveal their internal structure nor will they reveal the morpho-syntactic properties of their constituents. As a consequence, all relevant types of information on SVCs, given in Section 4, will have to be added by the NLP lexicographer.

5.2 Dictionary-based Acquisition of Support Verb Constructions

The dictionary resource used in our feasibility studies on dictionary-based methods was the Duden-Stilwörterbuch (Duden-2, cf. [8]). Duden-2 is a German

monolingual collocation dictionary containing descriptions of the combinatory potential of words in syntactic constructions and a well elaborated phraseological part. Parts of this dictionary were made available to us in machine-readable form for research purposes within the ELWIS project. The printsetting tapes were transformed into a lexical database by the dictionary entry parser LexParse (cf. [10]), so that the relevant phraseological items could be queried and accessed directly.

A quantitative comparison of dictionary and corpus results, as carried out for the verbs *kommen, setzen*, and *stellen*, showed that all SVCs and noun-verb-collocations which were detected in the MK1 by using statistical methods are also listed in the Duden-2. From a quantitative point of view, the dictionary is obviously a richer resource than the corpus. However, one should not conclude from this result that the use of text corpora is superfluous: the use of larger corpora than the MK1 might bring up some new collocations which are not recorded in the dictionary. In addition, corpus research offers the opportunity of deciding which collocations occur most frequently in a given text type. Thus, a combination of corpus and dictionary-based methods should produce good results, as long as the internal structure and components of the SVCs are specified in the dictionary.

However, the Duden-2 does not specify all types of information which are neccessary for an NLP treatment of SVCs: information about morpho-syntactic properties of the SVC, number and type of SVC arguments, and the semantic contribution of the SV is not explicitly accounted for. Moreover, there are inconsistencies with respect to the dictionary entry structure, which make it difficult to retrieve SVCs and noun-verb collocations automatically:

- In most cases, SVCs and noun-verb-collocations are not listed under the dictionary entry of the verb, but rather under the entry of the Npred. This is in line with metalexicographic claims (cf. [11]), which advise that collocations are to be included in the entry of the base component (which is the Npred) and not in the entry of the collocate part (which is the SV). Unfortunately, this strategy has not been followed consistently. As a consequence, one has to check the complete lexical database in order to retrieve all SVCs for a given SV.
- SVCs are to be found in different positions of the dictionary's microstructure: they may be part of the example group, part of the group on proverbial expressions, or they appear in the phraseological part. Systematic principles for the placement of SVCs and noun-verb-collocations within the dictionary entries would considerably facilitate their automatic extraction.

5.3 Acquisition of Verbal Idioms

We pointed out in Section 2 that verbal idioms quite often belong to informal and colloquial registers. This is most likely the reason why they do not appear frequently in a corpus of written language, such as the MK1. We made a KWIC (keyword-in-context) search for the headword *Kopf (head)* in the MK1 and found

examples for only 30 of the 74 idioms given in the Duden-2. 17 idioms with *Teufel (devil)* as a component are in the dictionary, but only four of them were found with the help of KWIC analyses in our corpus. As a consequence of the low frequency of instances, the overall result obtained by statistical methods is very unsatisfactory.

There are not only quantitative, but above all qualitative arguments in favour of the dictionary as the primary source: dictionaries contain explicit and implicit information on the internal and external structure of idiomatic expressions, which may, to a certain extent, be extracted by using pattern-matching methods. Another feasibility study was carried out using the Duden-2 to give an initial impression of the opportunities and limitations offered by such methods. The study showed that the following types of lexical information may be extracted (semi-)automatically:

- Argument slots are represented by indefinite pronouns like *jemandem* (abbreviated as *jdm.* and meaning *somebody*) and *etwas* (meaning *something*), which have the function of argument indicators carrying information on the case marking and the semantic type of the argument slot fillers. The argument indicator *jdm.* in the idiom description in (32)

  (32) jdm. ueber den Kopf wachsen
  Lit.: sb. over the head grow
  Engl.: to outgrow sb.

  indicates that the idiom has an argument slot filled by a noun phrase in dative case which denotes a living being. The idiom description in (33)

  (33) jdm. etwas auf den Kopf zusagen
  Lit.: sb. sth. to the head tell
  Engl.: to say sth. to sb.'s face

  indicates a dative slot for a living being (=*jdm.*) and an accusative slot for a non-living thing (=*etwas*). The argument indicators can be used to automatically extract information on the arguments of verbal idioms.

- The lexical components of idioms can be obtained if the argument indicators are removed from the idiom description. By removing the indicators *jdm.* and *etwas* from the idiom description in (33), one obtains the internal lexical components of the idiom in their canonical form, namely *auf den Kopf zusagen.*

- Morpho-syntactic flexibility can be inferred by the rule that no variation is possible unless specified otherwise. Possibilities of number and determiner variation are explicitly specified:

  (34) in [des] Teufels Kueche kommen
  Lit.: in the devil's kitchen come
  Engl.: to get into the hell of a mess

In (34), the brackets around *des* indicate that the definite article of the noun phrase *des Teufels* is optional and can also be omitted.

- For lexical variation a similar rule applies: lexical variation is not possible unless specified otherwise. It would, however, be helpful if two different types of lexical variants were explicitly differentiated:
  - Purely **idiom-internal variants** as in (35),

    (35) sich an den Kopf fassen/greifen
    Lit.: o.s. on the head grasp/seize
    Engl.: to throw one's hands up in despair

    in which the choice of the verbs *fassen* or *greifen* does not affect the denotative meaning of the idiom.
  - **Semantic variants** as in (36),

    (36) Kopf und Kragen riskieren/verlieren
    Lit.: head and collar risk/lose
    Engl.: to risk/lose one's head

    in which the verbs *riskieren* and *verlieren* make distinct semantic contributions to the meaning of the idiom.

  Both types of variants are separated by the same structure indicator, namely the slash. For MT purposes, however, the distinction is essential and has to be made explicit.

- Information on the modifiability of idiom components is implicitly given if one follows the rule "components cannot be modified unless specified otherwise" as a rule of thumb. However, the study showed that that the possibilities of component modification are not systematically accounted for in the Duden-2 and are not very reliable.

There are important types of information on verbal idioms which are not included in the dictionary at all:

The part of speech of idiomatic expressions is not explicitly specified nor can it be inferred from the part of speech of the headword of the dictionary entry in which the idiomatic expression is listed. This is due to the fact that an idiomatic expression is always listed in the dictionary entry of the first noun in its citation form or, if there is no noun, in the dictionary entry of the first content word. Thus, the verbal idiom *zwei Fliegen mit einer Klappe schlagen* is listed in the dictionary under the entry for the noun *Fliege*; and the adverbial idiom *mit Mann und Maus* is listed in the dictionary under the entry for the noun *Mann*.

No information is given on the syntactic operations which may or may not be applied. This information, like the information on part of speech, must be added manually relying on a theory concerning the syntactic and semantic properties of idioms.

6 Conclusion

The machine processing of MWLs requires detailed specifications of their morpho-syntactic properties in the lexicon. The task of drawing up such specifications is laborious and time-consuming, but is indispensable in view of the frequency of MWLs. We showed that the information on MWL properties which has to be encoded in the lexicon is quite complex. To manage this complexity the use of intelligent lexicon formalisms with deduction components supporting default inheritance would be of great benefit. This way, only information which cannot be derived from general principles must be specified explicitly. In order to take full advantage of the options offered by such formalisms, the general relationships between semantic structure, metaphoric content and morpho-syntactic flexibility should be further investigated.

The specifications of MWLs should be based on corpus-based research. However, the feasibility studies discussed in this paper showed that statistical corpus-based methods provide only partial information on SVCs and yield only poor results when applied to verbal idioms. The results demonstrate that statistical methods for extracting verbal MWLs which have proved successful for English cannot simply be applied to German.

If machine-readable dictionaries are available, the combination of corpus- and dictionary-based methods should produce better results than using just one type of source. From the quantitative and qualitative comparison of the Duden-2 dictionary and the MK 1 text corpus, we come to the conclusion that a dictionary, provided that it is of high quality, should be considered as the primary source for acquiring lexical knowledge.

A feasibility study using the Duden-2 has shown, however, that the information given on MWLs is neither extensive nor explicit enough to fully meet the requirements of NLP applications. Nevertheless, partial information on the morpho-syntactic properties of MWLs can be extracted and may then be completed by the MT lexicographer.

References

1. Breidt, E.: Extraktion von Verb-Nomen-Verbindungen aus dem Mannheimer Korpus I. SfS-Report 03-93, University of Tübingen, 1993
2. Breidt, E.: Extraction of V-N-collocations from text corpora: A feasibility study for German. First workshop on very large corpora, Ohio State University, Columbus Ohio, June 1993
3. Breidt, E.: Definition and classification criteria for verbal multiword lexemes. SfS Report, University of Tübingen, (to appear 1994)
4. Bresson, D.: Classification des verbes support (Funktionsverben) de l'allemand. In: Cahiers d'Etudes Germaniques 15, (1988) 53-65
5. Brundage, J., Kresse, M., Schwall, U., Storrer, A.: Multiword Lexemes: A monolingual and contrastive typology for NLP and MT. IWBS Report 232, IBM Institute for Knowledge Based Systems, Stuttgart 1992
6. Church, K., Gale, W., Hanks, P., Hindle, D.: Using statistics in lexical analysis. In: Zernik, U. (ed.): Lexical Acquisition. New York 1991

7. Danlos, L.: Support Verb Constructions. Linguistic properties, representation, translation. In: Journal of French Linguistic Study, Vol.2, No. 1 (1992), 1-32
8. Duden-2: Duden Stilwörterbuch der deutschen Sprache (ed. by G. Drosdowski, G.). 7th edition, Mannheim 1988
9. Engelke, S.: Eigenschaften von Phraseolexemen: Eine Untersuchung zu Möglichkeiten der Transformation und internen Modifizierbarkeit von somatischen verbalen Phraseolexemen. M.A. dissertation, University of Tübingen, 1994
10. Hauser, R., Storrer, A.: Dictionary entry parsing using the LexParse system. In: Lexicographica 9 (1993), Tübingen 1994, 174-219
11. Hausmann, F.K.: Kollokationen im Deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: Bergenholtz, H., Mugdan, J.(ed.): Lexikographie und Grammatik. Akten des Essener Kolloquiums 1984. Tübingen 1985, 118-129
12. Krenn, B., Volk, M.: DiTo Datenbank Datendokumentation zu Funktionsverbgefügen und Relativsätzen. DFKI-Document D-93-24, Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken 1993
13. Nicolas, T.: Internal modification of English V-NP-idioms. In: Everaert et al. (ed.): Proceedings of the Second Tilburg Workshop on Idioms 1992. Tilburg 1992, 85-96
14. Nunberg, G., Sag, I., Wasow, T.: Idioms. Manuscript, Stanford 1993 (to appear in Language 1994)
15. Schenk, A.: The syntactic behaviour of idioms. In: Everaert et al. (ed.): Proceedings of the Second Tilburg Workshop on Idioms 1992. Tilburg 1992, 97-110
16. Schwall, U.: Aspektualität - Eine semantisch-funktionelle Kategorie. Tübinger Beiträge zur Linguistik 344, Tübingen 1991
17. Smadja, F.A.: Macrocoding the lexicon with co-occurence knowledge. In: Zernik, U. (ed.): Lexical Acquisition: exploring on-line resources to build a lexicon. Hillsdale, NJ 1991
18. Storrer, A., Feldweg, H., Hinrichs, E.W.: Korpusunterstützte Entwicklung lexikalischer Wissensbasen. In: Sprache und Datenverarbeitung 17 (1993), Bonn 1994, 59-72
19. Van der Linden, E.-J.: A categorial, computational theory of idioms. OTS dissertation series, Rijksuniversiteit Utrecht 1993
20. Wiegand, H.E.: Kritische Lanze für das Fackelredensartenwörterbuch. Bericht und Diskussion zu einem Workshop in der österreichischen Akademie der Wissenschaften am 14.2.1994. In: Lexicographica 9, (1993), 1994
21. Wotjak, B: Verbale Phraseolexeme in System und Text. Tübingen 1992