Parsing decomposable idioms

Ingrid Fischer and Martina Keil
IMMDII, University of Erlangen
Martensstr. 3
91058 Erlangen, Germany
{idfische,keil}@informatik.uni-erlangen.de

Abstract

Verbal idioms can be divided into two main groups: non-compositional idioms as *kick the bucket* and compositional/decomposable idioms as *spill the beans*. In the following we will point to the fact that there are German decomposable idioms which can be decomposed into components, having identifiable meanings contributing to the meaning of the whole. These idiom components are taken to have referents. Taking these facts into account we propose an adequate way to represent the idiomatic meaning by Kamp's *Discourse Representation Theory (DRT)*. Furthermore, we show how to parse idiomatic sentences and how to process the proposed semantic representation. While parsing idioms, the necessary idiomatic knowledge of the idioms' syntax and semantics is extracted from a special idiomatic knowledge base called PHRASEO–LEX.

1 Introduction

Today it becomes more and more evident that a too restricted view on idiomatic phenomena is of limited use for the purpose of natural language processing. Therefore, it is now widely accepted that we have to distinguish at least two groups of figurative verbal phrasal idioms: first, there is a group of syntactically frozen idioms as *kick the bucket*, meaning "die", which are called *non–compositional*. Second, there is a group which shows more syntactic and semantic flexibility. An example for the latter group, often called *compositional* or *decomposable*[1] idioms, is *spill the beans* meaning "divulge information". With this group we are dealing here in depth. In this paper we propose an adequate semantic representation for idiomatic knowledge and show a way of processing syntax and semantics of decomposable idioms.

In the following we will first deal with the idea of decomposability of idioms in section 2. In section 3 we will present our proposal of an adequate representation of the idioms' meaning by means of DRT. Before we will outline a way of processing decomposable idioms in section 5, we will briefly introduce the necessary tools for the parsing process in a few brief words in section 4. Finally, in section 6 we show some possible extensions.

2 Decomposable idioms and the referential status of their idiom chunks

In contrast to *non-compositional* idioms, *decomposable* idioms are able to undergo several syntactic operations that lead to the opinion that "pieces of an idiom typically have identifiable meanings which combine to produce the meaning of the whole"(Wasow, 1982).

As example, we consider the syntactic behavior of the German verbal idioms *einen Bock schießen* (lit.: "shoot a buck", fig.: "make a mistake", fig. eq.: "make a bloomer")[2] and *jmdm. einen Bären aufbinden* (lit.: "tie sb. a bear on", fig.: "tell a tall tale to sb.", fig. eq.: "pull sb.'s leg"; "spin sb. a yarn")

In the following examples several modifications

[1] By classfying idioms with the terms *compositional* respectively *decomposable* the same property is described by two different point of views. The first notion is a more structural term, the second notion a more process–oriented term. See (Geeraerts, 1992).

[2] Since a high degree of language competence is necessary when judging about grammaticality of idiom constructions, we – as German native speakers – choose German idioms as examples. We establish the following convention for translation: literal: literal English word–by–word translation of the German idiom; figurative: English paraphrase of the figurative meaning; fig. equivalent: English idioms with an equivalent meaning.

(1) Tom hat auf der Sitzung *einen* *großen* *Bock* *geschossen.*
Tom has on the meeting a big buck shot.
Tom made a big mistake on the meeting.

(2) Tom *hat* in seinem Leben schon *einige* *Böcke* *geschossen.*
Tom has in his life already several bucks shot.
Tom already made several mistakes in his life.

(3) *Diesen* *Bock* hat Tom *geschossen.*
This buck has Tom shot.
Tom made this mistake.

(4) Tom bindet Kim *einen* *unglaublichen* *Bären* *auf.*
Tom ties Kim a unbelievable bear on.
Tom tells Kim an unbelievable tall tale.

(5) *Was* *für* *einen* *Bären* *hat* Tom Kim *aufgebunden?*
What for a bear has Tom Kim tied-on?
What kind of tall tale did Tom tell to Kim?

can be found: adjectival modifications in (1, 4), quantification in (2), and focussing by demonstrative determiner (3) and by question in (5) apply to the idioms internal NPs . It is important to notice that these operations and modifications in (1) (5) are not result of puns or word plays but grammatically and stylistically unmarked constructions.

Similar examples can be found in other languages, too. The first German example has a Dutch equivalent: *een bok schieten*, where internal modifications and quantification are possible. A french decomposable idiom is *lever un lièvre* (lit.: "raise a hare", fig.: "touch a delicate subject"); *prendere una cantonata* (lit.:"take a corner") meaning "to make a mistake" is an italian one. Therefore, internal modifiability of idioms seems not to be restricted on the German language.

2.1 Decomposable idioms are structured entities

It is evident that a component like *bucket* of a non compositional idiom as *kick the bucket* cannot undergo such kind of syntactic operations. Therefore, the meaning of non compositional idioms is seen as an unstructured complex. Components like *bucket* which do not carry any individual meaning are called *quasi-arguments* with a *non-referential function* (Chomsky, 1981). In opposite to this, components of decomposable idioms do carry some individual meaning! "Of course, these meanings are not the literal meanings of the parts" is stated in (Wasow, 1982). Then, the questions arise, which kind of meaning do these parts carry? Which is the *hidden semantic stuff* of *Bock* or *Bär* respectively, that is modified, inquired, quantified, and emphasized?

We adopt the point of view that items as *Bock* or *Bär* cannot be considered as quasi-arguments but as *figurative arguments*. Furthermore, we follow the opinion that such idiomatic strings are not unstructured complexes, but structured entities. Their structuring takes place in parallel to the structuring of the literally interpreted string (Burger, 1973). Our intuition suggests to paraphrase *einen Bock schießen* by "einen Fehler machen" (lit.: "make a mistake") and *jmdm. einen Bären aufbinden* by "jmdm. eine Lügengeschichte erzählen", (lit.: "tell a tall tale to sb."). It is evident for the paraphrase and the idiom to have at least the same syntactic structure as shown in the next table.

einen	Bock	schießen
a	*buck*	*shoot*
einen	Fehler	machen
a	*mistake*	*make*

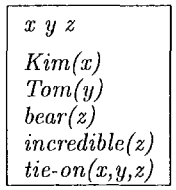
jmdm.	einen	Bären	aufbinden
to sb.	*a*	*bear*	*tie-on*
jmdm.	eine	Lügengeschichte	erzählen
to sb.	*a*	*tall tale*	*tell*

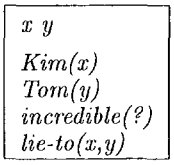
In addition it is important that also the semantics of the paraphrase and the idiom can be structured in parallel.

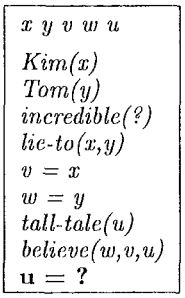
2.2 Figurative referents of idiom chunks

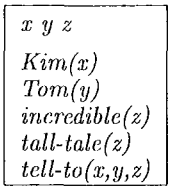
To explain this, let us now consider the problem from the referential point of view. We claim that individual components of decomposable idioms can be considered figurative arguments and that these figurative arguments have referents on their own.

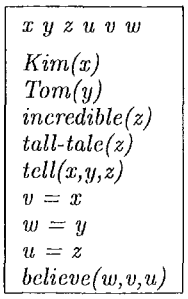
Following (Nunberg, 1978) who first discussed the referential aspect of idioms let "us consider that verb phrases 'refer' to states and activities, and transitive verb phrases normally refer to states and activities that are best identified as 'open relations' of the form *Rxb* where *'R'* stands for the relation referred to by the verb, *'x'* is a variable for the referent of the sentence subject, and *'b'* stands for the referent of the object NP."

(6)


(7)


(8)


(9)


(10)


On this basis, an idiom is called decomposable because the situation to which it refers can be seen as an open relation *Rxb*.

For the idiom *einen Bock schießen* this means that *schießen* is a two-argument relation with a variable for the subject NP, the noun phrase *einen Bock* referring to the concept *a mistake* and the verb *schießen* denoting a situation where someone is *acting*. Extending this idea to the decomposable idiom *jmdm. einen Bären aufbinden*, it is necessary to suppose a three-argument-relation *Rxyb* with two open variables: x represents the subject NP and y the indirect object NP. The idiom *jmdm. einen Bären aufbinden* is now decomposable into the noun phrase *einen Bären*, referring to a *tall tale*, and the verb *aufbinden*, referring to the activity of *telling*.

By paraphrasing decomposable idioms, the identifiable parts of meaning are taken into account. That means that the concept of the underlying referent, which often may be an abstract entity lacking a physical extension, should be verbalized and included into the paraphrase.

Notice that in the above cases the relation between the idiom components *Bock*, *Bär* and its paraphrased referents *Fehler*, *Lügengeschichte* is not a metaphorical one, but a conventional one. There are also decomposable idioms where decomposability is based on metaphoriocal knowledge.

Besides our introspective intuition, evidence for the proposed paraphrases is found through text analyses. The strongest support comes from the everyday usage of language being observed for example in textcorpora with newspapers, literature etc. (Keil, 1995).

3 Semantic representation of decomposable idioms via DRT

In the following, we will point out the problematic nature of meaning representation of idiomatic language with the help of DRT (Kamp, 1993). We will show the advantages of our theoretical considerations above, that can be best illustrated by DRT already including mechanisms to handle referents.

Consider example (4) *Kim bindet Tom einen unglaublichen Bären auf* (fig.:"Kim tells Tom an incredible tall tale"). DRS (6) shows the result of processing the — in this case senseless — literal reading of sentence (4) without any idiom handling procedures.[3] DRS (7) represents a non-compositional solution: after analysing the structure syntactically, the literal meaning of the multiword lexeme *jmdm. einen Bären aufbinden* is substituted by the "complex meaning" of the simple verb phrase as "jmdn. belügen" ("lie to sb."). Note that it is now a problem to represent the internal adjectival modifier *incredible* correctly. There is no discourse referent for that the condition *incredible* as semantic representation of the adjective *unglaublich* holds. Furthermore, if we want to represent the sentence *Er glaubte ihr die Lügengeschichte* ("He believes her the tall tale") — continuing example (4) —, the connection of the discourse referents cannot be made correctly as shown in DRS (8). The connection of the resumed constituent *einen unglaublichen Bären* and the resuming definite description *die unglaubliche Lügengeschichte*, which definitively exists, cannot be mapped into the DRS.

We claim that a more appropriate semantic representation of this idiom should respect its kind of composition and take its referents into consideration. On the base of the discussed paraphrase "eine Lügengeschichte erzählen", we offer the solution shown in (9). This representation now includes the condition *incredible(z)*, *tall-tale(z)*, *tell(x,y,z)* to represent the idiom. In (10) the continuation of our sentence is shown. Reference identity between *bear* and *tall-tale* is established by the equation *u=z*.

What decomposable idioms concerns, now the

[3]For the reason of simplification, we choose English predicate names for the conditions in the DRSs, e.g. instead of logical clauses as *bär(x)*, *aufbinden(x,y,z)*, or *belügen(x,y)* we present the sentence meaning with *bear(x)*, *tie-on(x,y,z)*, or *lie-to(x,y)*. This way the expenditure of translation can be reduced in this paper.

base for adequate anaphora resolution and resolution of definite descriptions resuming earlier introduced discourse material is created.

4 Used Tools: The Basic Parser and the Lexicon

In the following we introduce the tools we have used for parsing idiomatic sentences. We give a short description of the underlying chart–parsing–system (Fischer, 1995) and our idiomatic lexical database, called PHRASEO–LEX, that we use in the sense of an additional idiom list proposed by (Weinreich, 1969).

The design of our parsing system was governed by two main goals: *parallelism* and *incrementality*. Nevertheless different formalisms are used to represent syntactic and semantic features, having the advantage that for syntax as well as for semantics the most appropriate formalism can be chosen.[4] Consequently, to guarantee parallelism, this also requires a connection mechanism between these formalisms is necessary. In the following sections the structure of the parser will be described along these lines.

The **grammar formalism** of our system is an extension of the well known PATR-II. Syntactic information is encoded in feature structures. With the help of constraint equations these feature structures can be modified. The underlying unification mechanism is enriched with sequences as well as simple value disjunctions.

For our application the **semantic formalism** is of more interest. We decided to adopt Pinkal's approach (Bos, 1996) of DRT. In contrast to Kamp DRSs are not constructed in a top down fashion from a phrase structure tree, but bottom-up using a version of λ–calculus. When combining λ-calculus and DRT, two different kinds of abstraction are possible. First one can abstract over a complete DRS (*partial DRS*) or one can abstract only over a single discourse referent (*predicative DRS*). The following example shows both kind of abstraction with the λ–DRS for the indefinite determiner and the noun *mistake*.[5]

$$\lambda Q \lambda R \left[\begin{array}{|c|}\hline x \\ \hline \end{array}\right] \text{+Q(x)+R(x)} \quad \lambda x \left[\begin{array}{|c|}\hline \textit{mistake(x)} \\ \hline \end{array}\right]$$

Feature structures are used to encode the λ-DRSs. The main operation on λ- DRSs is the *functional compostion* on a partial DRS as functor and a predicative DRS as argument. It is implemented with the help of unification on the feature structures.

Our parsing engine is an **active chart parser**. The chart edges are marked, as usual, with category symbols. Additionally, syntactic feature structures and λ- DRSs are attached to every edge. For the extension of active edges according to the fundamental rule of active chart parsing, all syntactic and semantic constraints of the respective grammar rule must be satisfied.

A **grammar rule** consists of three parts: Context–free rules over category symbols constitute the backbone of every grammar. They are annotated with equations, the solutions of which result in syntactic feature structures. In these equations the category symbols are used as projections to mark the structures to be used. The category symbols are also used in the semantic operations on DRSs. For semantics, besides an operator *compose* for functional composition, an operator *id* for *identity* is used.

An interface module helps to connect different **lexicons** to the parser. At the moment, a syntactic lexicon containing feature structures, a semantic lexicon with λ–DRSs and a special lexicon for idioms, called PHRASEO-LEX (Keil, 1994), (Keil, 1995) are connected to the parser.

Phraseo-Lex is a computational lexicon which was specially developed for idiomatic knowledge. Of all diversed syntactic, semantic, and pragmatic information provided by PHRASEO-LEX, we only need for our purpose lemmata, base lexemes, (idiom participating lexical words: *Bock, schießen*), the internal syntactic structure encoded as a syntactic tree, the internal semantic structure encoded as predicate–argument–structure and the logical form.

As example, we show the lexical entries of our first examplary decomposable idiom.

lemma: `einen Bock schießen`
base lexemes: `Bock, schießen`
internal syntactic structure:

```
(vp11
      (np-acc (det einen)
              (n Bock))
      (v schießen))
```

internal semantic structure:

Idiom		Paraphrase
subject		subject
X	→	X
direct object		direct object
einen Bock	→	a mistake
predicate		predicate
schießen	→	make

logical form: `make(x,y), mistake(y)`

[4] This is the so called *co–descriptive approach*. Using the same formalism for syntactic and semantic construction is called the *integrated approach*, in the *descriptive approach* they are build up sequentially.

[5] + indicates the union of DRSs.

During the parsing process this necessary idiomatic information is extracted from PHRASEO-LEX and mapped into feature structures the parser can handle.

5 Processing decomposable idioms

When parsing decomposable idioms with the parser described in the previous section, the following steps are taken:

While **initializing the chart**, it is important to control whether potential parts of an idiom are found or not. For every word of a sentecne to be parsed it is checked if it is a base lexeme of an idiom. If this test was positive, an additional chart edge is inserted for every idiom the word can occur in. This edge is marked as usual, but with the syntactic feature structure and the λ–DRS built from the idiomatic information of PHRASEO-LEX.

The **feature structure** of this idiom edge contains information about how the idiom has to be completed and its underlying syntactic structure. This information is extracted from the PHRASEO-LEX syntax–tree. The following examples show the feature structures of *schießen* and *Bock* of our running example.

$$\begin{bmatrix} \text{agrm:} & \begin{bmatrix} \text{case:} & \text{nominative} \\ \text{number:} & \text{singular} \\ \text{person:} & \text{two} \end{bmatrix} \\ \text{stem:} & \textbf{schießen\_vpl13} \\ \text{val:} & \begin{bmatrix} \text{head:} & \begin{bmatrix} \text{stem:} & \text{bock\_vpl13} \\ \text{case:} & \text{accusative} \end{bmatrix} \\ \text{rest:} & \text{nil} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{agrm:} & \begin{bmatrix} \text{case:} & \text{nominative} \\ \text{number:} & \text{singular} \\ \text{person:} & \text{two} \\ \textit{gender:} & \text{masc} \end{bmatrix} \\ \text{stem:} & \textbf{bock\_vpl13} \\ \text{vpl:} & \begin{bmatrix} \text{verb:} & \text{schießen\_vpl13} \end{bmatrix} \end{bmatrix}$$

The features *val* (for valency) respectively *vpl* (for verbal phraseologism) contain the information necessary to find other relevant parts for building the idiom. While in the case of verbs the feature *val* just contains more information than usual, namely the stems of the missing parts of the idiom, the feature *vpl* is used to mark idiomatic information in other syntactic feature structures. Every part of the idiom is marked with an extra ending, in our example *_vpl13*. This is due to the fact that the same words can occur in different idioms and should not be mixed up during parsing, because of the corresponding semantic structures. For example, the words *Katze* and *Sack* occur as well in *die Katze aus dem Sack lassen* (fig. eq. "let the cat out of the bag") as in *die Katze im Sack kaufen* (fig. eq. "buy a pig in the poke").

The λ–**DRS** of the idiomatic edge already contains the literal referent of the part of the idiom they represent. This means the semantic entry for *schießen* as part of *einen Bock schießen* already contains the predicate *make(x,y)*.

$\lambda x \lambda y$ | *make(x,y)* $\qquad$ λx | *mistake(x)*

For the same reasons the λ–DRS for *bock* contains the predicate *mistake(x)*. This information is taken from the internal semantic structure of the idiom encoded in PHRASEO-LEX as shown above and translated into the λ–DRS.

It is important to notice that the information concerning decomposable idioms is distributed among all its base lexemes. Nevertheless, we only have *one* entry for every idiom in our idiomatic database. Only when initializing the chart, this information is spread over several edges.

In the **grammar**, special rules must be written to handle the idiomatic edges. In these rules it must be checked whether a complete idiom can be constructed. This is done with the help of extra equations over the special features *val* and *vpl* of the idiomatic feature structures. The following example shows a rule connecting an object and the verb phrase of a sentence, checking if both the verb and the noun are parts of the same idiom.[6]

```
(VP → V NP
 ((V val head)          =   (NP agrm)
  (V val head stem)     =   (NP stem)
  (NP vpl verb)         =   (VP stem)
  (VP val)              =   (NP val rest)
  (VP obj head)         =   NP
  (VP stem)             =   (V stem))
 (compose NP V))
```

No changes were necessary neither to the chart parser itself nor to the fundamental rule. All features concerning idioms are handled in the lexicons or the grammar.

The **result of the parsing process** are two readings of the sentence: the literal one, and the idiomatic one. The syntactic feature structures of the literal and the idiomatic reading are the same, as there is no pure syntactic difference between the two readings. Only the semantic structures differ: one DRS represents the literal idiomatic and one the idiomatic reading.

[6] Feature structures and rules are reduced to a minimum in our examples to keep the structures clear.

This technique allows us to parse sentences like (1)–(5) where one part of the idiom is modified and not the idiom as whole. A discourse referent for *bär* or *bock* respectively *tall tale* or *mistake* is already introduced during the initialization of the chart. This referent can serve as an anchor for an possible adjectival modifier as *unglaublich*. With the help of the rule connecting adjectives and nouns (not especially written for idioms!), the predicates *incredible(z)* and *tall-tale(z)* are inserted in the DRS. This approach also works for anaphoras. The discourse referent introduced for *B"ar* is the antecedent for the anapher in the next sentence.[7]

6 Extensions

It is quite simple to add the processing of non-compositional idioms to our parser. In this case, the whole literal meaning is bound to the main part of the idiom, in most cases the verb. The semantic of all the other parts is considered empty, the empty λ-DRS is bound to the corresponding edges. When parsing a sentence where a part of a non-compositional idiom is modified, the corresponding rules fail, because no discourse referent can be found this modification may be bound to. The only result will be the literal meaning of the sentence.

Our system starts processing a potential idiom as soon as one base lexeme was found. An improved version of our approach will handle an idiom after some more base lexemes appeared. This will reduce the number of lexical lookups to PHRASEO–LEX as well as the number of edges in the parser.

References

Bos, Johan; Gambäck, Björn; Lieske, Christian; Mori, Yoshiki; Pinkal, Manfred and Worm, Karsten. 1996 *Compositional Semantics in Verbmobil*, in this volume.

Burger, Harald. 1973 *Idiomatik des Deutschen*. Tübingen.

Chomsky, Noam. 1981 *Lectures on Government and Binding: The Pisa lectures*. Dordrecht/NL.

Fischer, Ingrid; Geistert, Bernd and Görz, Günther. 1995. Chart-based Incremental Semantics Construction with Anaphora Resolution Using λ-DRT. In *Proceedings of the fourth International Workshop on Parsing Technologies*, Prag and Karlsbad: 87–88.

Geeraerts, Dirk. 1992 Specialisation and Reinterpretation in Idioms. In M. Everaert, E.-J. van der Linden, A. Schenk, R. Schreuder (eds), *Proceedings of IDIOMS. International Conference on Idioms*, Tilburg/NL: 39–52.

Kamp, Hans and Reyle, Uwe. 1993 *From Discourse to Logic*. Kluwer Academic Press.

Keil, Martina. 1994 Systematische Repräsentation verbaler Phraseologismen und deren Eigenschaften im Lexikon. In Trost, Harald. (ed.) *Proceedings of KONVENS'94*, Springer, Wien/Austria: 181–190.

Keil, Martina. 1995 *Modell zur Repräsentation verbaler Phraseologismen (Phraseo-Lex)*. PhD thesis, University of Erlangen-Nuremberg.

Nunberg, Geoffrey D. 1978 *The Pragmatics of Reference*. PhD thesis. Reproduced by the Indiana University Linguistics Club, Bloomington.

Nunberg, Geoffrey D., Ivan A. Sag, Thomas Wasow. 1994 *Idioms*. Technical report, University of Stanford.

Pulman, Stephen G. 1993. The Recognition, Interpretation of Idioms In M. Everaert M, E.-J. Van der Linden (eds.) *Proceedings of the first Tilburg Workshop on Idioms*, Tilburg/NL: 249–270.

Stock, O. 1989 Parsing with flexiblity, dynamic strategies, idioms in mind. In *Computational linguistics*, 15, 1: 1–18.

Wasow, Thomas, Ivan A. Sag, and Geoffrey Nunberg. 1982 Idioms. An Interim Report. In S. Hattori, I. Kazuko, (eds.) *Proceedings of the XIIIth international congress of linguists*, Tokio: 102–115.

Weinreich, Uriel. 1969 Problems in analysis of Idioms. In J. Puhvel (ed.) *Substance and Structure of language*, Berkeley, University of California Press: 23–81.

[7] Similar results can be found in (Stock, 1989) where Italian idioms are the base. Here the mapping of parts of the idiom to parts of the paraphrase is done with the help of special referent substitutions between the literal and paraphrased meaning. Unfortunately, it is not described in detail how their mechanism would work for anaphoras or modifications of parts of the idiom.