# The Formation of Idiomatic Structures

André Schenk

Philips Research Laboratories
P.O. Box 80000
5600 JA Eindhoven The Netherlands

## 1 Introduction

In this paper we will discuss the formation of idiomatic syntactic structures in the grammatical framework of the Rosetta machine translation system. It will be shown that the syntactic representation of idiom structures has certain characteristics. To generate idiom structures in accordance with these characteristics, idiom formation has to meet certain criteria.

The paper has the following sectioning. In subsection 1.1, an informal description of idioms is presented, to give an indication of what is to be understood as idioms in this paper. Subsection 1.2 will discuss the central problems in a machine translation environment with regard to idioms and it will be indicated which of these problems will be discussed in this paper. In subsection 1.3 characteristics of idioms will be discussed and, on the basis of this, criteria will be formulated which a theory on idiom formation should meet. Subsection 1.4 will discuss the representation of idioms in syntax. Section 2 will present idiom formation in the grammatical framework of the Rosetta systems given the criteria stated in subsection 1.3; an example of the lexical entries for an idiom will be given and some apparent exceptions will be discussed. In section 3 some concluding remarks will be made.

### 1.1 Informal Description of Idioms

Informally, we can define idioms as expressions consisting of more than one word, for which a literal interpretation does not give the correct meaning. Examples are given in (1).

(1) a de plaat poetsen
(lit. polish the picture, id. sling one's hook)
b de pijp aan Maarten geven
(lit. give the pipe to Maarten, id. opt out)
c de pijp uitgaan
(lit. go out of the pipe, id. kick the bucket)
d iemand laten zitten
(lit. let sb. sit, id. leave sb. in the lurch)
e kick the bucket
f spill the beans
g break sb.'s heart
h lose one's cool

For example, (1a) in a literal reading can mean polish a certain picture, in the idiomatic reading the expression means approximately *leave*. The above characterisation is too general to indicate what is meant by the term idiom, because it also includes expressions such as semi-idioms or collocations as in (2a), composite predicates as in (2b), proverbs as in (2c), pragmatic idioms as in (2d), if the intended meaning is 'please, close the window' and metaphors as in (2e), if the meaning is 'the old man has his eyes on the lovely young girl'.

(2) a voorkeur hebben voor
(have preference for)
b een demonstratie geven
(give a demonstration)
c je moet de huid niet verkopen voordat de beer geschoten is
(do not sell the chickens before they hatch)
d het is koud hier!
(it is cold in here!)
e de oude buizerd aast op de kleine engel
(the old buzzard is preying on the little angel)

Expressions such as those in (2a-e) have characteristics that are different from those of idioms and thus they are not discussed in this paper.

Note that we have not given a definition of idioms. Such a definition is dependent on the grammar, similar to the definition of e.g. nouns, verb phrases or subjects. A definition of idioms would necessitate an extensive discussion of the grammatical framework of Rosetta and of the treatment of idioms in this framework. This is outside the scope of this paper. In this section, we have merely given a rough indication of what is to be understood to be an idiom in this paper.

## 1.2 Formulation of the Problem

The following four problems with regard to idioms are the most prominent ones in machine translation.

1. The translation relation between idioms in one language and expressions in another. This is evident, in a translation environment it has to be possible to translate an idiom.

2. The description of and the explanation for the syntactic behaviour of idioms. Idioms can undergo syntactic operations, but sometimes they are reluctant to do so.

3. The representation of idioms in syntax. This is linked to the previous problem; the representation should support an adequate description of the syntactic behaviour of idioms.

4. The formation of syntactic representations of idioms. How, and on the basis of what is the syntactic representation of idioms generated?

This paper will be on the fourth problem. Given of the representation of idioms in syntax, as discussed subsections 1.3 and 1.4, in section 2 it will be explained how this representation is generated.
For this purpose it is not necessary to discuss translation. Furthermore, the discussion will be purely from a generative point of view, i.e. aspects concerning analysis will not be considered. For a discussion of the treatment of idioms in the Rosetta machine translation system, cf. Schenk (1986) and Landsbergen, Odijk and Schenk (forthcoming).

## 1.3 Characteristics and Criteria

In this section we will give some characteristics of the syntactic representation of idioms that are relevant to the present discussion, since they lead to criteria that idiom formation should meet.

The syntactic representation is a base form that serves as the starting point for the derivation of the sentence the idiom occurs in. To this base form syntactic operations are applicable. The Rosetta grammars are based on Montague grammar, so the syntactic operations, applicable to the base form, are represented in a derivation tree.

We can distinguish the following four criteria.
(i) As already stated in e.g. Fraser (1970) and Jackendoff (1975), the representation of an idiom in syntax has to be similar to its literal counterpart. More in general, it has to be similar in relevant aspects to non-idiomatic representations with roughly the same form, i.e. it is constructed from existing lexical items

and it is constructed in accordance with existing syntactic rules (for apparent exceptions cf. section 2.3). The grammar should guarantee this.
Firstly, this is necessary to account for the fact that for instance the sequence noun, article, verb, e.g. *bucket the kick* is not an idiom. By using existing rules this can be accounted for.
Secondly, in principle, idioms or idiom parts can undergo any syntactic operation a literal counterpart or, more in general, a similar non-idiomatic construction can undergo.
Consider the examples in (3). There is passivisation in (3a); in (3b) the verb is in the final position in the sentence, while it is moved to the second position in the sentence under verb second in (3c). (3d) gives an example of topicalisation and (3e) of wh-movement.

(3) a het bijltje werd er bij neergegooid
(lit. the small axe was thrown down next to it)
b ...omdat hij de pijp uit ging
(lit. ... because he went out of the pipe)
c hij ging de pijp uit
(lit. he went out of the pipe)
d Marie's hart brak hij
(lit. Marie's heart he broke)
e Wiens hart zei je dat hij gebroken had?
(lit. Whose heart did you say he had broken?)

It is easy to see that the syntactic operations in the examples are the same as for non-idiomatic sentences with a similar syntactic structure. By using existing rules and lexical items this can be accounted for.
The syntactic representation should reflect for the larger part the syntactic behaviour of idioms. We will not discuss this here because it is outside the scope of this paper.
So, the syntactic representation of idioms should be similar in relevant aspects to non-idiomatic representations with roughly the same form, but not exactly the same, because the representation of an idiom should indicate for the larger part the non-applicability of syntactic operations. There is no difference at all between the syntactic surface structures of an idiom and its literal counterpart.
(ii) Furthermore, it follows from the examples in (3) that the representation of idioms should be in a canonical form, i.e. the form to which no syntactic transformations have applied (for example, if there is a passive transformation, then the active form is canonical). If this were not the case, then idiom formation should be such that every possible order of the idiom parts is generated taking into account any constituents which may intervene.
(iii) The representation of the idiom has to indicate exactly what the constituent structure looks like and where free arguments have to be realized. Thus, the NP *het bijltje* in (3a) has to be represented as a direct object and not, for example,

as a temporal adverbial, because then it cannot be the focus of passivisation. The free argument *Marie* in (3d) has to be realized in the genitive position in the object NP and not, for example, in the VP.
(iv) Idioms have a limited format. We will not give an explanation for the limited format of idioms, but, as will be shown below, it is possible to indicate what the format of idioms can be[1].
In section 2, we will show that idiom formation in the Rosetta grammars meets the above five criteria[2].

## 1.4 Representation of Idioms in Syntax

As an example, we will give the syntactic representation of the idiom *de plaat poetsen* in figure (1).
In figure (1), a constituent structure is shown with categories such as VERBP or NOUN, and relation names such as subj or head. V1 is a syntactic variable. Note that the example is simplified; in the actual Rosetta grammars the syntactic structure is more complex, e.g. the attribute value pairs associated with each category have been left out.
Though this is the syntactic representation of the idiom *de plaat poetsen*, it should not be listed in the dictionary as such, since it does not meet criteria (i) and (iv) and there are some further problems.

---

[1] In fact, we do not have an explanation for the limited format of idioms and in the linguistic literature we have not come across fully adequate accounts for the format. Attempts to give an explanation for certain idiom types are made in e.g. Coopmans and Everaert (1988) and Arnold and Sadler (1987).

In Rosetta, we have a more pragmatic approach. By looking at idiomatic data, we extract the syntactic rules that are used in the formation of idioms and in that way we compile a grammar that defines the idiomatic structures of a language, cf. below.

[2] As far as we can see, none of the proposals made in the linguistic or computational linguistic literature meets all five criteria given here. For example, Abeillé and Schabes (1989) propose an approach to parsing idioms in a tree adjoining grammars framework. Criterion (i) is not met, because the lexical elements in an idiom are not related to existing lexical items. Furthermore, they do not express the fact that the form of articles follows from the noun that is the head of the NP in which the article occurs. Consider the examples in (i) below, in which selection of the article follows from properties of the noun. Due to lexical properties, the noun in (ia) takes the definite article *de* and in (ib) *het*. In (ic), while, due to lexical properties, the noun takes *de* as article, here, because *bijltje* is in the diminutive form the article has to be *het*.

(i) a de plaat poetsen
(lit. polish the picture, id. sling one's hook)
b het paard achter de wagen spannen
(lit. harness the horse behind the car, id. put the cart before the horse)
c het bijltje erbij neergooien
(lit. throw the small axe down next to it, id. call it a day)

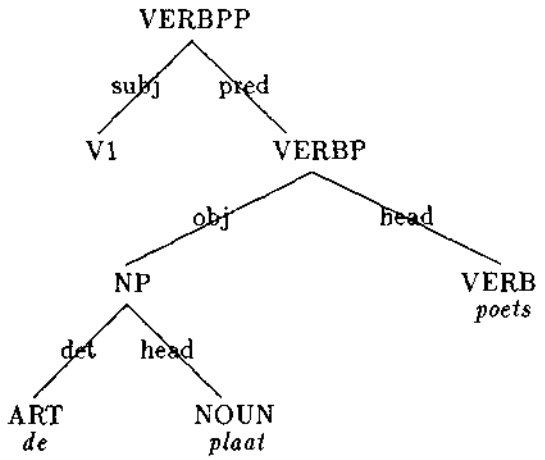


Figure 1: Syntactic representation of *de plaat poetsen*

Firstly, if the grammar has no explicit control over the representation, then it is impossible to impose some of the restrictions given in section 1.3, since:

1. there is no guarantee that the structure is similar to non-idiomatic structures.

2. without an extension of the lexical formalism, it is impossible to state restrictions on the format of idioms.

Secondly, the values of the attributes on the non-terminal nodes have to be specified explicitly, while some of these values follow directly from attribute values of the daughters of the non-terminal. For example, the NP that contains *de plaat* carries attribute values that follow from the attribute values specified at the article *de* and the noun *plaat*. These values cannot percolate because the structure is listed in the dictionary as such. These values have to be specified explicitly at the NP. Thus, from a linguistic point of view, the drawbacks are that information has to be stipulated that normally follows from rule application and, from a computational point of view, the drawbacks are that it is a lot of work to fill the dictionary and that it is easy to make mistakes.
In this section we have given an example of a syntactic representation of an idiom. Such a representation has to meet certain criteria. It has been shown that listing such a representation in the dictionary does not meet the criteria and that there are other problems. From this it follows that the syntactic representation of idioms should somehow be generated on the basis of the lexicon. We will call this process idiom formation. In the following section we will present extensions of the lexical formalism by means of which syntactic representations of idioms can be formed in accordance with the above criteria.

# 2 Idiom Formation

In this section, we will describe the formation of idiom structures. In subsection 2.1, we will describe the syntactic lexicon from the point of view of idioms. In subsection 2.2, we will give an example of the entries in the syntactic lexicon for an idiom. Finally, in subsection 2.3, we will discuss some apparent exceptions.

## 2.1 The Syntactic Lexicon

The syntactic component of a compositional grammar as used in a Rosetta system specifies a set of basic expressions and a set of syntactic or compositional rules. The basic expressions are the smallest meaningful units, i.e. they correspond to a primitive basic meaning. The set of compositional rules defines how larger expressions and, ultimately, sentences can be formed, starting with basic expressions. For a more elaborate discussion of the Rosetta framework and the organisation of the Rosetta grammars, cf. e.g. Appelo, Fellinger and Landsbergen (1987), Landsbergen (1987) and Odijk (1989).
The set of basic expressions is listed in a dictionary, which we will call the syntactic lexicon.

Some expressions are introduced syncategorematically by rules. Instead of incorporating such an expression with a full specification in a rule, they are listed in the syntactic lexicon and in the rule reference is made to this lexicon.
The syntactic lexicon, conceptually, now consists of two types of expressions: (i) expressions that do not correspond with a primitive meaning, i.e. the class of expressions that are introduced syncategorematically, and (ii) expressions that correspond to a basic meaning, i.e. the basic expressions.

In subsection 1.1 it was shown that literal interpretation of idioms does not lead to the correct meaning. The way this is accounted for in the Rosetta grammars is that an idiom is considered a basic expression with a primitive meaning.
Thus, conceptually, the set of basic expressions can be split up into two types of expressions: (i) the simple basic expressions, i.e. the basic expressions that do not have an internal structure and (ii) the complex basic expressions, i.e. the expressions that do have an internal structure; this type comprises idiomatic expressions.

For the formation of complex basic expressions the syntactic lexicon specifies: (i) the idiom grammar and (ii) a set of idiom derivation trees.
Idiom grammar specifies the set of rules that defines well-formed idiomatic representations. Thus, it defines the set of possible idiomatic structures in a language. The set of rules of idiom grammar is a subset of the set of compositional rules. In this way it is possible to guarantee minimally that representations of

idioms are well-formed and that they are similar to non-idiomatic constituent structures with approximately the same form. Furthermore, the subset is chosen in such a way that (i) it expresses the restrictions on the format of idioms and (ii) only canonical forms of idioms are made. The syntactic primitives, i.e. the leaves of an idiom, are specified in the syntactic lexicon.
Every element of the set of idiom derivation trees defines a representation of a specific idiom. Thus, an element of idiom derivation trees specifies which rules are applicable to which syntactic primitives to form the syntactic representation of a specific idiom. As the name indicates, such an element is represented as a derivation tree. Below, in figure 3, an example of an idiom derivation tree will be given.
The conceptual structure of the syntactic lexicon is represented schematically in figure (2).

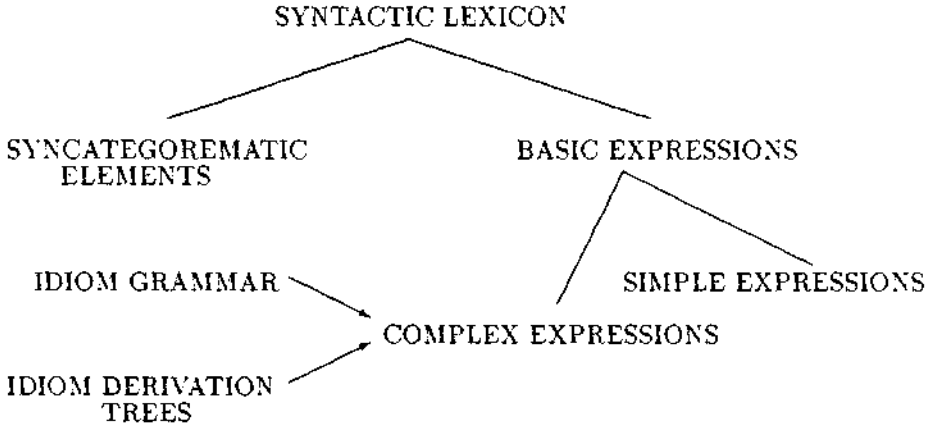


Figure 2: Syntactic Lexicon

## 2.2 The Entries for 'de plaat poetsen'

In this section an example will be given of the entries in the syntactic lexicon for the idiom *de plaat poetsen*. In figure (3) the idiom derivation tree for *de plaat poetsen* is given.
In figure (3), a start rule RSTARTVERB2 is applicable to a verb that takes two arguments, in this case *poets*, and it specifies that the first argument has to be realised in subject position, yielding (4a). A verb pattern rule TVERBPATTERN8 specifies that the second argument has to be represented as an object in the verb phrase, giving (4b). A substitution rule RSUBSTITUTION3 substitutes the NP *de plaat*, represented in (4c), which is made by rule RNP1 on

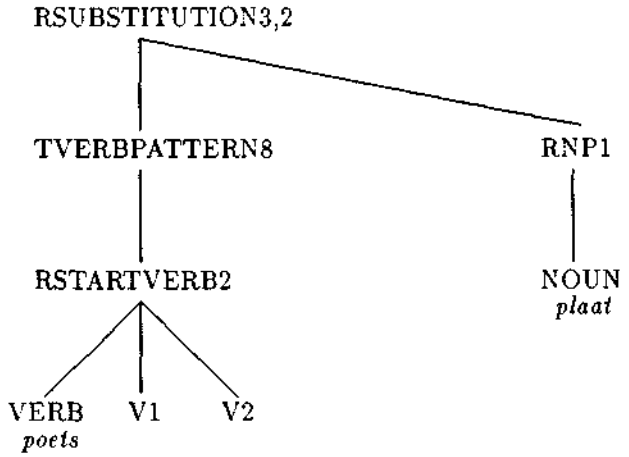


Figure 3: Idiom Derivation Tree for *de plaat poetsen*

the basis of the noun *plaat*, for the variable with index 2, yielding (4d), which was also represented in figure (1). The rules are taken from the set of syntactic rules and the lexical elements, cf. below, are taken from the syntactic lexicon.

(4) a VERBPP[subj/V1, pred/VERBP[arg/V2, head/VERB]]
b VERBPP[subj/V1, pred/VERBP[obj/V2, head/VERB]]
c NP[det/ART, head/NOUN]
d VERBPP[subj/V1, pred/VERBP[obj/NP
[det/ART, head/NOUN], head/VERB]]

Furthermore, there are the simple entries in the syntactic lexicon that are necessary for the idiom. Only some of the attribute value pairs are given:

**de** {category: article, definiteness: definite, ...: ...}

**plaat** {category: noun, gender: masculine, ...: ...}

**poets** {category: verb, reflexivity: notreflexive, ...: ...}

The article *de* is introduced syncategorematically in rule RNP1 of (3).

## 2.3 Apparent Exceptions

There appear to be exceptions to the claims made above, but at a closer look they may not be exceptional. These counterexamples can be found at word level and at the level of syntactic constructions.
(i) Word Level. Above, it has been stated that the syntactic lexicon consists of two types of expressions, i.e. the syncategorematic expressions and the basic

expressions. There is, however, a third class. For example, the word *brui* in (5a) and the word *lurch* in (5b) can only occur in the idioms given here.

(5) a er de brui aan geven
(lit. give the *brui* to it, id. chuck it in)
b leave sb. in the lurch

The way we account for this phenomenon in a Rosetta grammar is that this type of expression is listed in the syntactic lexicon with the appropriate attribute value specification, but it is not mapped on a basic meaning.

Words such as *brui* and *lurch* are an exception to the conceptual structure of the lexicon given above, so this structure has to be extended with this class. However, grammatically, the words of this class are completely 'natural'. For example, *lurch* behaves like any other noun; it takes *the* as an article and it is the head of an NP that behaves like other definite NPs (disregarding the deficient syntactic behaviour of idiom parts in general). Furthermore, *lurch* is natural from a morphological and a phonological point of view; it is a possible word of English (for example a word like *rluch* cannot occur in an English idiom, because it would violate phonological constraints).
It is also not exceptional that a syntactic element does not correspond to a basic meaning. For example, in English, there is existential *there* in the class of syncategorematic expressions, which is a noun or a pronoun, while *there* in the class of basic expressions is an adverb. The noun or pronoun *there* is not mapped onto a basic meaning.
So, words in idioms like *lurch* are not exceptional.
(ii) Syntactic Constructions. There are also exceptions at the level of syntactic constructions. Above, we claimed that idiom grammar is a subset of the set of compositional rules, but sometimes it looks as if in idioms different types of structures can appear when compared to non-idiomatic expressions. (6a) is a case of coordination of a preposition (*by*) and an adjective (*large*); in (6b) there is coordination of a noun (*kant*) and an adjective (*klaar*).

(6) a by and large
b kant en klaar
(lit. lace and ready, id. ready-made)

In actual Rosetta grammars, we account for this phenomenon by incorporating the rules that deal with these constructions in idiom grammar and not in the set of compositional rules. These rules then are not mapped on an expression in the meaning representation.

However, there is more to be said about these cases. In (6a-b), there is coordination of different categories. One can doubt whether such a coordinated construction is syntactically anomalous. Consider (7).

(7) a he is both wealthy and out of his mind
b this proposal is absurd and out of proportion

In (7), there is coordination of an adjective and a prepositional phrase. Linguistic theory should give an explanation for the fact that these sentences are well-formed. It is conceivable that it follows from this explanation that the coordination cases in idioms are not syntactically anomalous.

# 3 Conclusions

In this paper, we have discussed the formation of the syntactic representation of idioms. We have argued that this representation should meet certain criteria. Idiom formation, as described in this paper, guarantees that these criteria are met:
(i) By using existing syntactic rules and lexical items, the syntactic representation of an idiom is similar to its literal counterpart or, more in general, similar to non-idiomatic representations with roughly the same form. Although we have not dealt with this in this paper, the syntactic representation reflects for the larger part the syntactic behaviour of idioms.
(ii) Idiom grammar guarantees that only canonical forms of idioms are made.
(iii) An idiom derivation tree describes exactly the constituent structure of an idiomatic representation.
(iv) Idiom grammar guarantees that idiomatic representations have a limited format.

Furthermore, the problem of the percolation of information as described in subsection 1.4 has been solved by using rules to make the non-terminal nodes.

A final remark is that although the implementation of idiom formation in the actual Rosetta grammars at some points differs from the theory of idiom formation presented here, the implementation is consistent with the theory.

# Acknowledgements

The author would like to thank Jan Landsbergen, René Leermakers, Jan Odijk and Margreet Sanders for their comments on an earlier version of this paper.

# References

Abeillé, Anne and Yves Schabes (1989), *Parsing Idioms in Lexicalized Tags*, in: Proceedings of the European ACL Conference, pp. 1-9.

Appelo, L., C. Fellinger and J. Landsbergen (1987) *Subgrammars, Rule Classes and Control in the Rosetta Translation System*, Philips Research M.S. 14.131, in: Proceedings of the European ACL Conference, Copenhagen.

Arnold, Doug and Louisa Sadler (1987), *Non-compositionality and Translation*, Working Papers in Language Processing No. 1, Department of Language and Linguistics, University of Essex.

Coopmans, P. and M. Everaert (1988), *The Simplex Structure of Complex Idioms: The Morphological Status of 'laten'*, in: Everaert, M., A. Evers, R. Huybregts and M. Trommelen (eds), Morphology and Modularity, Foris, Dordrecht, pp. 75-103.

Fraser, B. (1970) *Idioms within a transformational Grammar*, in: Foundations of Language, Vol. 6, pp. 22-43.

Jackendoff, R. (1975), *Morphological and Semantic Regularities in the Lexicon*, Language, Vol. 51, no. 3, pp. 639-671.

Landsbergen, J. (1987), *Isomorphic grammars and their use in the Rosetta translation system*, Philips Research M.S. 12.950. Paper presented at the Tutorial on Machine Translation, Lugano, 1984, in: King, M. (ed), Machine Translation the state of the art, Edinburg University Press.

Landsbergen, J., J. Odijk and A. Schenk (forthcoming), *The Power of Compositional Translation*, Philips Research M.S. 15.427. To appear in: Literary and Linguistic Computing.

Odijk, J. (1989), *The organisation of the Rosetta grammars*, in: Proceedings of the European ACL Conference, Manchester.

Schenk, A. (1986), *Idioms in the Rosetta Machine Translation System*, Philips Research M.S. 13.508, in: Proceedings of the 11th Conference on Computational Linguistics, Bonn.