Compound Nouns in a Unification-Based MT System

Pierrette Bouillon **Katharina Boesefeldt** **Graham Russell**
ISSCO, 54 route des Acacias
1227 Geneva, Switzerland
pb@divsun.unige.ch

Abstract

This paper describes an approach to the treatment of nominal compounds in a machine translation project employing a modern unification-based system.

General problems connected with the analysis of compounds are briefly reviewed, and the project, for the automatic translation of Swiss avalanche bulletins, is introduced. Avalanche bulletins deal with a limited semantic domain and employ a sublanguage in which nominal compounds occur frequently. These and other properties of the texts affect the treatment of compounds, permitting certain simplifications, while leaving a number of possible alternative analyses.

We discuss the different problems involving the translation of compounds between German and French, and show how the computational environment in use permits two different approaches to the problem: an interlingua-based approach and a transfer-based approach. Finally, we evaluate these approaches with respect to linguistic and computational considerations applicable in a MT-system dealing with a limited semantic domain and describe the solution that has actually been implemented.

1 Compound Nouns

Compound words pose well-known problems for linguistic description in general, and some additional ones for natural language processing in particular:

Identification: how can compounds be distinguished from other words and phrases?

Segmentation: what are the components of a compound? In many languages, including German, orthographic convention is such that compounds are written as single units.[1]

Disambiguation: what is the correct analysis of a compound? On the widespread assumption that compounds have a recursive binary structure, any occurrences with more than two basic elements will admit multiple analyses,[2] from which, normally, a single candidate must be selected.

Interpretation: how can the meaning of a compound word be derived from the meanings of its parts? For many purposes, there is little point in performing any of the other tasks unless this is feasible.

It is clear that solutions to these four problems may be closely interrelated; ill-formed interpretations may permit unwanted analyses to be filtered out, the correct analysis will constrain possible segmentations, and so on.

In what follows, we outline an approach to the treatment of compounds within a specific limited application, the automatic translation of Swiss avalanche warning bulletins, which exploits the nature of the texts involved in order to translate compounds efficiently and correctly. We first give a brief overview of the project, paying special attention to phenomena related to compounds and their translation. We then describe ELU,[3] the software employed for translation, and discuss a number of the different treatments that it permits, motivating our choice of analysis with illustrations of their weak and strong points.

2 Compounds in Avalanche Bulletins

The avalanche bulletins are issued by IFENA (the Federal Institute for the Study of Snow and Avalanches)[4] a number of times a week during the winter season, the exact frequency of their appearance depending on weather conditions. Bulletins are prepared in German, and are translated into the other official Swiss languages, French and Italian, before publication. The current state of affairs, in which the source language is exclusively German, may change in future, and for this reason it has been decided to implement a reversible system (Bouillon and Boesefeldt, 1991a; Bouillon and Boesefeldt, 1991b).

Avalanche bulletins describe a fixed and specifiable semantic domain, and employ a language restricted in both vocabulary and syntactic variety. They contain a large number of compounds, with differing grammatical

[1] Elements of German compounds may however be separated by the so-called "Fugenzeichen" (*s*, *en*, etc.).

[2] See Church and Patil (1982: p.140ff.) for discussion.

[3] "Environnement Linguistique d'Unification" (Estival, 1990). See also Johnson and Rosner (1989) for a description of the earlier UD system on which ELU is based.

[4] The project is partially supported by IFENA.

properties. It is the interpretation of compounds, together with the closely related issue of structural disambiguation, which most researchers have addressed (Finin, 1980; Isabelle, 1984; Sparck Jones, 1983). In an application such as avalanche bulletin translation, however, the necessity for a deep or sophisticated analysis of compound meanings is not obvious. On the one hand, the number of compounds appearing in the sublanguage is highly restricted (approximately 400); they may therefore be listed exhaustively, and the question of disambiguation does not arise. And on the other, the interpretation of a compound is given by its established translation – the meaning of *Lawinengefahr* ('danger of avalanches') is just the information necessary to produce the required target language expression *danger d'avalanches* (or *pericolo di valanghe*, etc.).

In the current system, the main difficulty thus consists of establishing a systematic and general relation between single words in German and complex nominal structures in French, without defining the semantic and pragmatic roles of the different parts of the compound.

2.1 Variation in German Compounds

A purely categorial analysis of the German compound nouns contained in the avalanche warning bulletins reveals the following internal combinations of nouns, adjectives, verbs and prepositions:

[ADJ N]	*Neuschnee* 'new snow'
[N N]	*Lawinengefahr* 'danger of avalanches'
[V N]	*Schwimmschnee* 'floating snow'
[N [ADJ N]]	*Alpensüdhang* 'southern slope of the Alps'
[[ADJ N] N]	*Neuschneezuwachs* 'increase in new snow'
[[V N] N]	*Schwimmschneebildung* 'development of floating snow'
[[N N] N]	*Schneebrettgefahr* 'danger of snow patches'
[[PREP N] N]	*Überschneesprengung* 'explosion above the snow'

The 'head' of a compound is its final constituent which determines its properties (in the examples above, *-schnee*, *-gefahr*, *-zuwachs*, etc.), and may itself be a compound (*-südhang*, for example). Constituents of a compound are not inflected, although, naturally, the compound as a whole may be.

There is a very strong tendency to restrict the length of compounds to three elementary constituents; of approximately 400 compound nouns in the bulletins, issued over the past few years, only one (*Naßschneelawinengefahr*: 'danger of wet-snow avalanches') contains four.[5] Complex meanings that might be expressed by a fourth element within a compound generally occur in the form of a noun taking the compound as complement (e.g. *Abgang von Naßschneelawinen*, literally: 'going-down of wet-snow avalanches'), or conventional adjectival modification (e.g. *sehr grosse allgemeine Schneebrettgefahr* 'very great general danger of snow patches').

[5] This distribution can also be found in other domains (Boesefeldt, 1989) as well as in general language (Fleischer, 1982).

2.2 The French Translation

Compound nouns in the German original texts are usually translated into French as complex nominal phrases containing adjectival or prepositional subparts (e.g. *Neuschnee: neige fraîche*; *Schneebrettgefahr: danger de plaques de neige*). The details of translation differ in a number of aspects. First of all, the French phrases do not always contain the same syntactic categories as the corresponding German compounds. The noun-noun compound *Oberflächenschicht* ('surface layer'), for instance is generally translated by the sequence of noun and postnominal adjective *couches superficielles*.

Secondly, prepositions within French translations of German compounds vary from case to case: *Schneebrettgefahr* is translated as *danger de plaques de neige* *Schneeraster* as *grille à neige* and *Lawinenforschung* as *recherche sur les avalanches*. The use of articles within these PPs also varies: *température de la neige* ('snow temperature') has the definite article, while *quantité de neige* ('snow quantity') has none. Even though the use of the different prepositions and articles might contain clues concerning the possible semantic and pragmatic relation between the different elements of the compound (*à* denoting an instrument, an omitted article an abstract or partitive, etc.), the size of the corpus (only about 1000 words including nouns, verbs, articles etc.) makes it impossible to draw any general conclusion which could be implemented.

Finally, certain parts of German compounds are systematically omitted from the corresponding French phrase: *Schneebrettanriß* ('rupture of a snow patch' translates as *rupture de plaque*, where the word *neige* does not appear.

3 The treatment of compounds with ELU

Since 1990, the ELU system has been used at ISSCO for the implementation of a machine translation system for Swiss avalanche warning bulletins (Boesefeldt and Bouillon, 1991; Bouillon and Boesefeldt, 1991a; Bouillon and Boesefeldt, 1991b). ELU is a unification-based grammar development environment in the style of PATR-II (Shieber, 1986), designed especially for experimentation with machine translation.

ELU is composed of three modules: a parser, a generator and a reversible transfer module. Since it is the last of these that chiefly concerns us here, we shall say nothing about the parser or generator. ELU supports various methods of translation. By writing grammars that share meaning representations, one may implement an interlingua-based system; the result of analysing a source-language text is used directly as the basis for gen

erating the target-language text, and transfer is unnecessary. At the other extreme, ELU transfer rules are capable of performing arbitrary transformations on an input structure, and thus permit whatever variety of transfer the user desires.[6] A transfer-based approach to translation has been selected for the current project.

Transfer rules are written which state binary relations over the representations of texts from the source and target languages so as to associate the analysis of one language with the synthesis of another. There are two kinds of transfer rule, treating atomic and complex feature structures. For example, the following two FSs, representing the German phrase *die Gefahr* and its French equivalent *le danger*,

$$\begin{bmatrix} \text{PRED} & \text{gefahr} \\ \text{DETYPE} & \text{definite} \end{bmatrix} \qquad \begin{bmatrix} \text{PRED} & \text{danger} \\ \text{DETYPE} & \text{definite} \end{bmatrix}$$

are related by means of the following four rules:

```
:TA: gefahr danger

:TA: definite definite

:T:  pred            :T:  detype
:L1: <* pred> = X    :L1: <* detype> = X
:L2: <* pred> = Y    :L2: <* detype> = Y
:X:  X = Y           :X:  X = Y
```

The transfer rules establish a relation between the pairs of atoms **gefahr – danger** and **definite – definite** as the values of the attributes PRED and DETYPE in the two representations. The rules **pred** and **detype** abstract away from the particular values which their paths may take; a statement of the form "**:X: X = Y**" indicates that the success of the rule in which it appears is conditional on the success of other rules involving whatever structures are bound to **X** and **Y** during the transfer process. Transfer of a FS thus proceeds recursively, beginning with the root, and, provided that no failure occurs, terminating with the atoms which form the 'leaves' of the FS. For further details, see Estival et al. (1990) and Russell et al. (1991).

Two broad classes of solutions suggest themselves for the treatment of compounds in ELU: one, interlingua-oriented (cf. Section 3.2), in which the representations in the two languages are broadly similar, as in the *Gefahr – danger* example, and the other exploiting more fully the transfer mechanism (cf. Section 3.1). In the former case, it is possible to pass from a German compound to a complex French nominal phrase by employing general transfer rules of the kind required by other aspects of the translation task, while in the latter, different representations are related by complex transfer rules specific to the treatment of compounds.

3.1 The transfer-based approach

At first view, the transfer-based approach seems to be the more natural.

[6] "Codescription" analyses of the type proposed by Kaplan et al. (1989) are also possible.

In German, compounds can be introduced in the lexicon as simple words and treated in the same way by the grammar. The NP *eine Schneebrettgefahr*, for example, will be represented by a feature structure similar to that provided for the simple *eine Gefahr*. The semantics of adjectival modifiers is encoded as elements in a list, since the number of these may vary. Lists are indicated in attribute-value diagrams by '⟨...⟩' and in transfer rules by '[...]'; those shown here are empty:

$$\begin{bmatrix} \text{PRED} & \text{schneebrettgefahr} \\ \text{DETYPE} & \text{indefinite} \\ \text{MOD} & \langle\,\rangle \end{bmatrix} \qquad \begin{bmatrix} \text{PRED} & \text{gefahr} \\ \text{DETYPE} & \text{indefinite} \\ \text{MOD} & \langle\,\rangle \end{bmatrix}$$

The French grammar constructs the translation of a German compound from individual words, the head of the phrase being lexically specified for its complement and/or modifiers. For example, *danger* subcategorizes for a PP headed by the preposition *de*, and containing a NP which lacks an article and whose head noun is one of a class that includes *plaques*. The representation constructed for the French equivalent *danger de plaques de neige* will be considerably more complex than the German, with sub-FSs relating to the head (*danger*) and complement, which itself will be a complex structure with information concerning a head (*plaques*) and complement (*neige*). Since the prepositions in these constructions are determined by the lexical properties of the nouns, they convey no information in these cases, and do not appear in the representation; the French grammar accounts for their correct distribution.

$$\begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{PRED} & \text{neige} \\ \text{DETYPE} & \text{non} \\ \text{MOD} & \langle\,\rangle \end{bmatrix} \\ \text{PRED} & \text{plaques} \\ \text{DETYPE} & \text{non} \\ \text{MOD} & \langle\,\rangle \end{bmatrix} \\ \text{PRED} & \text{danger} \\ \text{DETYPE} & \text{indefinite} \\ \text{MOD} & \langle\,\rangle \end{bmatrix}$$

This technique makes it necessary to write a transfer rule for every compound in order to pass from the German compound to the different French words:

```
:T:  Schneebrettgefahr
:L1: <* pred> = schneebrettgefahr
:L2: <* args args pred> = neige
     <* args args detype> = non
     <* args args mod> = []
     <* args pred> = plaques
     <* args detype> = non
     <* args mod> = []
     <* pred> = danger
:X:  -
```

It presents a number of disadvantages. In the first place, certain generalizations which we consider important cannot be taken into account. For example, an element may appear in a number of different compounds, and the same information has to be repeated in each case.

For every compound including *Gebiet* ('region'), for example, we have to repeat the fact that the noun *Gebiet* translates as *région*:

```
:T:  Simplongebiet
:L1: <* pred> = simplongebiet
:L2: <* args pred> = simplon
     <* args detype> = non
     <* args mod> = []
     <* pred> = region

:T:  Engadingebiet
:L1: <* pred> = engadingebiet
:L2: <* args pred> = engadine
     <* args detype> = non
     <* args mod> = []
     <* pred> = region
```

Moreover, as this term also exists as a simple word, we also have to add a transfer rule to translate the simple word:

```
:TA: gebiet region
```

No account is taken of the fact that the elements inside the compounds correspond to isolated words and translate in the same way; moreover, a larger number of transfer rules is required, which, in turn, decreases the efficiency of the system.

The use of this technique leads to an additional problem. It is possible to add an adjective to a compound which already contains an adjective, e.g. *ganze Alpensüdhang* ('whole southern slope of the Alps'). This phrase has the following representation:

$$\begin{bmatrix} \text{PRED} & \text{alpensudhang} \\ \text{MOD} & \langle\ \text{ganz}\ \rangle \end{bmatrix}$$

The representation of equivalent French expression *tout le versant sud des Alpes*, however, is as follows:

$$\begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{PRED} & \text{alpes} \\ \text{DETYPE} & \text{definite} \\ \text{MOD} & \langle\ \rangle \end{bmatrix} \\ \text{PRED} & \text{versant} \\ \text{MOD} & \langle\ \text{tout, sud}\ \rangle \end{bmatrix}$$

For the transfer between the two languages it is necessary to establish a relation between the elements inside the lists. We cannot be sure, however, that the adjective which is a part of the German compound will always be at the same place inside the list in French. For this reason, it was decided to distinguish within the French description between adjectives which compose in German and adjectives which do not. This distinction which might seem undesirable at first sight has been proved satisfactory and generalizable for the sublanguage we treat and could be quite easily established.

An adjective cannot be found inside a German compound if it is part of a coordination, if it is modified by an adverb, if it has a particular morphological suffix or if it is a participle which is used adjectivally, if it has been given a certain semantic type[7] (adjectives qualifying snow always compose, adjectives expressing a degree of danger never compose), if there already is another adjective inside the compound or if the compound contains more than three elements. In all these cases it will therefore become the value of the path MOD : NOM and not the value of the new path MOD : N_COMP.

The representation of the nominal phrase *tout le versant sud des Alpes* is then modified as follows:

$$\begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{PRED} & \text{alpes} \\ \text{DETYPE} & \text{definite} \\ \text{MOD} & \langle\ \rangle \end{bmatrix} \\ \text{PRED} & \text{versant} \\ \text{MOD} & \begin{bmatrix} \text{NOM} & \text{tout} \\ \text{N\_COMP} & \text{sud} \end{bmatrix} \end{bmatrix}$$

The attribute MOD has thus been subdivided into two parts, MOD : NOM for adjectives like *tout* that do not compose and MOD : N_COMP for adjectives like *sud* that do compose.

The transfer rules for the transfer between *ganze Alpensüdhang* and the French *tout le versant sud des Alpes* can then be simplified and made more general:

```
:T:  Alpensudhang
:L1: <* pred> = alpensudhang
:L2: <* args pred> = alpes
     <* args detype> = non
     <* args mod> = []
     <* mod n_comp> = sud
     <* pred> = versant
:X:  -

:T:  nom
:L1: <* mod nom> = X1
:L2: <* mod nom> = X2
:X:  X1 = X2

:TA: ganz tout
```

This technique has been successfully implemented and was the first step towards the second solution to be presented here, the interlingua-oriented solution.

3.2 An interlingua-oriented approach

In order to obtain the same representation in German and in French, the German compounds have to be assigned in the lexicon a complex representation in which the word is semantically decomposed into head and complement(s). Moreover, certain information concerning the internal structure of the French compounds has to be added to the German representation during the analysis. This process can be done quite easily in ELU by means of macros.[8] The desired structures need only to be defined once, the adequate values being instantiated at the place of the variables.

[7] Lexical items are typed according to the contexts in which they may appear.

[8] ELU macros resemble a more powerful version of PATR-II 'templates', permitting the use of arguments, and multiple or recursive definitions.

The NP *Alpensüdhang* will thus receive the same representation as its French equivalent:

$$\begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{PRED} & \text{alpen} \\ \text{DETYPE} & \text{definite} \\ \text{MOD} & \begin{bmatrix} \text{N\_COMP} & \langle\,\rangle \\ \text{NOM} & \langle\,\rangle \end{bmatrix} \end{bmatrix} \\ \text{PRED} & \text{hang} \\ \text{MOD} & \begin{bmatrix} \text{N\_COMP} & \text{süd} \\ \text{NOM} & \langle\,\rangle \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{PRED} & \text{alpes} \\ \text{DETYPE} & \text{definite} \\ \text{MOD} & \begin{bmatrix} \text{N\_COMP} & \langle\,\rangle \\ \text{NOM} & \langle\,\rangle \end{bmatrix} \end{bmatrix} \\ \text{PRED} & \text{versant} \\ \text{MOD} & \begin{bmatrix} \text{N\_COMP} & \text{sud} \\ \text{NOM} & \langle\,\rangle \end{bmatrix} \end{bmatrix}$$

The paths ARGS : MOD : NOM and ARGS : DETYPE in the German representation have been added because they are added by grammar rules constituting French NPs and PPs.

This technique makes it possible to simplify the transfer rules, which can then be used for the transfer of not only compounds, but also other items:

```
:T:  ncomp
:L1: <* mod n_comp> = X1
:L2: <* mod n_comp> = X2
:X:  X1 = X2

:T:  nom
:L1: <* mod nom> = X1
:L2: <* mod nom> = X2
:X:  X1 = X2

:T:  pred
:L1: <* pred> = X1
:L2: <* pred> = X2
:X:  X1 = X2

:T:  args
:L1: <* args> = X1
:L2: <* args> = X2
:X:  X1 = X2

:T:  detype
:L1: <* detype> = X1
:L2: <* detype> = X2
:X:  X1 = X2

:TA: alpen alpes
:TA: hang versant
:TA: sud sud
:TA: definite definite
```

The semantic decomposition of the German compounds not only decreases the number of transfer rules, but also makes the transfer more general and coherent, as the different elements of a compound are translated as simple words.

The first problem of this technique consists of deciding what complex semantic representation to give German compounds. The decomposition could be performed according to the categories in either German or the target language. A disadvantage of the former is that, for the reasons given in section 2.2, transfer rules for irregular cases must be added. The latter approach avoids this difficulty by assigning the noun-noun compound *Sonnenlagen* ('sunny places'), for example, a representation in which the contribution of the element *sonnen-* and *-lage* are classified as modifier and predicate respectively, rather than predicate and argument:

$$\begin{bmatrix} \text{PRED} & \text{lage} \\ \text{MOD} & \begin{bmatrix} \text{NOM} & \langle\,\rangle \\ \text{N\_COMP} & \langle\,\text{sonnig}\,\rangle \end{bmatrix} \end{bmatrix}$$

This representation corresponds to the representation for the French equivalent, *endroit ensoleillé*:

$$\begin{bmatrix} \text{PRED} & \text{endroit} \\ \text{MOD} & \begin{bmatrix} \text{NOM} & \langle\,\rangle \\ \text{N\_COMP} & \langle\,\text{ensoleillé}\,\rangle \end{bmatrix} \end{bmatrix}$$

As *sonnig* is not the value of the path MOD : NOM, this representation does not interfere with the representation of the possible nominal phrase *sonnige Lage*:

$$\begin{bmatrix} \text{PRED} & \text{lage} \\ \text{MOD} & \begin{bmatrix} \text{NOM} & \langle\,\text{sonnig}\,\rangle \\ \text{N\_COMP} & \langle\,\rangle \end{bmatrix} \end{bmatrix}$$

But there is another reason to treat the irregularities between German and French in the German lexicon. Two different words can also be used as synonyms inside and outside of a compound (e.g. *Grisons* is translated by *Graubünden* in German, but *Grisons nord* is translated by *Nordbünden* and not by *Nordgraubünden*). *Nordbünden* has therefore been assigned a representation suitable for the generation of the French nominal phrase, *nord des Grisons*:

$$\begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{PRED} & \text{bunden} \\ \text{DETYPE} & \text{definite} \\ \text{MOD} & \begin{bmatrix} \text{NOM} & \langle\,\rangle \\ \text{N\_COMP} & \langle\,\rangle \end{bmatrix} \end{bmatrix} \\ \text{PRED} & \text{nord} \end{bmatrix}$$

$$\begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{PRED} & \text{grisons} \\ \text{DETYPE} & \text{definite} \\ \text{MOD} & \begin{bmatrix} \text{NOM} & \langle\,\rangle \\ \text{N\_COMP} & \langle\,\rangle \end{bmatrix} \end{bmatrix} \\ \text{PRED} & \text{nord} \end{bmatrix}$$

This representation for the German compound, however, does not seem very satisfactory because it would be necessary to write a special atomic transfer rule in addition to the transfer rule for *Graubünden*:

```
:TA: bunden grisons

:TA: graubunden grisons
```

Moreover, this would introduce an unwanted ambiguity when translating from French into German, since the French *Grisons* would now be related to both *Bünden* and *Graubünden*. *Nordbünden* is therefore assigned a representation in which *nord* and *Graubünden* are related as shown here:

$$\begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{PRED} & \text{graubunden} \\ \text{DETYPE} & \text{definite} \\ \text{MOD} & \begin{bmatrix} \text{NOM} & \langle\,\rangle \\ \text{N\_COMP} & \langle\,\rangle \end{bmatrix} \end{bmatrix} \\ \text{PRED} & \text{nord} \end{bmatrix}$$

Changes to lexical specifications can be made very easily by altering the value of the relevant path in the lexicon entry, without affecting the surface form of the lexical item.

Similarly, certain parts of the German compound have to be eliminated from the representation in the case of French nominal phrases in which certain parts of the German compound have been excluded (cf. section 2.2). This technique, however, cannot be applied if a compound without the element that is missing in French exists in German because the same French representation would then lead to the generation of two different nominal phrases in German.

In a number of cases the content of a compound can also be expressed by a complex nominal phrase with only slight stylistic differences (Fleischer, 1982: p. 20–21). The semantic representation used for compounds, however, is also used to express complex NPs in German; these constructions are also translated into French nominal phrases. In constrast to the case of adjectives within compound, which are assigned a special path in the representation (MOD : N_COMP), a complex NP in German, e.g. *Abgang von Lawinen* ('going-down of avalanches') will get the following representation:

$$\begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{PRED} & \text{lawinen} \\ \text{DETYPE} & \text{indefinite} \\ \text{MOD} & \begin{bmatrix} \text{NOM} & \langle\,\rangle \\ \text{NCOMP} & \langle\,\rangle \end{bmatrix} \end{bmatrix} \\ \text{PRED} & \text{abgang} \end{bmatrix}$$

which is the same as the representation for the compound *Lawinenabgang*, containing exactly the same information. In order to avoid over-generation resulting from the same representation being used for two different syntactic constructions, the possible syntactic structures in German have been restricted.

Analysis of the corpus of avalanche bulletin texts brought to light two types of case in which the same representation could be obtained for a compound and a complex NP. In the first, a fourth element either appears with a compound in a complex NP or has been included as an element of the compound itself. As both constructions contain exactly the same information, it was decided to eliminate the second possibility, and allow no more than three elements within a compound. For example, the word *Naßschneelawinengefahr* ('danger of wet snow avalanches') which is sometimes used instead of *Gefahr von Naßschneelawinen* does not appear in the lexicon, and the phrasal version is used instead.

It was also observed that, according to the degree of lexicalisation, nouns originally derived from a verb such as *Abgang* can form a part of a compound (e.g. *Neuschneeablagerung* – 'deposit of wet snow') as well as a part of a nominal phrase (e.g. *Verfestigung der Schneedecke* – 'consolidation of the snow cover'). As no general syntactic rule could be found to deal with this phenomenon it was decided to treat it on a case by case basis according to the occurrence of derived nouns in the avalanche bulletins.

The restrictions introduced into the German grammar, which are due to general language tendencies as well as to the language used in the avalanche bulletins, thus make it possible to avoid over-generation for German compounds without introducing an additional path for complex NPs. In case of the sublanguage we are treating, semantic decomposition therefore does not complicate the transfer as it is carried out according to the requirements of the target language. If phenomena of another language which might be added later contradict the current decomposition, specific transfer rules can be added to the system.

4 Conclusion

To conclude, we return to the four problems stated in section 1:

Identification of compounds presents no problem; a compound is listed in the lexicon just like any other noun, except that its representation is more complex.

Segmentation of compounds is not performed; the number and variety of compounds in the texts is sufficiently restricted, and the facilities of ELU sufficiently flexible, to make morphological analysis redundant.

Disambiguation of the structure of compounds is not necessary, since they are not morphologically decomposed, and since the

Interpretation of compounds is given by their lexical entry in combination with the transfer rules which accept meaning representations of sentences in which compounds appear.

In this paper we have outlined our approach to the treatment of compounds in a machine translation system dealing with the limited domain of avalanches in Switzerland. We have shown how we used ELU, a unification-based linguistic environment which has been developed for the implementation of machine translation systems for the testing of two different approaches to the translation of compounds: a transfer-based approach and an interlingua-oriented approach. The implementation of the these two approaches has shown that even though they have both proved satisfactory the interlingua approach seems more efficient and more general for the treatment of the sublanguage because it permits a rapid and coherent reversible transfer. The decomposition of the German compounds which is indispensible in an interlingua approach and did not complicate the system made it possible for us to solve the problems related to the translation of compounds which are the changing of

categories, the use of synonymy and the removal or addition of certain elements in the target language.

References

Boesefeldt, K. (1989) *Le problème de noms composés en allemand*, Ecole de Traduction et d'Interpretation, University of Geneva.

Boesefeldt, K. and P. Bouillon (1991) "Le rôle de la représentation sémantique dans un système de traduction multilingue," Working Paper no. 58, ISSCO, Geneva.

Bouillon, P. and K. Boesefeldt (1991a) "La Traduction Automatique des Bulletins d'Avalanches de la Suisse," *Colloque International sur l'Environnement Traductionnel*, Mons.

Bouillon, P. and K. Boesefeldt (1991b) "Applying an experimental MT system to a realistic problem," *Proceedings of Machine Translation Summit III*, Washington D.C., 1st–4th July 1991, 45–49.

Church, K. and R. Patil (1982) "Coping with Syntactic Ambiguity, or How to Put the Block in the Box on the Table," *Computational Linguistics* 8, 139–149.

Estival, D. (1990) "ELU User Manual," Technical Report 1, ISSCO, Geneva.

Estival, D., A. Ballim, G. Russell and S. Warwick (1990) "A Syntax and Semantics for Feature-Structure Transfer," *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, Austin, Texas, 11th–13th June 1990, 131–143.

Finin, T. W. (1980) "The Semantic Interpretation of Nominal Compounds," *Proceedings of the First National Conference on Artificial Intelligence*, Stanford, August 18th–21st, 1980, 310–312.

Fleischer, W. (1982) *Wortbildung der deutschen Gegenwartssprache*, Niemeyer, Tübingen.

Isabelle, P. (1984) "Another Look at Nominal Compounds," *Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Annual Meeting of the Association for Computational Linguistics*, Stanford, July 2nd–6th 1984, 509–516.

Johnson, R. and M. Rosner (1989) "A Rich Environment for Experimentation with Unification Grammars," *Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics*, Manchester, April 10th–12th 1989, 182–189.

Kaplan, R. M., K. Netter, J. Wedekind and A. Zaenen (1989) "Translation by Structural Correspondence," *Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics*, Manchester, April 10th–12th 1989, 272–281.

Russell, G., A. Ballim, D. Estival and S. Warwick-Armstrong (1991) "A Language for the Statement of Binary Relations over Feature Structures," *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, April 9th–11th 1991, 287–292.

Shieber, S. M. (1986) *An Introduction to Unification-Based Approaches to Grammar*, CSLI, Stanford.

Sparck Jones, K. (1983) "So what about parsing compound nouns?," in *Automatic Natural Language Processing*, K. Sparck Jones and Y. A. Wilks, eds., Ellis Horwood, Chichester, 164–168.