Patterns of Derivation

Oliver Streiter and Antje Schmidt-Wigger
IAI
Martin-Luther Straße 14
66111 Saarbrücken Germany
catler@iai.uni-sb.de

Contrary to inflection and compounding, derivation is a largely neglected topic m the field of MT. In this paper we stress the importance of morphological and syntactic patterns of derivation, the treatment of which a powerful NLP system, especially if it is MT-oriented, cannot dispense with. It will be shown that the suggested approach of full derivational analysis not only augments the consistency of lexical representations by an overall reduction in the size of monolingual and bilingual lexicons but also opens insight into the semantic nature of different parts of speech. Examples of implementation are taken from the CAT2 MT system as it is currently used in various MT projects funded by the CEC.

1   Introduction

Since the earliest days in the history of machine translation, the importance of the treatment of inflection has been recognized, especially in work on highly inflective languages like Russian and German. Full form dictionaries have been replaced by dictionaries with basic word forms as entries while the treatment of the inflected forms is carried out by autonomous morphological modules.

Contrary to inflection, compounding still causes notorious problems for MT applications. An exhaustive listing of compounds in the lexicon for languages such as Dutch or German is not possible, as any speaker can freely create new compounds (e.g. *Baufirma, Autofirma, Lederfirma* etc. . .). To cope with them systems must be able to identify the parts of these non-lexicalized compounds and their unmarked semantic relation, the recognition of which is necessary for the correct choice of the part of speech for the translated non-head (noun: **building** *firm* vs. adjective: *impresa* **edile)** and its functional markers (e.g. *entreprise* **de** *construction)* (cf. (STREITER *et al.* (1994)), (GAWROŃSKA *et al.* (1994))).

Derivation belongs to the most neglected topics in the field of MT, which may be due to the assumption that, like inflection, it has little impact on the construction of an MT system since the syntactico-semantic component can treat different derivations as independent entities. It is only recently that suggestions have been made to use patterns of derivation for MT purposes, in order to develop a concise lexicon which is flexible enough to respond to the requirements of translation (cf. (NOMURA *et al.* (1994)), (TALLOWITZ (1994))). In our approach derivationally related words are represented by the same lexeme, resulting in a coherent and reduced form of lexical representation. With respect to transfer, this approach allows for the implementation of a truly semantically driven transfer in which morpho-syntactic properties of the source language are of no importance except for the semantic representation they have triggered.

2   Transposition in Transfer

In this section we shall discuss phenomena of transfer, focusing on translations in which the part of speech changes. This **transposition** of the part of speech, as it is called in translation theory (cf. (PODEUR (1993))), is the most frequent translational operation to avoid an otherwise ill-formed word-for-word translation. We shall show how the traditional approach to transfer tries to model such operations, contrasting this approach to ours, which, we claim, overcomes the traditional shortcomings.

2.1 Constraints on the Part of Speech

In most modern transfer-based MT systems, the smallest transfer units are content words understood as concepts denoted by these words. Function words are no longer represented at the transfer stage and are typically replaced by an interlingual semantic annotation on the content words. Transfer rules are used to identify the target language item which denotes the same concept as denoted in the source language. This strategy for the selection of the lexical item, however, has to be such that mechanisms for paraphrasing are possible and controlled according to their semantic validity. Although linguistic models of paraphrases are available (cf. (MEL'ČUK AND PERTSOV (1987))), implementations often fall short of such mechanisms and stick to quite rigid methods of transfer, determining, for example, the part of speech within the transfer component. However, the choice of the part of speech, we shall argue, cannot be predicted within the transfer rules but must be determined by monolingual syntactic, semantic, stylistic and lexical constraints on the part of speech.

Syntactic constraints: To give an example of syntactic constraints on the part of speech: A verbal argument must be rephrased as a noun group in the target language if the target matrix verb does not subcategorize for verbal phrases, as can be seen in example (la).

> (la) Sie begrüssen es, daß ein Auto angeschafft wird.
> * They welcome that a car will be acquired.
> They welcome the acquisition of a car.

Another syntactic constraint comes from the obligatory realization of argument position; if obligatory arguments are not realized, paraphrases supporting this gap have to be found.

> (1b) Il embrouille. (He confuses)
> * Er verwirrt. / Er stiftet Verwirrung.
> (1c) Il embrouille les adversaires. (He confuses the opponents)
> Er verwirrt die Gegner. / Er stiftet bei den Gegnern Verwirrung.

Semantic Constraints: A semantic constraint is shown in (2a-b). The impossibility of a nominal construction to actualize the denoted event prevents the verb being translated as a deverbal noun, which is possible in a context like (la)[1]. While the verbal concept actualizes the proposition with the help of tense values, nominal concepts can not express tense unless they are carried by a support verb (e.g. 2a-b).

> (2a) Ein neues Auto wurde angeschafft.
> + The acquisition of a new car. [2] / A new car has been acquired.
> (2b) Er geht spazieren. (He takes a walk)
> + Sa promenade.[3] / Il fait une promenade.

Stylistic constraints: Besides syntactic and semantic constraints, stylistic considerations can also motivate a paraphrase. Where German, for example, frequently makes use of noun phrases, French prefers verbal constructions. Where French uses prepositional phrases, Italian prefers strongly adjectives (examples 3b-c come from (PODEUR (1993)) pg.43).

---

[1] By actualization we understand the passage from the virtual system (langue) to the actual process (parole) by the embedding of the prepositional content in a complex system of relations that are based on the speech act (cf. (GREIMAS AND JOSEPH (1989))). Tense, aspect and mood belong to the actualizing function of the verb.

[2] This phrase is well formed and excluded only as a translation of (2a).

[3] This translation would be perfect in a title or caption under a cartoon.

(3a) Er schlägt einen Spaziergang vor. (He proposes a walk)
    Il propose de faire une promenade.
(3b) entreprise de construction (building firm)
    impresa edile
(3c) esprit de compétition (spirit of competition)
    spirito competitivo

Lexical Constraints: A fourth constraint may come from the possible usages a lexeme is allowed to acquire in a language. Thus, while *technisch* can be used as an adverb in German, this usage is impossible in French.

(4a) Die Autos werden technisch kontrolliert.
    (The cars are technically controlled)
    Les voitures sont controllées concernant leur fonctionnement technique.
(4b) Sie sind technisch gut.
    (They are technically good)
    Ils ont une bonne technique.

In Russian, nearly any noun can form a relational adjective, something which in languages like German, French and English is less common. As a consequence, the adjective *fruktovyj* (fruit+SUFF) in (4c) has to be translated as a noun in these languages. As a further example for lexical constraints, Russian has few nouns denoting languages, e.g. the notion of the Russian language can be expressed only by adverbs (4d) or adjectives (4e).

(4c) fruktovyj sok - fruit juice
(4d) He speaks Russian. - On govorit po-russki.
(4e) Russian teacher – prepodavatel' russkogo jazyka

The transposition of the part of speech may affect beside the item in question all the other constituents it has to combine with. In such cases we speak of a **chain transposition.** If, for example, the concept underlying an adverb finds no adverbial but an adjectival realization in the target language, a nominal paraphrase of the verbal predicate saves the translation (4f). It goes without saying that in such cases prepositions and complementizers have to become mutually translatable (4g) and articles must be generated without a formal pendant in the verbal construction (4h).

(4f) travailler efficacement (to work efficiently)
    *nützlich arbeiten / eine nützliche Arbeit leisten
(4g) Do not open **before** train **stops!**
    Ne pas ouvrir **avant l'arrêt** du train!
(4h) Lavatory should not be used!
    L'usage du WC est interdit!

To sum up, in all these cases from (la) to (4h), an MT system has, in order to cope with such syntactic, semantic, stylistic and lexical constraints on realization, to fall back on words and phrases which have to be linked to the target concept by way of an appropriate derivation realized in syntax or morphology. How this can be done is shown in the following sections.

2.2 Representing Lexical Transfer

In most transfer-based MT systems, the relations between the source language unit, the underlying concept and the target language unit are formally represented by a Lexical Transfer Rule (LTR) (5a), the conceptual identity (6a) being implied.

(5a) **{lex=acquire} <=> {lex=anschaffen}**
(6a) **({lex=acquire} <=>** *Concept $_{acquire}$* **<=>** {lex=anschaffen})

Because of the general assumption that nominals are semantically different from verbal concepts, incompatible semantic descriptions are assigned to verbs and nouns in the linguistic literature (cf. (Zellnsky-Wlbbelt (1988)) and (Pollard and Sag (1994))). It is only natural that in this tradition most MT systems employ different LTRs for the translation of verbs and nouns (cf. (Schmidt (1988)), (Apresjan *et al.* (1989)), (Arnold and Sadler (1990)), (Whitelock (1992))):

(5a) **{lex=acquire} <=> {lex=anschaffen}**

(6a) **({lex=acquire} <=> *Concep $_{acquire}$* <=> {lex=anschaffen})**

(5b) **{lex=acquisition} <=> {lex=anschaffung}**

(6b) **({lex=acquisition} <=> *Concept$_{acquisition}$* <=> {lex=anschaffung})**

As for the conditions for the transposition of the part of speech, such information cannot be coded at the level of LTRs, since information about the context this word has to be integrated into (e.g. as argument, modifier or predicate) is necessary (see 1a-4h). The delegation of the change of part of speech to structural transfer rules, as suggested in (Somers *et al.* (1988)), is not possible since the governors, arguments and modifiers the item in question has to combine with are subject to the same degree of uncertainty, i.e. it is possible that they themselves must undergo a transposition, depending on their context. In addition, structural transfer rules, e.g. transfer rules which mention more than one constituent, are severely restricted in their possibility to combine with each other as outlined in (Arnold and Sadler (1987)) and (Dorr (1994)), indicating that such rules offer no definitive solution to the transfer problem. As a consequence, the possible change of the part of speech must be foreseen within the traditional transfer approach in all possible combinations by adding the corresponding translation rules to the transfer component.

(5a) **{lex=acquire} <=> {lex=anschaffen}**

(5b) **{lex=acquisition} <=> {lex=anschaffung}**

(5c) **{lex=acquire} <=> {lex=anschaffung}**

(5d) **{lex=acquisition} <=> {lex=anschaffen}**

It goes without saying that such a redundant representation is not desirable for various reasons, including considerations concerning memory size and man power necessary for the construction of such lexicons. In reality, most systems foresee a limited rank of transfer pairs to cope with the encountered problematic cases, running the risk of never completing the description of all possible paraphrases.[4]

In addition, assuming different semantic types for the different parts of speech signifies a renouncement of any semantic control when the part of speech shifts in transfer. In view of such facts, it is only natural to take into consideration derivation as a powerful and semantically motivated means which helps to avoid such redundant representation. Accordingly, we assume two principles:

P1 **Words related via the operation of morphological derivations are represented by the same lexeme.**

P2 **The semantic description of all parts of speech is of one semantic type, with potentially different semantic tokens, allowing for a meaningful match and mismatch of semantic information between different parts of speech.**

---

[4] This incompleteness can be found in all traditional transfer dictionaries. In the Russian to German transfer lexicon of SUSY, for example, we find LTRs which translate relational adjectives as adjectives (e.g. 1,3) or as nouns (e.g. 2,5), but in many cases necessary variations are missing (e.g. 4,6).

1.   *kvadratnyi $\Rightarrow$ quadratisch*
2.   *kvadratnyi $\Rightarrow$ Quadrat-*
3.   *gosudarstvennyi$\Rightarrow$ staatlich*
4. MISSING*: gosudarstvennyj $\Rightarrow$ Staats- (schulden)*
5.   *policejskij $\Rightarrow$ Polizei-*
6. MISSING: *policejskij $\Rightarrow$ polizeilich (-e Ermittlungen)*

Thus completely equivalent to (5a-d) is the CAT2-LTR which relates all concepts morphologically related to *acquire* to all concepts morphologically related to *anschaffen*.

(5) **{lex=acquire} <=> {lex=anschaffen}**

Taken alone, such a rule would be too unspecific to ensure valid translational relations. But in virtue of P2, we can assume a Semantic Transfer Principle (STP) (7) which states that the semantic space foregrounded in the source language must be foregrounded in the target language as well. Otherwise it is dismissed as an invalid translation. Thus (7) functions as meta-principle to rules of the type (5).

(7) **{head={ehead={sem=SEM}}} <=> {head={ehead={sem=SEM}}}**

It is this semantic control in transfer which has to limit the change of the part of speech and has to trigger the morpho-syntactic reflexes of the semantic content. How such a common semantic type can be described is subject to current investigations.

3   Semantic Components of Derivation

In our implementation, the following aspects of semantic description are held to be valid across all parts of speech: the notional domain and its reflexes in the argument structure and the referential argument.

3.1   The Notional Domain

The notional domain is an unanalysed semantic kernel for which the lexeme functions as a label. Strictly speaking, *"A notion can be defined as a complex bundle of structured physico-cultural properties and should not be equated with lexical labels or actual items. (...) they epitomize properties (...) derived from interaction between persons and persons, persons and objects, biological constraints, technical activity,* etc."((CULIOLI (1990)), pg.69). Such notional domains are derived from predicative relations such as events or states, including a number of participants which interact in a manner typical for this notional domain.

Within the notional domain we localize one or more concepts: *"It should be obvious that notions have status of predicable entities and could be described as unfragmented solid wholes; but they are apprehended through occurrences, i.e. distinguished through separate events, broken down into units (...) with variable properties"*((CULIOLI (1990)), pg.69-70) These concepts may refer differently to the notion as such (e.g. *to paint, painting),* its complement (e.g. *happiness* vs. *unhappiness)* or to participants of the predicative relation (e.g. *painter, a painting).*
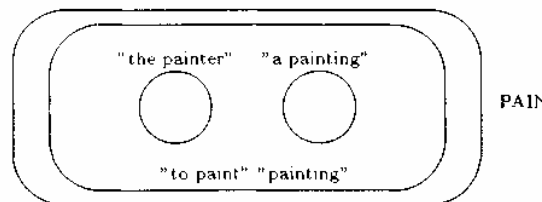

Figure l: The notional domain PAINT, containing the concepts 'to paint', 'painting', 'the painter', 'a painting'.

In view of these considerations we have to reinterpret LTR (5) and STP (7). LTR (5) must be regarded as activating the appropriate notional domain, while STP (7) selects within the notional domain the right concept.

3.2 Argument Structure and Referential Argument

Every lexical item can potentially have thematic arguments to which it assigns thematic roles. Within transfer, arguments maintain their thematic roles, which we assume to be valid across languages. As a consequence, different assignments of syntactic functions in source and target language, so called 'argument switching' (e.g. /l *me plait - I like him)* are not to be handled in transfer as recurrently suggested in literature (e.g. (VAUQUOIS AND BOITET (1985)), (BEAVEN (1992))), but result from different assignments of syntactic functions to identical thematic roles within the monolingual lexicons.

Different parts of speech have different possibilities of realizing syntactic functions for a thematic role. Support verb constructions, one type of syntactic derivation which we shall discuss later, may allow the omission of certain argument slots which are obligatory with the derivationally related verbal constructions. Causative derivations on the other hand represent a derivation where one additional thematic role can be mapped onto a syntactic function. Thus, without additional rules or annotations in the LTR, the presence or absence of realized arguments associated with a thematic role allow or disallow one paraphrase or another, determined only by the possibilities and necessities of the syntactic head to realize a syntactic function for that thematic role (cf. 1b-e).

Every concept of a notional domain can be specified according to a semantic classification which describes (i) restrictions on the denotation of the concept in question (i.e. selectional restrictions on the referential argument) and (ii) restrictions on the denotation of the thematic arguments [5] .
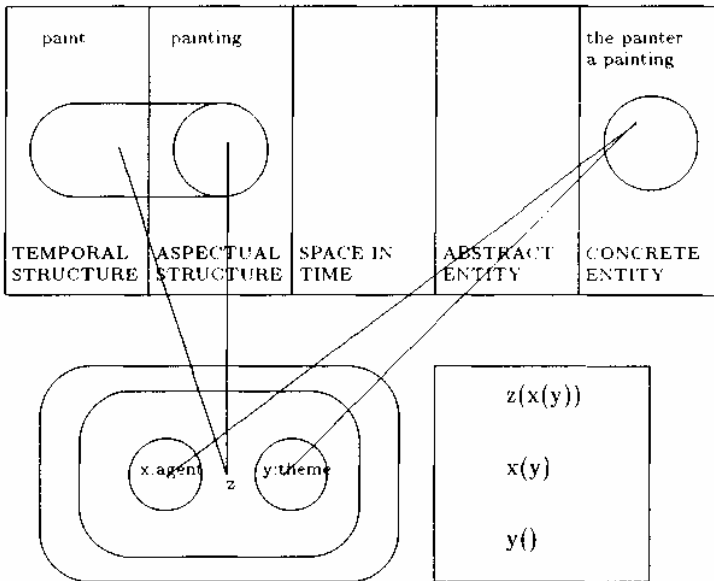
Figure 2: The relation of the notional domain PAINT and its related concepts, the regions their denotations occupy in the semantic space, and the argument structure they realize.

While *to paint* is distinguished from *painting* by means of selectional restrictions on the referential argument, the agent *(the painter)* and the theme (a *painting)* both may occupy the same region in the semantic space (e.g. localized as CONCRETE). In such cases the argument structure helps to distinguish the concepts since the agent may realize the theme argument (e.g. *the painter of a painting)* while *a painting* cannot

---

[5] We shall use the term 'referential' argument to refer to both 'referential' and 'eventual' argument (cf. (GRIMSHAW (1990))), according to our assumption of one common semantic type.

realize the agent argument, due to the fact that only concepts with an aspectual structure (ASPECT) may realize agent arguments. This becomes relevant for the reduction of lexical ambiguity. While action nominalizations are often ambiguous with respect to action versus result reading, the result reading is automatically excluded when the relevant argument slots are filled. For to realize a thematic argument within a phrasal structure, animateness (ANIM) or abstractness (ABSTRACT) of the concept in question suffices, while in sub-phrasal argument structures (e.g. in compounds) these constraints may not apply.[6]

Abstractness and Temporal Extension: As abstract concepts we regard all those entities the denotation of which we cannot touch with our hands and which are not part or a collection of entities which can be touched. *Stones, scree* and *atoms* are concrete, *ideas* and *good* are not. Adjectives derived from concrete nouns (e.g. *stony)* are coded as concrete in their literal meaning since they inherit the referential argument from the noun, i.e. they refer to the same concept (cf. (CARULLA (1994))). Due to the nature of their referential argument and their empty argument structure, such adjectives find no predicative usage. Concrete concepts denote entities **(sem={abstract=nil})** , while abstract concepts denote either entities **(sem={abstract={temp=nil}})**, a space in time **(sem={abstract={temp={struct=nil}}})**[7] or events/states
(sem={abstract={temp={struct={aspect=_}}}}), where the struct feature introduces an aspectual structuring of the event.
Concerning the hierarchy of actualization functions, as it is reflected in our semantic classification, only finite indicative verbs arrive at a full actualization of a concept by their being embedded in the speech act with tense values (cf. (MAINGUENAU (1981))). In other cases, where actualization is indirect, e.g. by reference to events actualized elsewhere in the text, verbs may be translated by event nouns or adjectives or by non-finite or subjunctive verb forms. If nominal concepts or adjectival concepts are to be actualized directly, they have to fall back on support verb constructions and copulative structures (cf. 2a-b).

Conceptual boundedness: As for temporal extension and abstractness, we assume conceptual boundedness to be a category applicable to all parts of speech (cf. Krifka's theory of homomorphism of objects and events (KRIFKA (1991))). We adopt Quine's definition of conceptual boundedness (cf. (QUINE (1960))), according to which concepts refer **cumulatively** if any sum of the concept is the concept itself, e.g. *any sum of parts which are water is water,* (pg.91). Otherwise concepts refer **quantized.** This

---

[6] An overview of the semantic classification currently used in CAT2 is given below. For a detailed description see (STREITER (1994B)).
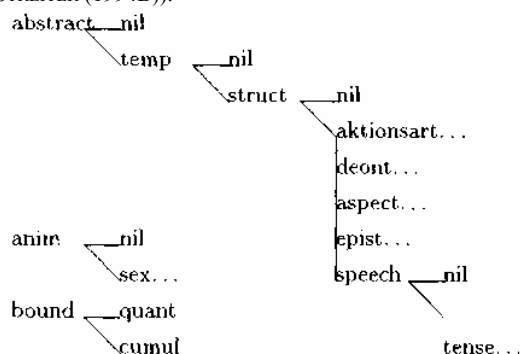
[7] A lexical item having a temporal extension denotes the relation of notional parts of that item to the temporal axis, which makes the whole concept abstract. Accordingly, we represent the temporal description as included in the description of abstraction.

cumul/quant distinction, applied initially only to the semantics of nominals, can be extended to verbal semantics, capturing the well-known distinction between bounded and unbounded processes. Just as 'water' plus 'water' is 'water', 'to run' and 'to run' means 'to run' and not 'to run two times'. Bounded verbal concepts, that is verbal concepts which are inherently lexically bounded (e.g. *to reach, to attain),* or which become bounded by the presence of an object (e.g. *to run a mile, to drink a cup of tea)* refer quantized.

Controlling the conceptual boundedness in transfer is not only essential when parts of speech remain identical, but has repercussions for the transposition of the part of speech. If, for example, the concept expressed by a verb is bounded, due to its inherent lexical semantics or contextual influences, this boundedness is transferred to the target language, where, according to the monolingual specification, it may trigger the generation of a definite or indefinite article on a noun, with the article functioning, among other things, as a marker for the conceptual boundedness. As the following examples show, conceptually unbounded adjectives can be paraphrased only by cumulative PPs (9a-b) while conceptually bounded adjectives need a quantized paraphrase (9c-d).

(9a) Eine anspruchsvolle Waise (a demanding orphan)

   eine Waise mit Ansprüchen / * eine Waise mit **den** Ansprüchen

(9b) Eine bedeutsame Weise (a meaningful tune)

   eine Weise mit Bedeutung / * eine Weise mit **der/einer** Bedeutung

(9c) Auf väterliche Weise (in a paternal way)

   so wie **ein** Vater / * so wie Väter

(9d) Die bucklige Weise (the hunchbacked wise woman)

   eine Weise mit **einem** Buckel. / * eine Weise mit Buckeln

The conceptual boundedness is equally important for the choice of the derivation type. In example (l0a) the conceptually unbounded verb is translated into the gerund nominalization and not into the zero-derivation, clue to the boundedness implied by this derivation type, while in the case of the conceptual bounded verb in (l0b), the zero-derivation is possible as well.

(l0a) Er wandert gerne.

   He likes walking. / *He likes the walk.

(l0b) Er wandert gerne nach Bonn.

   He likes walking to Bonn. / He likes the walk to Bonn.

4   Morphological Patterns of Derivation

In order to accommodate the facts described, the CAT2 dictionary contains a comprehensive description of various uses of the basic lexeme understood as the notional domain, including (i) a description of the referential argument (**sem={. . .}**) of every concept realized by the respective derivation and (ii) the underlying argument structure from which the concepts may realize their respective arguments.

4.1   Action Nominalizations

Action Nominalizations refer as nouns to their base verbal process, or to the resulting state of that process. The base together with its derived nominals is represented in the CAT2 lexicon as shown in Figure 3, where every head disjunct refers to one concept within the notional domain.
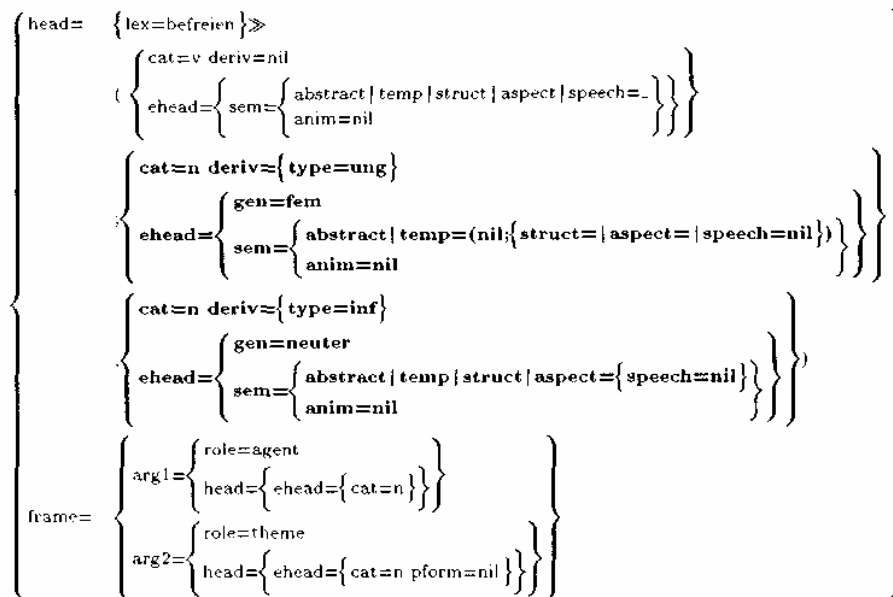
```
┌ head=  {lex=befreien}»                                                              ┐
│         ┌ cat=v deriv=nil                                              ┐            │
│       ( │          ┌      ┌ abstract | temp | struct | aspect | speech=_ ┐ ┐ │            │
│         │ ehead=  │ sem= │                                        │ │ │            │
│         └          └      └ anim=nil                              ┘ ┘ ┘            │
│         ┌ cat=n deriv={type=ung}                                           ┐        │
│         │          ┌ gen=fem                                              ┐ │        │
│       , │ ehead=  │        ┌ abstract | temp=(nil;{struct=|aspect=|speech=nil}) ┐ │ │        │
│         │          │ sem= │                                              │ │ │        │
│         └          └        └ anim=nil                                  ┘ ┘ ┘        │
│{                                                                                    }│
│         ┌ cat=n deriv={type=inf}                                     ┐              │
│         │          ┌ gen=neuter                                     ┐ │              │
│       , │ ehead=  │        ┌ abstract | temp | struct | aspect={speech=nil} ┐ │ │ )            │
│         │          │ sem= │                                        │ │ │              │
│         └          └        └ anim=nil                              ┘ ┘ ┘              │
│         ┌          ┌ role=agent                              ┐                       │
│         │ arg1=   │                                          │                       │
│         │          └ head={ehead={cat=n}}                    ┘                       │
│ frame= {                                                                            }│
│         │          ┌ role=theme                                     ┐                │
│         │ arg2=   │                                                 │                │
│         └          └ head={ehead={cat=n pform=nil}}                 ┘                │
└                                                                                      ┘
```

Figure 3: The lexical entry for the concepts *befreien, die Befreiung* and *das Befreien.*[8]

This lexical entry exemplifies the principle of an underlying argument structure: Common to both the verb and its nominalizations is the **frame={. . .}** feature which list possible arguments. Differences in the surface realization of arguments with respect to case marking or prepositional form depend on the syntactic principles of the respective languages (e.g. structural case assignment, *of*-insertion, adjacency restrictions etc. . .). Thus case and preposition, if they change in a regular manner, are not entered in the lexicon but are determined after the part of speech has been selected.

4.2 Potential Passive Derivations

To the Action Nominalization can easily be added the Potential Passive derivation. This derivation, which yields in most cases an adjective with a modal value 'ability', coindexes in attributive structures the modified noun with its second argument, while in predicative structures the second argument becomes the subject of the sentence by an 'argument transfer' from the adjective to the copula (see the section on copulative structures). The impossibility of realizing the first argument of the notional domain together with the Potential Passive is controlled in predicative structures by the copula verb which assigns no surface realization to it. In attributive structures the first argument of an adjective is never accessible.

---

[8] The feature **head** refers to Head Features, which are standard assumptions in theories as GPSG and HPSG. **ehead** refers to Extended Head Features as they are defined in (GRIMSHAW (1991)) and (STREITER (1994A)): they describe the subset of Head Features which are shared between the lexical head (e.g. a noun) and its functional extensions (determiner and preposition), **frame** describes the argument structure with **arg1 ,arg2** etc. . .representing the different syntactic functions. Following Prolog conventions, a disjunction of values is represented '(...;...)'. The » operator is used to link different values of an attribute, in this case a positive and a disjunctive value (cf. (SHARP AND STREITER (1995))).
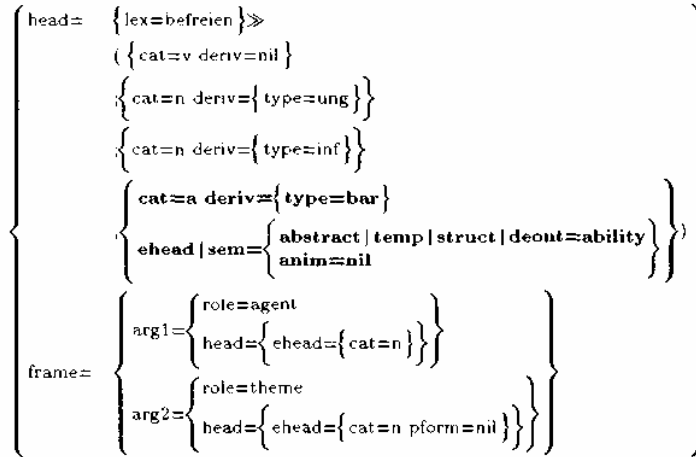
$$
\begin{bmatrix}
\text{head} = & \{\text{lex=befreien}\} \gg \\
& (\ \{\text{cat=v deriv=nil}\} \\
& \{\text{cat=n deriv=}\{\text{type=ung}\}\} \\
& \{\text{cat=n deriv=}\{\text{type=inf}\}\} \\
& \begin{cases}
\text{cat=a deriv=}\{\text{type=bar}\} \\
\text{ehead|sem=}\begin{cases}\text{abstract|temp|struct|deont=ability} \\ \text{anim=nil}\end{cases}
\end{cases})\ ) \\
\text{frame=} & \begin{cases}
\text{arg1=}\begin{bmatrix}\text{role=agent} \\ \text{head=}\{\text{ehead=}\{\text{cat=n}\}\}\end{bmatrix} \\
\text{arg2=}\begin{bmatrix}\text{role=theme} \\ \text{head=}\{\text{ehead=}\{\text{cat=n pform=nil}\}\}\end{bmatrix}
\end{cases}
\end{bmatrix}
$$

Figure 4:The lexical entry for *befreien,* including the Potential Passive Derivation *befreibar.*

Lexicalized derivations, that is derivations to which either no or a modal value other than 'ability' is associated, are accounted for in the lexicon. In these cases the derivation type **deriv={type=bar}** is associated with its 'lexicalized' semantic description as for example **deont=obligation** in *payable.*

4.3   Agent Nominalizations

The Agent Nominalization realizes the agent concept of the notional domain receiving the relevant semantic description from the agent argument of the underlying argument structure (**sem=SEM**)[9]. By coindexing the referential argument of the noun and the agent slot in the lexicon (**index=INDEX**) the agent noun is prevented from externally expressing the first argument of the notional domain due to the unique referential value of index.

$$
\begin{bmatrix}
\text{head} = & \{\text{lex=befreien}\} \gg \\
& (\ \{\text{cat=v deriv=nil}\} \\
& \{\text{cat=n deriv=}\{\text{type=ung}\}\} \\
& \{\text{cat=n deriv=}\{\text{type=inf}\}\} \\
& \{\text{cat=a deriv=}\{\text{type=bar}\}\} \\
& \begin{cases}
\text{cat=n deriv=}\{\text{type=agent}\} \\
\text{ehead=}\{\text{sem=SEM index=INDEX}\}
\end{cases})\ \\
\text{frame=} & \begin{cases}
\text{arg1=}\begin{bmatrix}\text{role=agent} \\ \text{head=}\{\text{ehead=}\{\text{cat=n sem=SEM index=INDEX}\}\}\end{bmatrix} \\
\text{arg2=}\begin{bmatrix}\text{role=theme} \\ \text{head=}\{\text{ehead=}\{\text{cat=n pform=nil}\}\}\end{bmatrix}
\end{cases}
\end{bmatrix}
$$
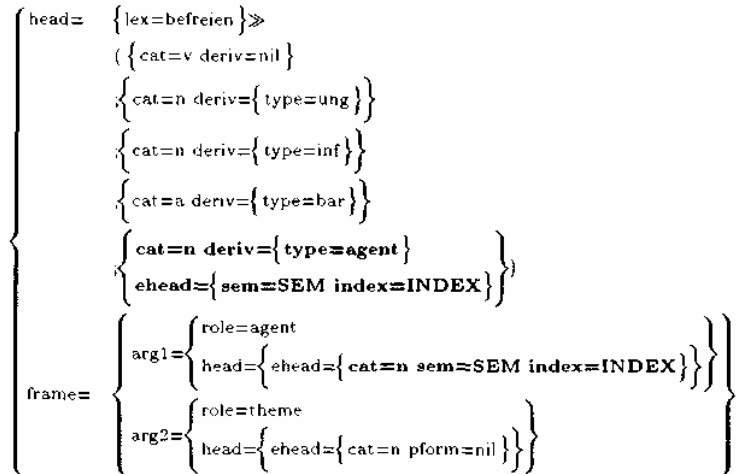
Figure 5: The lexical entry for *befreien,* including the agent derivation *Befreier.*

In our implementation, no distinction is made between the agentive vs. the instrumental reading of the agent nominalization if the same morphological pattern of derivation is involved. Limitations of interpretation come from the anim-feature shared with the first argument. In phrasal structures the thematic argument is not available to the instrumental reading due to its referential argument.

---

[9] Following Prolog convention, logical variables beginning with uppercase letters are used to bind two values.

5 Syntactic Patterns of Derivation

In syntactic patterns of derivation the same principles as in morphological derivation are preserved: (i) the argument structure is that of the underlying notional domain, ii) the syntactic realization of the arguments depends on the part of speech of the syntactic head and (iii) the switch of the part of speech allows semantic, syntactic and stylistic realizations which are not possible for the base form.

5.1 Support Verb Constructions

Support verb constructions (SVCs) are the first type of syntactic derivation we want to mention in this context. They consist of two parts, (i) a verb (SV) or its deverbal derivation which has lost its capacity to assign thematic roles and (ii) a predicative noun, which is the bearer of the propositional content and the argument structure. The main function of the verb in a SVC is restricted to the actualization of the propositional content by tense, aspect, mood values. Since the choice of the support verb is determined by lexical idiosyncratic specifications of the predicative noun, SVCs cannot be translated compositionally (examples from (MESLI (1991))). Accordingly, as for bound morphemes, the support verb is not presented at the Interface Structure (IS) and the place of the predicate is occupied by the predicative noun.

  (11a) prendre une décision - einen Beschluß fassen
  (11b) prendre l'initiative - die Initiative ergreifen
  (11e) prendre peur - Angst bekommen

The lexical entry of a support verb and its morphological derivatives is exemplified in Figure 6. The entry already contains all variable linking necessary to copy the semantic arguments of the predicative noun (**role=pred**) onto the 'syntactic' argument slots of the SV (cf. (GRIMSHAW AND MESTER (1988))).
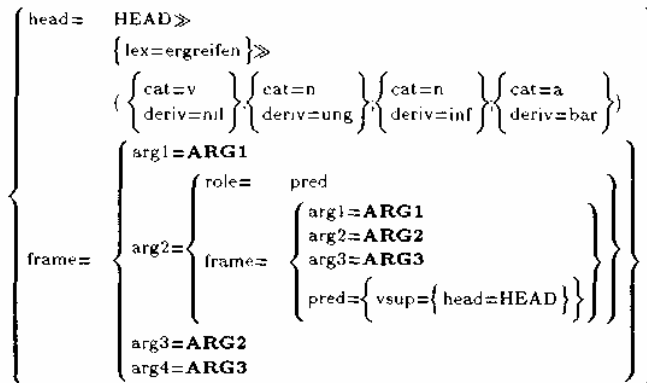
Figure 6: An example of a support verb which realizes the argument transfer by means of variable binding.

In our example, the support verb subcategorizes the predicative noun as its second argument position, by means of which the noun is assigned accusative case in active and nominative case in passive sentences. In other SVCs, other argument slots are opened for the predicative noun.

The entry of the predicative noun (**lex=macht**), i.e. the second component of a support verb construction, predicts the SV necessary for its predicative use in a sentence (**vsup={. . .}**), as well as its own form when appearing together with the predicative verb (**xpred={. . .}**). This information (e.g. **pform**) cannot be coded directly as an extended head feature of the noun since the predicative properties can be delegated to a dependent relative pronoun, in which case these constraints apply to the relative pronoun and not to the predicative noun itself (e.g. *He was proud of the decision which his daughter had made).*

$$\left\{\begin{array}{l} \text{head}=\left\{\begin{array}{l}\text{lex=macht, cat=n}\\ \text{ehead}=\{\text{type=def}\}\end{array}\right\}\\ \text{frame}=\left[\begin{array}{l}\text{arg1}=\left\{\begin{array}{l}\text{role=agent}\\ \text{head}=\{\text{ehead}=\{\text{cat=n}\}\}\end{array}\right\}\\ \text{arg2=nil}\\ \text{arg3=nil}\\ \text{pred}=\left\{\begin{array}{l}\text{vsup}=\left\{\text{head}=\left\{\begin{array}{l}\text{lex=ergreifen}\\ \text{ehead|sem|abstract|temp|struct|akt=incho}\end{array}\right\}\right\}\\ \text{xpred}=\{\text{head}=\{\text{ehead}=\{\text{pform=nil}\}\}\}\end{array}\right]\end{array}\right]\end{array}\right\}$$
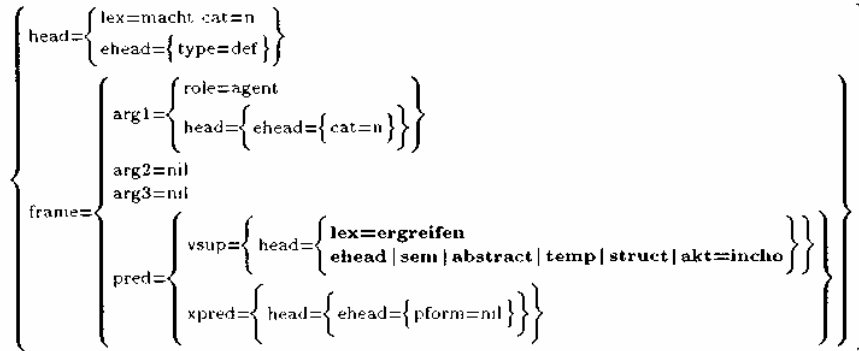
Figure 7: An example of a predicative noun which selects *ergreifen* for an inchoative SVC.

The result of the subcategorization of the predicative noun by the support verb is represented in the following structure, where the argument slot for the subject is still accessible.

$$\left\{\begin{array}{l}\text{head}= \quad \text{HEAD}\gg\\ \left\{\begin{array}{l}\text{lex=ergreifen}\\ \text{cat=v}\\ \text{deriv=nil}\\ \text{ehead|sem|abstract|temp|struct|akt=incho}\end{array}\right\}\\ \text{frame}= \left\{\begin{array}{l}\text{arg1}=\left\{\begin{array}{l}\text{role=agent}\\ \text{head=|ehead=|cat=n}\end{array}\right\}\gg\text{ARG1}\\ \text{arg2}=\left[\begin{array}{l}\text{role=pred}\\ \text{head}=\{\text{lex=macht}\}\\ \text{frame}=\left\{\begin{array}{l}\text{arg1}=\left\{\begin{array}{l}\text{role=agent}\\ \text{head|ehead|cat=n}\end{array}\right\}\gg\text{ARG1}\\ \text{arg2=nil}\\ \text{arg3=nil}\\ \text{pred|vsup|head=HEAD}\\ \text{xpred|head|ehead}=\{\text{type=def pform=nil cat=n}\}\end{array}\right\}\end{array}\right]\\ \text{arg3=nil}\\ \text{arg4=nil}\end{array}\right\}\end{array}\right\}$$
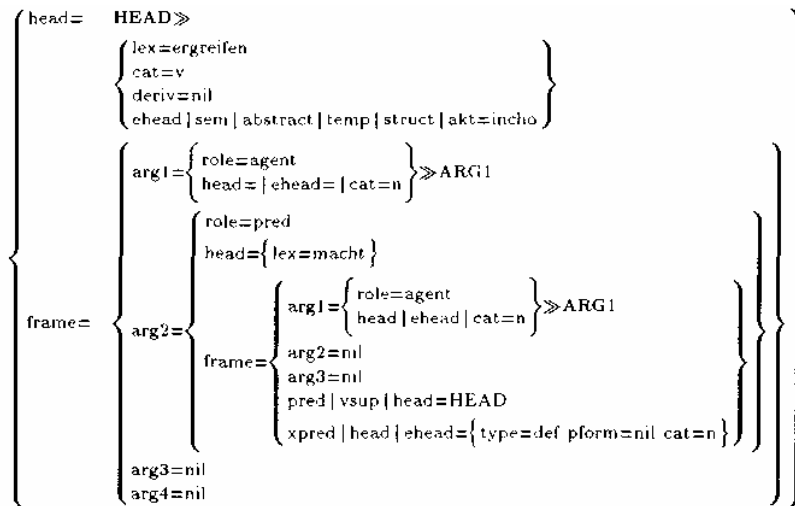
Figure 8: The SV *ergreifen* after having subcategorized the predicative noun *Macht*.

To link syntactic to morphological types of derivation, it should be noted that the predicative noun of the SVC is often itself derived from a verb which can be interpreted as a paraphrase of the whole SVC. In these cases we are confronted with double derivation, where first a noun is derived from a verb for reasons of greater variation regarding arguments and modifiers. Secondly a syntactic derivation of this noun to a verbal construction takes place, in order to achieve the appropriate actualizations.

5.2 Copulative Structures

Copulative structures placing an adjective in a predicative position are treated in our implementation along the same lines as support verb constructions, implying the adjective to be the semantic predicate of the sentence. The lexical entries of the copula and the adjective are equivalent to those of the support verb and the predicative noun respectively, describing an argument shift from the second argument position of the adjective to the first argument position of the verb. Supplementary arguments can be realized following the description given by the adjectival frame. The referential argument of the adjective and the copula must unify, so that only adjectives with an aspectual structure can form copulative structures. When the copula verb is dropped at the interface level (IS), all the semantic information of the verb is stored on the adjective.

5.3   Generic Support

Generic Support as described in (MEL'ČUK (1974)) is a third type of syntactic derivation. Adjectives may be defective with respect to the possibility of deriving nouns from them by morphological means, so that reference has to be achieved by other means. In these cases, as for stylistic purposes, adjectives may build a nominal structure with the help of their hyperonyms (e.g. red $\Rightarrow$ colour). In English for example, support nouns are required for singular nationality adjectives which end in dental fricatives. A 'support noun' such as *man* and *woman* must be inserted in order to establish reference. Further examples come from Russian and Serbo-Croatian.

(12a) ein Japaner/eine Japanerin - a Japanese **man/woman**
(12a) ein Weiser/eine Weise - a wise **man/woman**
(12c) in Russian - na russkom **jazyke**
(12d) in Serbian - na srpkom **jeziku**

In any of these cases a noun is translated by a adjective plus a supporting noun. The LTR maps *Russian* onto the adjective *russkij* and not onto the noun *jazyk* (language). The support noun *jazyk* has to be regenerated according to monolingual lexical specifications in the same way as the support verb and copulative verbs. If more than one support is possible for a given modifier, the selection of the right support depends on the referential argument of the modifier. Only a support that is compatible with the referential argument of the modifier can fulfil this function. Thus restrictions on the referential argument of the adjective (**{animate=hum**, **sex=male}** or **{animate=hum, sex=female}**) trigger the corresponding support noun, *man* and *woman* respectively.

$$\left\{ \begin{array}{l} head= \quad \left\{ \begin{array}{l} lex=japan \\ ehead=\left\{ cat=a\ sem=SEM \right\} \end{array} \right\} \\ frame= \quad \left\{ nsup=\left\{ head=\left\{ \begin{array}{l} lex=(man;woman) \\ ehead=\left\{ sem=SEM \right\} \end{array} \right\} \right\} \right\} \end{array} \right\}$$

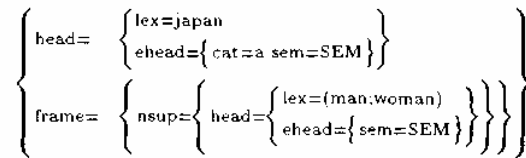Figure 9: Generic support for "Japanese" $\Rightarrow$ "woman" and "Japanese" $\Rightarrow$ "man".

6   An Example of Translation

In order to show how translation works within CAT2, let us consider the German sentence *Der Mann ist erpreßbar*. Its syntactic structure and the resulting interface structure are shown in Figure 10. Functional categories and the copula are no longer present at IS.
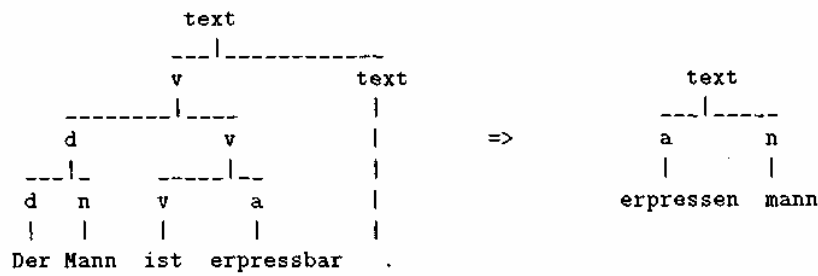
268

```
                        text
                   ___|_____
                    v            text
              _____|____         |
             d         v           |
          ___|_      ____|__       |
         d    n     v     a        |    =>
         |    |     |     |        |
       Der  Mann   ist  erpressbar  .
```
```
        text
     ___|____
      a     n
      |     |
  erpressen  mann
```

Figure 10: The syntactic structure and the interface structure of *Der Mann ist erpreßbar.*

The feature structure of *erpreßbar is* reproduced in Figure 11, where the predicative adjective bears all semantic values including temporal and aspectual information coming from the copula.
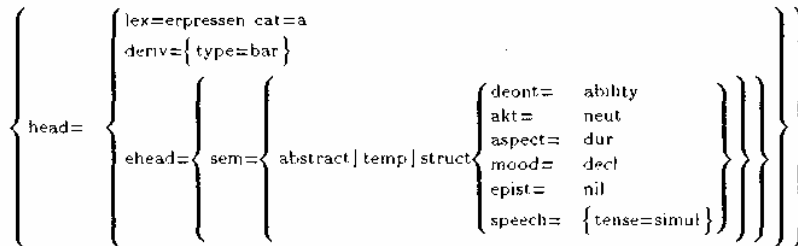
$$
\left\{ head= \left\{ \begin{array}{l} lex{=}erpressen\ cat{=}a \\ deriv{=}\{ type{=}bar\} \\ ehead{=}\left\{ sem{=}\left\{ abstract\,|\,temp\,|\,struct \left\{ \begin{array}{ll} deont{=} & ability \\ akt{=} & neut \\ aspect{=} & dur \\ mood{=} & decl \\ epist{=} & nil \\ speech{=} & \{tense{=}simul\} \end{array} \right\} \right\} \right\} \end{array} \right\} \right\}
$$

Figure 11: The semantic value of *erpreßbar* in *Der Mann ist erpreßbar.*

By virtue of the lexical LTR **{lex=erpressen} <=> {lex=blackmail}** this structure is mapped onto the domain 'blackmail' (cf. ENTRY). Since there is no morphological derivation *blackmailable* and the nominalization has conflicting semantic values the application of the STP selects the verbal concept as the translation of the German adjective.
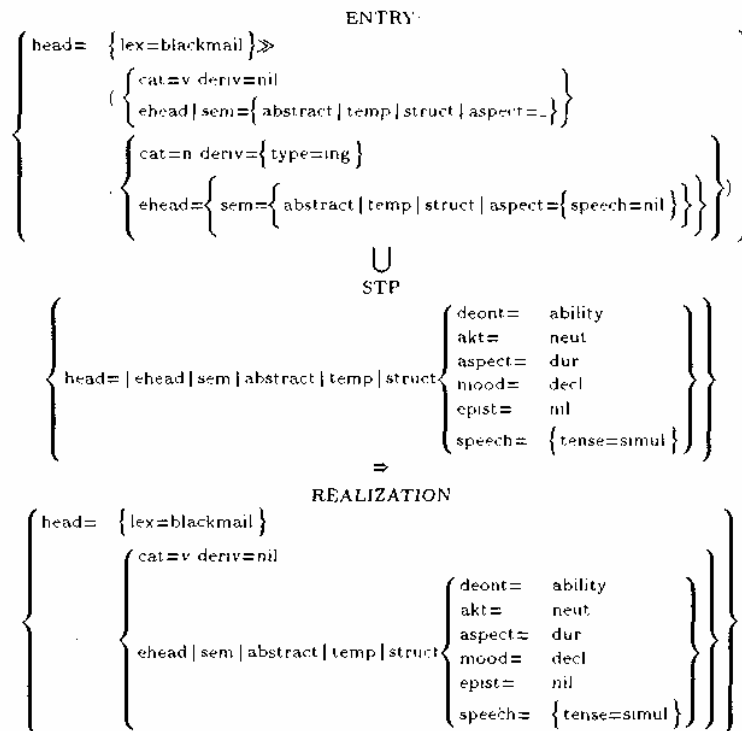
ENTRY

$$
\left\{ \begin{array}{l} head= \{ lex{=}blackmail\} \gg \\ \left(\begin{array}{l} \left\{ \begin{array}{l} cat{=}v\ deriv{=}nil \\ ehead\,|\,sem{=}\{ abstract\,|\,temp\,|\,struct\,|\,aspect{=}\_\} \end{array} \right\} \\ \left\{ \begin{array}{l} cat{=}n\ deriv{=}\{type{=}ing\} \\ ehead{=}\left\{ sem{=}\{ abstract\,|\,temp\,|\,struct\,|\,aspect{=}\{speech{=}nil\}\} \right\} \end{array} \right\} \end{array} \right) \end{array} \right\}
$$

∪

STP

$$
\left\{ head{=}\,|\,ehead\,|\,sem\,|\,abstract\,|\,temp\,|\,struct \left\{ \begin{array}{ll} deont{=} & ability \\ akt{=} & neut \\ aspect{=} & dur \\ mood{=} & decl \\ epist{=} & nil \\ speech{=} & \{tense{=}simul\} \end{array} \right\} \right\}
$$

⇒

REALIZATION

$$
\left\{ \begin{array}{l} head= \{ lex{=}blackmail\} \\ \left\{ \begin{array}{l} cat{=}v\ deriv{=}nil \\ ehead\,|\,sem\,|\,abstract\,|\,temp\,|\,struct \left\{ \begin{array}{ll} deont{=} & ability \\ akt{=} & neut \\ aspect{=} & dur \\ mood{=} & decl \\ epist{=} & nil \\ speech{=} & \{tense{=}simul\} \end{array} \right\} \end{array} \right\} \end{array} \right\}
$$

Figure12: The selection of the verbal concept within the domain *blackmail.*

From this, the translation *The man can be blackmailed* is generated. The modal value coming from the German adjective derivation is expressed with the help of the auxiliary can, which is introduced between IS and the syntactic structure. Passive is generated in order to allow the agent to remain not explicitly expressed.
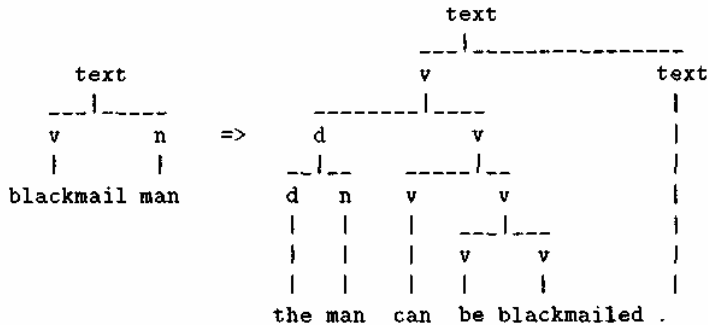
```
                                        text
                                   ___|_____
                                     v                 text
           text                                          |
        ___|_____                 _____|____          |
        v        n     =>         d           v          |
        |        |              __|__      _____|__       |
     blackmail  man             d   n     v      v        |
                                |   |     |    ___|___     |
                                |   |     |   v     v      |
                                |   |     |   |     |      |
                               the man   can  be blackmailed  .
```

Figure 13: The interface structure and the syntactic structure of the target sentence *The man can be blackmailed.*

7 Conclusion

The treatment of derivation suggested here has been implemented in the current CAT2 prototype for English, French and German, with about 10.000 basic lexemes per language. Other languages as Chinese, Spanish and Russian have been developed to a lesser extent, but follow the principles outlined here. By means of derivation these 10.000 lexemes correspond to ca. 20.000 concepts, for which about 11.000 transfer rules are needed for each language pair, due to the existence of many-to-many translation equivalences. Other transfer-based systems thus would need between 40.000 and 80.000 transfer rules in order to have a similar coverage in transfer if they were to allow the part of speech to change.
The advantages of the notion-driven approach to transfer is thus the reduction in the size of bilingual dictionaries, resulting at the same time in a greater variability in transfer and a nearly complete coverage of possible translational equivalents. The monolingual lexicons represent in a coherent form the different patterns of derivation, allowing a systematic development of the lexicons. Since each pattern of derivation is associated with its semantic value, the target language can choose the appropriate morphological and syntactic structure for the expression of the transferred semantic representation. Such an autonomous mechanism of paraphrasing seems to us necessary in order to cope with at least the basic problems of MT.

References

JURIJ D. APRESJAN, IGOR M. BOGUSLAVSKIJ, LEONID L. IOMDIN, ALEXANDRE V. LAZURSKIJ, VLADIMIR Z. SANNIKOV, and LEONID L. TSINMAN. 1989. *Lingvističeskoe obespečenie sistemy ETAP-2.* Izdatel'stvo "Nauka", Moskva.

DOUG ARNOLD and LOUISA SADLER. 1987. Non-Compositionality and translation. Working paper in language processing no.l, University of Essex, Department of Language and Linguistics, Essex.

DOUG ARNOLD and LOUISA SADLER. 1990. The theoretical basis of MiMo. *Machine Translation,* 5(3): 195-222.

JOHN L. BEAVEN. 1992. Shake-and-bake machine translation. In *Actes de COLING-92,* pages 603-609, Nantes, August.

MARTA CARULLA. 1994. Relational adjectives in the translation from Germanic nominal compounds into Romance languages. In Pierrette Bouillon and Dominique Estival, editors, *Proceedings of the Workshop on compound Nouns: Multilingual Aspects of Nominal Composition,* pages 103-107, Geneva, 2-3 December. ISSCO.

ANTOINE CULIOLI. 1990. *Pour une linguistique de l'énonciation, tom 1.* Collection L'Homme dans la Langue. OPHYS, Paris.

BONNIE J. DORR. 1994. Machine translation divergences. *Computational Linguistics,* 20(4):597-633.

BARBARA GAWRONSKA, ANDERS NORDNER, CHRISTER JOHANSSON, and CAROLINE WILLNER. 1994. Interpreting compounds for machine translation. In *COLING 94, The 15th International Conference on Computational Linguistics. Proceedings,* Kyoto, Japan, August 5-9.

ALGIRDAS JULIEN GREIMAS and COURTÉS JOSEPH. 1989. *Sémiotique, Dictionnaire raisonné de la théorie du langage.* Langue Linguistique Communication. Hachette, Paris.

JANE GRIMSHAW and ARMIN MESTER. 1988. Light verbs and θ-marking. *Linguistic Inquiry,* 19(2):205-232.

JANE GRIMSHAW. 1990. *Argument Structure.* Linguistic Inquiry Monographs. The MIT Press, Cambridge, Massachusetts.

JANE GRIMSHAW. 1991. Extended projection. Brandeis University, Waltham MA 02254, ms, July.

MANFRED KRIFKA. 1991. Thematic relations as links between nominal reference and temporal constitution. In Ivan Sag and Anna Sabolesi, editors. *Lexical Matters.* Chicago University Press, Chicago.

DOMINIQUE MAINGUENAU. 1981. *Approche de l'Enonciation en linguistique Française, Embrayeurs, "Temps", Discours rapporté.* Langue Linguistique Communication. Hachette, Paris.

IGOR ALEKSANDROVIC MEL'ČUK and NIKOLAJ V. PERTSOV. 1987. *Surface Syntax of English, a formal model within the meaning-text framework.* Linguistic & Literary Studies in Eastern Europe. John Benjamins Publishing Company, Amsterdam, Philadelphia.

IGOR ALEKSANDROVIC MEL'ČUK. 1974. *Opyt teorii lingvističeskix modelej Smysl ⇔Tekst. Semantika, sintaksis.* Izdatel'stvo "Nauka", Moskva.

NADIA MESLI. 1991. Analyse et traduction automatique de constructions à verbe support dans le formalisme CAT2. EUROTRA--D working paper 19b, IAI, Martin-Luther-Straße 14, 66111 Saarbrücken, BRD.

NAOYUKI NOMURA, DOUGLAS A. JONES, and ROBERT C. BERWICK. 1994. An architecture for a universal lexicon. In *COLING 94, The 15th International Conference on Computational Linguistics. Proceedings,* Kyoto, Japan, August 5-9.

JOSIANE PODEUR. 1993. *La Pratica della Traduzione. Dal francese in italiano e dell'italiano in francese.* Liguori editore, Napoli.

271

CARL POLLARD and IVAN SAG. 1994. *Head-Driven Phrase Structure Grammar.* Studies in Contemporary Linguistics. The University of Chicago Press, Chicago & London.

WILLARD VAN ORMAN QUINE. 1960. *Word and Object.* The M.I.T. Press, Cambridge, Massachusetts, 14 edition.

PAUL SCHMIDT. 1988. Transfer strategies in EUROTRA. In Erich Steiner, Paul Schmidt, and Cornelia Zelinsky-Wibbelt, editors, *From Syntax to Semantics, Insight from Machine Translation.* Pinter Publishers Ltd., London.

RANDALL SHARP and OLIVER STREITER. 1995. Applications in multilingual machine translation. In *Proceedings of The Third International Conference and Exhibition on Practical Applications of Prolog, Paris, 4th-7th April.*

HAROLD SOMERS, HIDEKI HIRAKAWA, SEIJI MIIKE, and SHINYA AMANO. 1988. The treatment of complex English nominalizations in Machine Translation. *Computers and Translation,* 3(1):3-21, March.

OLIVER STREITER, RANDALL SHARP, JOHANN HALLER, CATHERINE PEASE, and ANTJE SCHMIDT-WIGGER. 1994. Aspects of a unification based multilingual system for computer-aided translation. In *Proceedings of Avignon '94, 14th International Conference.*

OLIVER STREITER. 1994a. Komplexe Disjunktion und Erweiterter Kopf: Ein Kontrollrnechanismus für die MÜ. In *Konvens '94 Tagungsband, 2. Konferenz "Verarbeitung natürlicher Sprache" Wien, 28-30 September 1994.*

OLIVER STREITER. 1994b. *Linguistic Reference Manual of the CAT2 Machine Translation System.* Martin-Luther-Straße 14, 66111 Saarbrücken, BRD, April.

ULRIKE TALLOWITZ. 1994. Die Behandlung von Nominalisierungen in der automatischen Übersetzung. CELE-UNAM Mexico-City, ms, August.

BERNARD VAUQUOIS and CHRISTIAN BOITET. 1985. Automated translation at Grenoble University. *Computational Linguistics,* 11(l):28-39.

P. WHITELOCK. 1992. Shake-and-bake translation. In *Actes de COLING-92,* pages 784-789, Nantes, August.

CORNELIA ZELINSKY-WIBBELT. 1988. From cognitive grammars to the generation of semantic interpretation in machine translation. In Erich Steiner, Paul Schmidt, and Cornelia Zelinsky-Wibbelt, editors. *From Syntax to Semantics. Insight from Machine Translation.* Pinter Publishers Ltd., 25 Floral Street, London.