INTERPRETING COMPOUNDS FOR MACHINE TRANSLATION

BARBARA GAWROŃSKA
ANDERS NORDNER
CHRISTER JOHANSSON
CAROLINE WILLNERS

Dept. of Linguistics, University of Lund,
Helgonabacken 12, S-223 62 LUND, Sweden

SUMMARY:

The paper presents a procedure for interpretation of English compounds and for automatic translation of such compounds into Slavic languages and French. In the target languages, a compound nominal is as a rule to be rendered by an NP with an adjective or genitive attribute, or with an attributive participle construction. The model is based on Bierwisch's theory of word formation, which in turn is inspired by categorial grammar. The procedure is applied to a specific domain (asthma research).

0. INTRODUCTION

The need of a component interpreting complex lexical items in an MT system translating from Germanic languages into e.g. French or Slavic languages is obvious. Many rules (or patterns) of word formation are highly productive, which makes it impossible to store all complex lexical entries in a static lexicon.

An effective MT system must also be able to match the interpretation of a complex entry with the correct morphosyntactic pattern in the target language. For example, a program translating from German into Polish must distinguish the relations between the parts of a compound like *Universitätslehrer* (university teacher) from the relations holding between *Musik* and *Lehrer* in *Musiklehrer* (teacher of music). The first mentioned compound is to be translated as a noun followed by an adjective (*nauczyciel uniwersytecki*–'teacher university+adjective ending'), the later one as a noun and a genitive attribute (*nauczyciel muzyki*–'teacher music+gen'). Similar problems occur when translating into French or Czech: cf. *Musikabend*–Fr. *soirée musicale* (n a), Cz. *hudební večer* (a n), *Musiklehrer*–Fr. *professeur de musique* (n prep n), Cz. *učitel hudby* (n n+gen).

The models for compound interpretation and generation proposed by general linguists (cf. Lees 1960, Selkirk 1982, Fanselow 1988, Bierwisch 1989) require as a rule several modifications in order to be applicable in an MT system. Since, in our opinion, a model aimed to serve as an efficient tool for NLP and MT must be linguistically valid, we will discuss a number of theoretical questions and relate our model to general linguistics before presenting our experimental procedure for domain-restricted compound translation.

1. THE STATUS OF WORD FORMATION RULES

1.1. 'Where's morphology?'

The above question, put by Stephen Anderson 1982, is still waiting for a definitive answer. Word formation rules have been claimed to obey syntactic principles and hence being a part of UG (Lees 1960, Pesetsky 1985), to form a grammatical level on their own (Di Sciullo & Williams 1987), to be explainable in semantic terms solely (Fanselow 1988) or to belong to the lexicon (Chomsky 1970, Jackendoff 1975, Bierwisch 1989).

We will propose a quite simple answer to Anderson's question: morphology shall be seen as a component of the grammar, the notion 'grammar' to be understood as an integrated model where no borders are drawn between syntax, morphology, and semantics.

1.2. Towards an integration of syntax, semantics and morphology

Fanselow (1985, 1988) argues, on the basis of psycholinguistic evidence, for treating word formation rules not as generative processes, but as a 'primitive' process of concatenating morphematic items, a very easily learnable procedure. His argumentation is restricted to morphology in the traditional sense of the term. We would like to go even further and claim that the grammar as a whole can be regarded as a set of patterns for concatenation or cooccurrence of lexical items, each concatenation pattern associated with principles of semantic interpretation. This approach is to some extent inspired by (but far from identical to) categorial grammar and Bierwisch's lexicon theory (Bierwisch 1989). At the same time, it is in its very essence not incompatible with Constraint Grammar (Karlsson 1990, Koskenniemi 1990).

1.3. Compounds as collocations

English compounds provide an argument in favour of our approach to grammar. It seems impossible to draw a clear-cut borderline between strings traditionally labelled as

compounds and those classified as noun phrases. Cf. the following examples, taken from a corpus of medical abstracts:

ragweed allergic rhinitis
house-dust-allergic asthma
house dust asthma
patient daily symptom diary cards
fluticasone propionate aqueous nasal spray

In most grammatical descriptions, strings consisting of nouns (like *house dust asthma*) are treated as compound nouns, whereas a complex including an adjective followed by a noun is normally labelled as an NP. The above examples show, however, that such a distinction is not unproblematic. Phrases like *house-dust-allergic asthma* and *fluticasone propionate aqueous nasal spray* may be analysed either as NPs containing a compound adjective and a head noun, or as compounds including optional adjective constituents (*house dust asthma* and *fluticasone spray* are perfectly well-formed). Furthermore, parts of an English compound may provide referents for elliptic constructions, as in the following examples:

The variations in provocation concentrations ... were small during both ***placebo*** *and active drug* ***treatment***

the difference between a single ***allergen provocation*** *and continuous* ***exposure***...

Thus, a noun included in a compound can still have a referent on its own, an ability normally associated with nominal phrases. Such facts indicate that there is no absolute distinction to be drawn between compound nominals and complex nominal phrases in English. It seems more appropriate to talk about more or less lexicalized collocations. However, in the following the traditional term 'compound' will be used.

2. Automatic interpretation and translation of compounds

2.1. The theoretical foundation

Bierwisch (1989; cf also Olsen 1991) regards the process of compounding as a functional application, where one of the thematic roles of the head noun becomes 'absorbed'. For example, a noun like *payer* is supposed to have the following interpretation:

$$\lambda y \lambda x[zINST[xPAYy]],$$

where y is the external theta-role, x the internal one, and z represents the 'referential role'. In a compound like *bill payer*, the internal role of *pay* becomes instantiated:

$$\lambda x[zINST[xPAY\ BILL]].$$

Our analysis of compounds is not incompatible with Bierwisch's approach. However, for the purpose of MT, a classification of valency in terms of three kinds of theta roles only (external, internal and referential) seems insufficient. A procedure for compound interpretation must also take into account optional thematic roles, e.g. location (*university teacher*). It must in addition be able to deal with compounds that do not include deverbal components. Hence, we decided to modify the theory proposed by Bierwisch at two main points:

a. the valency of a verbal stem is to be represented not in terms of external and internal theta roles, but in terms of the components of the event or situation the verb may refer to

b. the interpretation of compounds that do not contain deverbal elements is based on morpho-semantic patterns specifying the default readings of combinations that include members of different semantic categories.

2.2 An experimental procedure for understanding derived nouns and compounds

In an experimental program, implemented in LPA MacProlog, we structured a very restricted lexicon of Swedish stems and affixes (basal lexical entries, BLA) according to the approach outlined above. Each verbal stem was provided with a list of elements of its typical event referent, e.g.:

```
lex([lär],m(teach,stem),v,vt,[agent,
    sem_object,domain,place,time,result],[]).
```

Affixes were specified with respect to the following features:

◊ the category or categories of stems the affix may be combined with
◊ the resulting category, including the morpho-syntactic specification
◊ the default semantic interpretation of the affix.

For example, the Swedish agentive suffix *-are* was represented as:

```
slex([are],suff(n,agr(sg,re,indef)),v,agent,[]).
```

Underived nouns got a quite simplified semantic specifcation formulated in traditional terms like 'human', 'animate', 'abstract', 'concrete', 'potential location' etc. On this basis, the interpretation procedure tried to match the semantic specification of the affix or of the noun and associate the morphematic entries attached to the verb stem with the most probable elements of

the stem's semantic valency. The program distinguished correctly between compounds like *grammatiklärare* (teacher of grammar) and *universitetslärare* (university teacher), as shown in the following outprint.

```
:- analyse([grammatiklärare])

m([domain(grammar),
   head(m([agent(suff),
           head(teach)]))])

category: n agr(sg, re, indef)

constituents [grammatik, lärare, [lär, are]]

:- analyse([universitetslärare])

m([place(university),
   head(m([agent(suff),
           head(teach)]))])

category: n agr(sg, re, indef)

constituents [universitet, lärare, [lär, are]]
```

The program was also able to interpret somewhat unusual, but fully possible compounds like *universitetsmördare* (university killer). In the case of 'university killer', three alternative interpretations were given, all of them acceptable in Swedish: 1) a person who kills in university buildings, 2) somebody who causes destruction of a university, 3) somebody who uses a university for destructive purposes. The flexibility of the quite simple interpretation procedure and its ability to 'understand' even unusual complex words encouraged us to apply the method tested by means of the toy program for a more serious goal, viz. for interpretation and translation of medical abstracts dealing with asthma and allergy research.

2.3. Translation of compounds within a restricted domain (medical texts on asthma and allergy research)

2.3.1. Domain-related requirements

In order to construct a domain specific lexicon and to design appropriate parsing and translation algorithms, we investigated a corpus of about 140 medical abstracts. Already the preliminary inspection provided evidence for the need of a special procedure for compound interpretation. The frequency of compounds in the texts was extremely high. Cf. the following sample:

*A **large-scale multicenter investigation** was undertaken in 3 cities with comparable **pollen seasons** and atmospheric **pollen concentrations** in order to obtain more definite information about the safety and efficacy of **cromolyn sodium** in the treatment of **pollen-induced** seasonal rhinitis.*

Complex names of chemical substances, as *cromolyn sodium*, do not pose especially great problems to an MT system, since chemical symbols may be efficiently used as interlingual representations. Highly lexicalized and highly idiosyncratic compounds, like *airways* or *hay fever*, may also be stored in the basic lexicon. The main difficulty lies rather in the translation of productive compounds referring to different allergic syndroms, types of medical treatment and patient groups (*ragweed pollen asthma, late-summer rhinitis, flunisolide test, flunisolide patient group* etc.). In different texts, the same syndrom may be referred to by different phrases, e.g. *ragwood asthma, ragwood-induced asthma, ragwood pollen asthma, ragwood-allergic asthma* etc. A correct interpretation of the semantic relations between the constituents of such collocations is necessary for correct translation. Otherwise, a phrase like *childhood asthma* would be translated into French not as *asthme des enfants*, but as *asthme induit par enfance* (lit. asthma induced by childhood–by analogy to e.g. *pyrethrum asthma–asthme induit par pyréthrines*). A procedure for interpretation of compounds and complex NPs must therefore include a kind of domain knowledge, preferably encoded in the lexicon.

2.3.2. The lexicon

An MT system aimed at translation of scientific texts should give the user a possibility of adding new entries to the lexicon in a simple way. A system for medical abstract translation would not be really useful, if the user could not introduce names of new medicines, new terms denoting syndroms, symptoms, treatment methods etc. Since the users of such a system would, with a high degree of probability, be a non-linguist, the linguist designing the method for lexicon extension must adapt the form of interactions to the expected competence of the user.

It would be naive to belive that a non-linguist could manage to specify the lexical items in terms of internal and external theta-roles. Even terms like agent, theme and semantic object would probably cause confusion. Hence, it seems most reasonable to formulate the semantic classification in domain-specific texts (in our case, in terms like allergen, syndrom, body-part etc.). There are actually linguistic reasons for this solution, as scientific sublanguages differ semantically from each other as well as from the everyday conversation language. For a botanist, pyrethrum is primarily a plant belonging to the chrysanthemum family, whereas an allergy re-

searcher regards pyrethrum as an allergy-inducing factor, having much in common with grass pollen and house dust.

In the preliminary model of the lexicon developed until now we classify nouns as members of the following categories:

- syndrom (asthma, rhinitis)
- symptom (sneezing, irritation)
- allergen (pyrethrum, ragwood)
- body part (airways, skin)
- body function (inhalation)
- chemical substance:
  medicine (antihistamine) or
  not used as medicine (histamine)
- medical treatment (injection)
- scientific method (measurement, test)
- time period (season, childhood)
- human: patient or not (the later distinction is needed for correct interpretation of e.g. *asthma patient* and *asthma researcher*)
- amount: mass or countable (dose, group)
- others: concrete or abstract

2.3.3. Interactive lexicon extension

The user has the possibility to classify new nouns to be added to the lexicon by marking the desired alternative in an interaction window. The same entry may be marked as belonging to several categories. For example, *inhalation* may be regarded as both body function and medical treatment (*house dust inhalation/steroid inhalation*). When adding a compound, the user is asked to specify its constituents according to the category list above. New words may be typed in by the user or read in from a text file.

It is assumed that the lexical entries to be added will belong to open lexical classes: nouns, verbs and adjectives. To distinguish between these three classes is not an impossible task for a non-linguist, especially if an appropriate instruction is provided. Adjectives are classified in a way similar to nouns, e.g. *nasal, bronchial*–denoting body part; *stuffy, runny* (as in *stuffy nose*)–denoting symptom and attribute of body part.

A user-adapted classification of verbs is more difficult to achieve. In our preliminary model, the user is presented questions combined with example patterns, for instance: 'Does the verb take an object, like *investigate* ***the effects***?' 'Does it also take a complement with a certain preposition like: *shield the patient* ***from*** *house dust?*' 'What preposition is required?' If the verb in question turns out to be transitive, a further question is asked about the semantic category of the typical object, according to the standard category list. The specification of verbs takes more time than the one of nouns and adjectives. However, the need of introducing new verbs is usually not as great as the need of adding new nouns.

2.3.4. Compound interpretation and generation of target equivalents

The present program covers the most frequent types of compounds found in the corpus. After having filtered out the most frequent verbs (auxiliaries, modals) and items belonging to closed lexical classes (pronouns, articles, prepositions etc.), we first investigated word frequencies, and then the (unfiltered) environment of about the thirty most frequent words. On this basis, we could state that the most usual compounds containing the most frequent nouns (disregarding names of chemical substances) display the following patterns:

i. (attribute, concrete)-allergen-(adj)-syndrom

house dust (allergic) asthma
(grass) pollen (seasonal) asthma

ii. medicine/allergen-medical treatment

antigen injection
allergen injection
steroid treatment

iii. (time period)-adj/allergen/medicine-(body part)-scientific method

allergen (skin) test
9 week double-blind study

iv. syndrom-patient-(countable amount)

hay fever patient group

v. medicine-(patient)-countable amount

steroid patient group
flunisolide group

vi. body part-body function/symptom

skin hyperresponsiveness
airway patency

vii. (attribute, concrete)-allergen-time period

grass pollen season

viii. (medicine/allergen)-medical treatment/body function-time period

steroid treatment period
house dust inhalation period

The procedure for compound interpretation is based on a Prolog formalization of the most frequent patterns. The following program fragment shows what the format for basal lexical entries looks like and how the interpretation rules are constructed.

```
lex([asthma],n,[syndrom],_,_,_).
lex([dust],n,[allergen],_,_,_).
lex([pollen],n,[allergen],_,_,_).
lex([patient],n,[patient],_,_,_).
lex([season],n,[time_period],_,_,_).
lex([steroid],n,[medicine],_,_,_).
lex([grass],n,[concrete],_,_,_).

/* pattern: grass pollen */

tlex([G,P],mean([G,P]),n,
  [allergen],F1,F2,F3):-
      lex([G],_,[concrete],_,_,_),
        lex([P],n,[allergen],_,_,_).

/*pattern: allergen-syndrom:
          ragwood asthma*/

tlex(Tlex,mean(Complex),n,
  [syndrom],A,B,C):-
    append(All,Dis,Tlex),
      lex(All,n,[allergen],_,_,_),
lex(Dis,n,[syndrom],A,B,C),
  append(Dis,[because_of],New),
append(New,[allergen(All)],Complex).

  /* pattern: allergen,complex- (a) - syndrom:
       grass pollen (allergic) asthma */

tlex(Tlex,mean(Complex),n,
  [syndrom],A,B,C):-
    append([Attr,All],Dis,Tlex),
      (lex(Dis,n,[syndrom],A,B,C),
tlex([Attr,All],M,n,[allergen],_,_,_),
  append(Dis,[because_of],New),
    append(New,[allergen([Attr,All])],
      Complex));
        (lex(Dis,a,Sem,_,_,_),
append([Attr,All,Dis],Next,Tlex),
  lex(Next,n,[syndrom],A,B,C),
    tlex([Attr,All],M,n,[allergen],_,_,_),
      append([attr(Dis)],Next,Head),
        append(Head,[because_of],New),
append(New,[allergen([Attr,All])],Complex)).
```

lex = basic lexical entry
tlex = temporary lexical entry

The rules simply specify the default interpretation of a sequence of nouns and deliver a semantic representation coded in 'Machinese English', as shown in the outprints below:

```
:- interpret([house, dust, asthma, patient])

mean([patient,
  suffering_from,
    syndrom ([asthma,
        because_of,allergen([house, dust])])])
grammatical category : n
semantic category : [patient]

:- interpret([house, dust, inhalation])

mean([inhalation,
  of_object,
    allergen([house, dust])])
grammatical category : n
semantic category : [body_function]
```

The Machinese representations can without difficulties be matched with the appropriate target morphosyntactic patterns. For example, the semantic representation of *grass pollen asthma patient* becomes associated with the Polish pattern (simplified notation):

```
patient,suffering_from,
syndrom([X,because_of,allergen(Attr,All)]) -->

n(patient,Agr,nom),
  prtact(suffer,Agr,nom),
    prep(suffer,Prep,Case),
      n(X,Agr2,Case),
        prtpass(cause,Agr2,Case),
          n(allergy,Agr3,ins),
            prep(_,na,ack),
              n(All,Agr3,ack),
                n(Attr,agr(Gen,pl),gen).
```

ins = the instrumental case

The pattern above correctly generates the Polish equivalent of *grass pollen asthma patient*:

pacjent cierpiący na astmę
patient suffering prep asthma-acc

spowodowaną
caused-acc

uczuleniem na pyłek kwiatowy traw
allergy-ins prep pollen-acc flower-adj grass-gen

In a similar way, the program disambiguates *ragwood asthma* and *childhood asthma* when translating into French. Still, certain ambiguities may remain: the present program can, for example, not decide whether *grass pollen asthma* should be translated into French as *asthme in-*

duit par pollen des graminées or *par pollen de l'herbe*. The decision has to be made by the user.

Translation of frequent compounds of the type noun+past participle (*allergen-shielded, allergen-tested, placebo controlled*) is handled in a way similar to the one used in the prototype program when translating compounds like *university teacher* and *university killer*. The semantic category of the noun is compared with the semantic specification of the valency of the verb stem and the noun is associated with the most probable verbal argument. Thus, *allergen shielded room* is interpreted as 'a room shielded from allergen', while *allergen tested skin* gets the reading 'skin tested by exposure to allergen'.

3. Conclusions and implications for further research

3.1. Remaining problems

The method proposed here has so far led to good translation results. However, the problem lies not only in interpreting a compound, but also in identifying an English word sequence as a compound. For the time being, we use a parsing procedure based on a combination of dependency grammar and categorial grammar. The main parsing difficulty, when dealing with an English input, is to decide whether a lexical stem functions as a finite predicate or as a nominal. We try to remove the ambiguity by starting the parsing by a procedure called 'verbfinder', searching for possible candidates for the predicate function. The function of ambiguous items, like *result, control* etc., may often be identified on the basis of their evironment: if the word in question is immediately preceded by a preposition and/or an article, it can be easily identified as a nominal element. The parsing procedure may still be made more efficient by utilizing results of statistic investigations of the corpus (Steier & Belew 1991, Johansson 1993).

3.2. Future plans

The advantage of the model outlined here lies in the fact that the general approach to the grammar underlying the translation system may be adapted to different domains without violating any theoretical assumptions. However, the theory solely does not guarantee a high-quality translation. The preliminary system outlined above is to be developed and improved along the following lines:

◊ statistical methods will be used in order to reduce ambiguities and to discover coocurrence patterns on the basis of larger corpora
◊ the medical vocabulary will be enlarged by using large computational medical data-bases (e.g. MEDLINE) and by consulting specialists who are native speakers of the languages involved in the system
◊ the interactive procedures will be evaluated and refined by testing their usefulness in experiments with non-linguists.

The results of the corpus investigations and the experiments with translation of abstracts are to be used in a system for automatic abstracting and multilingual abstract generation.

References:

Anderson, S. 1982. Where's morphology? *Linguistic Inquiry* 13. pp. 571-612.

Bierwisch, M. 1989. Event nominalization: Proposals and problems. Motsch, W. (ed.): *Wortstruktur und Satzstruktur*. Berlin: VEB (=Linguistische Studien, Reihe A 194). pp. 1-73.

Chomsky, N. 1970. Remarks on nominalizations. R. Jacobs & P. Rosenbaum (eds.): *Readings in English Transformational Grammar*. 184-221. Waltham: Ginn & Co.

Di Sciullo, A.M. & E. Williams. 1987. *On the definition of word*. Cambridge: MIT Press.

Fanselow, G. 1985. What is a possible complex word? J. Toman (ed.): *Studies in German Grammar*. Dordrecht: Foris. pp. 289-318.

Fanselow, G. 1988. 'Word Syntax' and semantic principles. G. Booij & J. v. Marie (eds.): *Theorie des Lexikons*. Universität Düsseldorf. pp. 1-32.

Jackendoff, R. 1975. Morphological and semantic regularities in the lexicon. *Language* 51. pp. 639-71.

Johansson, C. 1992. Using a statistical measure to find relations between words: Semantics from frequency of co-occurrence? (MS)

Karlsson, F. 1990. Constraint grammar for parsing running texts. *CoLing '90*, Helsinki. pp. 168-173

Koskenniemi, K. 1990. Finite-state parsing and disambiguation. *CoLing '90*, Helsinki. pp. 229-32.

Lees, R. 1960. *The grammar of English nominalizations*. The Hague: Mouton.

Olsen, S. 1991. Zur Grammatik des Wortes: Argumente zur Argumentstruktur. *Theorie des Lexikons*. Universität Düsseldorf. pp. 31-58.

Pesetsky, D. 1990. *Experiencer predicates and universal alignment principles*. Cambridge: MIT Press.

Selkirk, E. 1982. *The syntax of words*. Cambridge: MIT Press.

Steier, A.M. & R.K. Belew. 1991. A statistical analysis of topical language. R. Casey & B. Croft (eds.): *2nd Symposium on Document Analysis and Information Retrieval*.