Automatic translation of nominal compounds
A case study of Danish and Italian

Patrizia Paggio & Bjarne Ørsnes

This paper presents a case study on the automatic translation of a number of endocentric Danish nominal compounds into Italian. Two main issues are focussed on, namely what constitutes a nominal compound in Danish, and whether a semantic account of non-argumental compounds can help find the correct Italian equivalents. As regards the former issue, we claim that the traditional definition of a Danish compound as a typographical unit is inadequate, as it fails to treat sequences of relational adjectives and nouns as compounds. We then show that accepting to treat such sequences as nominal compounds has interesting consequences for a general theory of compounding, since it means acknowledging the fact that the non-head can function as the external argument, and that internal arguments can be realised outside the compound itself. As for the controversial issue of a semantic treatment of non-argumental compounds, we present a semantic model inspired by Levi (1978), and show how it can be profitably used to capture some regularities in the translation of the examples occurring in our corpus into Italian.*

0. Introduction.

The object of this paper is twofold. On the one hand, it provides an account of a case study on the automatic translation of a set of Danish nominal compounds into Italian.[1] Although we focus on analysis of the Danish data, our strategies are clearly motivated by translational needs, and translational evidence is continually provided. On the other hand, the paper relates the problems encountered and the solutions adopted in this research project to other work on compounds, not necessarily

* *Acknowledgements:* this paper is based on two parallel research projects carried out under the Eurotra programme, the first one in cooperation with Gruppo Dima (Torino), and the other with Eurotra-GB (Umist, Manchester), Eurotra-ES (Barcelona), Eurotra-DE (University of Saarbrücken, Eurotra-BE (Leuven). We would also like to thank Michael Mudrow, Indiana University, Peter Molbæk Hansen, University of Copenhagen, and two anonymous referees for their invaluable comments.

[1] This article is inspired by work carried out in the framework of the Eurotra Machine Translation Programme. A fairly detailed description of the linguistic specifications of the system can be found in Copeland *et al.* (1991a).

 (ricevuto nel dicembre 1992)

connected to Machine Translation. We believe in fact that an elucidation of the concrete problems encountered in an applied context such as MT, especially in connection with naturally occurring data, may provide fruitful feedback to the general theoretical discussion.

In line with this philosophy, we base our observations on the empirical study of a corpus of Danish nominal compounds extracted from real texts dealing with telecommunications and information technology.[2] The choice of a limited domain has important consequences for the results obtained in our work, as will be shown below. It will suffice here to stress the fact that we have tried to substantiate our claims with naturally occurring linguistic material as far as possible. In particular, the experimental data discussed in section 3 relate exclusively to our corpus.

As Ananiadou & McNaught (1990) put it, "MT systems have generally failed to tackle the problems of compounds in anything other than an ad hoc fashion", although the same authors admit that exceptions to this general failure may be found. Indeed, treating compounds in an MT system means addressing a number of quite difficult problems. Limiting ourselves to the European languages, the most crucial ones are perhaps how to cope with the productivity of compounding, especially in Germanic languages, and how to map Germanic compounds onto corresponding phrasal expressions in Romance languages and vice versa. The strategy adopted in this work to cope with the first problem is that of analysing compounds into their component parts. Thus, the word

(1) Datakommunikation
'data communication'

will be decomposed into *data* and *kommunikation*. In this way, the lexicon needs only to contain the single components for the system to be able to accept unknown compound forms. However, a compositional treatment is not possible for words such as

(2) Sommerfugl
lit. summer bird 'butterfly'

which, although morphologically complex, have acquired a specialised meaning not derivable from the meanings of the component parts. Assuming that such forms must be listed in the lexicon in their entirety, complete with the relevant translations, we shall exclude them from our discussion.

[2] We have used the following text corpora:
* ESPRIT - a EEC text consising of 25,000 running words, translations of which exist in all European languages.
* Text corpus on telecommunications, 17,000 running words, parallel versions of which exist in all European languages.

Once a compositional strategy for the analysis of compounds in Germanic languages is chosen, a number of other issues arise concerning their identification and interpretation. A discussion of these issues, focussing on the morphological and syntactic analysis of Danish endocentric compounds, constitutes the first section of our paper. In particular, we show how the definition of compound traditionally used in Danish grammars fails to account for all the forms that a theory of compounding ought to be able to describe. Consequently, we propose a revised definition, and assess its consequences in light of well-known theories of compounding such as that described in Selkirk (1982).

The discussion of the syntactic properties of Danish compounds is concluded by a syntactic typology of the nominal compounds occurring in our corpora, which forms the second section. Finally, the third part of this paper deals with a rather controversial issue, namely semantic analysis of compounds, and shows how, for the specific domain chosen, a semantic classification of a certain class of nominal compounds seems both viable and useful for achieving a better translation.

This section has an experimental character, as the procedures described can be seen as an emulation of what the computer's behaviour would be under controlled circumstances. However, the specifications provided are clear and detailed enough to constitute the basis for a real implementation.

The choice of Danish and Italian as the two languages treated in this work may seem unusual. However, the problems that automatic translation of Danish compounds into Italian causes are typical of translation of Germanic compounds into any Romance language. Therefore, we hope this paper can be of some interest to a larger public than those privileged few who happen to be familiar with the two languages under consideration.

1. Some general issues.

Dealing with compounds in an MT environment means facing the following four problems (cf. Bouillon *et al.* 1992):

1) *Identification*
What distinguishes compounds from other words and phrases?

2) *Segmentation*
What are the constituents of a compound, and how can a compound be delimited considering the fact that it is not always written as a single unit?

3) *Disambiguation*
What is the correct structural representation of a compound?

4) *Interpretation*
How can the meaning of a compound be derived in a compositional fashion and what role does the semantic relation between the head and the non-head in a compound play in translating a Germanic compound into a Romance noun phrase?

In this section we shall address the first three issues, paying special attention to the definition of a compound, as our criteria for identifying compounds in certain respects differ from the traditional view. But even though our approach is biassed by its foundation in MT we shall try to relate our findings to the ongoing theoretical discussion.

1.1. Criteria for compoundhood in Danish.

In many recent discussions on compounds (Selkirk 1982, Bauer 1988) the matter of the very definition of a compound is not given much space, leaving compounds to be identified on a rather intuitive basis. Instead, focus is put on the properties of compounds (such as Headedness), the formal devices for compound formation, the place of compound formation in the grammar, the cyclic ordering of compound formation in relation to other processes of word formation, the base of compound formation, etc.

However, automatic identification and translational considerations call for an assessment of the definitional issue. In this paper, we shall only be concerned with the definition of compounds in Danish and not address the issue of how to distinguish between compounds and syntactic phrases in Italian, or Romance languages in general.

In distinguishing a compound from a syntactic phrase several language specific criteria can be applied for Danish.

1.1.1. Phonological factors.

Two phonological factors distinguish a compound from a syntactic phrase, namely a heavy stress, and the loss of the glottal stop on the first element. However, there are exceptions to both rules (cf. Bauer 1978). Thus the presence of these phonological features can be used as additional evidence of compoundhood rather than a defining characteristic. However, in this article we consider automatic translation of written text, and therefore phonological criteria have no place in the definition and delimitation of compounds for our purposes.

1.1.2. Words vs. Stems.

In Danish compounds the non-head is usually non-declinable. This is for example the case in adj + noun compounds, where the adjective takes the form of a stem, i.e. a common generic form. Thus, *totalbudget* 'total budget' and not *totaltbudget* is correct, although the word *budget* is neuter, and the neuter form of an adjective ends in *t*. Similarly, *storbyer* and not *storebyer* is the plural of *storby* 'big city'/'metropolis', although the plural of *stor* would be *store*. Rare exceptions, where the adjective appears to be inflected, are such lexicalised compounds as *storetå* 'big toe' and *Storebælt* 'the Great Belt' (further examples can be found in Hansen 1967). In these cases, the adjective appears in the form of a singular definite adjective (a form which is identical with the above-mentioned plural form). Interpreting these forms as containing a linking element is problematic, as linking elements are very rare in Danish adj + noun compounds and, contrary to noun + noun compounds, no linking element occurs which cannot simultaneously be interpreted as a regular inflectional affix for the adjective in question, showing agreement with the head noun. As noted above, however, these inflected adjectives only occur in few established lexicalised compounds, and not in newly coined ones, e.g. *nyvognssalg* lit: 'new car sale', so inflection of adjectives cannot be considered a valid generalisation concerning the productive word formation process of Danish compounding.

Also in noun + noun compounds, the general rule that the first element is non-declinable applies in an overwhelming majority of cases.

These well-known facts (Hansen 1967, Scalise 1986) have led to the Ordering Hypothesis of word formation rules in the lexical component of the grammar, according to which compounding rules apply before inflectional affixation. As has been shown by Selkirk (1982) and Scalise (1986), this hypothesis is untenable. Selkirk uses this insight to show that compounds in her context-free rewriting system are rewritten as consisting of Words and not Stems, and Scalise shows that the output of the inflectional rules component in his lexical model can be input for the compounding rules.

The problem of inflection internal to compounds does require some consideration in Danish, as the non-head component of a Danish noun + noun compound frequently occurs with the affix *-s* or *-e*. It can, however, be argued that these affixes constitute linking elements and not inflectional affixes. We thereby maintain the overall rule that the non-head part of a Danish compound is composed of Stems and not Words (unlike English compounds according to Selkirk's analysis). This implies that the category of the compound components is a language-specific parameter.

1.1.3. A traditional morphosyntactic criterion.

The largest number of compounds in our texts (and in Danish in general) are written as one word, with the possibility of various linking elements between the component parts (as outlined above). Our primary criterion for defining a compound is thus: 'a typographical unit that can be split into individual lexemes'.

As a number of nouns systematically undergo a morphological change in the process of compounding, thereby turning into bound morphemes, this criterion may have to be relaxed. For example, in the compound *arbejdsdeling* 'division of work', the left-hand component, *arbejd-* is a truncated version of the noun *arbejde* 'work'. However, *arbejdsdeling* is treated as a compound because of the obvious morphological and semantic similarity between *arbejd-* and *arbejde*. The *-s-* affix between the two components is considered a linking element, as explained in section 1.1.2.

Capturing compounds automatically means either listing the compounds in the dictionary (cf. Bouillon *et al.* 1992) or applying a segmentation procedure. Here, we assume an affix-stripping procedure, segmenting the input word into all possible morphologically relevant parts (including linking elements). On the basis of this segmentation compounds are assigned an internal structure by the morphological grammar.

1.1.4. Relaxing the traditional definition: anglicisms.

A compound can also be defined on syntactic grounds as "A sequence of two or more nouns with no overt manifestation of subordination of one of the parts where the first noun cannot be interpreted as a specifier"

(3) printer outlet

(4) data kommunikation
'data communication'

A tendency to write nominal compounds with a space between the single components can be observed if one or both of the nouns are English words, as in (3), but also when all the components are Danish words (although loanwords) as in (4). These items would normally be written as one orthographical unit and must be interpreted as compounds. In fact in Danish, a sequence of two nouns with no overt manifestation of subordination (e.g. genitive marking or prepositional phrase) is a grammatical phrase only in very few types, e.g. *en kop kaffe* 'a cup of coffee'. So, unless the first noun is coded as a specifier in the lexicon, the sequence will be treated as a compound.

1.1.5. More controversy: relational adjectives.

The final criterion for identifying a Danish compound relates to cases which have otherwise been analysed as phrases. The criterion is syntactic in nature because it applies to a sequence of words as well as to syntactic properties of the non-head part, but also semantic because it makes reference to semantic properties of the components involved. It states that "A sequence of adjective + noun, where adjective is a relational adjective, is a compound"

The most prominent treatment of relational adjectives can be found in Levi (1978), whose terminology we shall adopt.[3]

Levi establishes a set of tests to identify relational adjectives, of which, however, only a subset apply to Danish.

a) Relational adjectives are denominal (although the opposite is not always true)

(5) ministerium 'Ministery' => ministeriel 'ministerial'

(6) nation 'nation' => national 'national'

A very frequent way of forming relational adjectives is by means of the suffix *-mæssig* 'related to':

(7) bymæssig
'urban'

(8) undersøgelsesmæssig
'related to an inquiry'

(9) ressourcemæssig
'related to resources'

b) Relational adjectives are non-predicating

(10) * Angrebet er ministerielt
'the attack is ministerial'

(11) * Bebyggelsen er bymæssig
'the built-up area is urban'

[3] A recent discussion of these adjectives is contained in Giorgi and Longobardi (1989) where referential adjectives are contrasted with predicating ones. From the discussion in Giorgi and Longobardi it is however doubtful whether we are dealing with two terms for the same concept or two groups of adjectives where one group is a proper subset of the other. We shall not pursue the matter here, as it has no consequence for our basic assumptions.

c) Relational adjectives show non-degreeness

(12) * Et meget ministerielt angreb
'a very ministerial attack'

(13) * En meget bymæssig bebyggelse
'a very urban built-up area'

d) In combination with deverbal nouns relational adjectives manifest case-relations.

(14) Et ministerielt angreb - agentive
'a ministerial attack'

In fact, (14) above can be paraphrased as shown in (15):

(15) Et angreb fra ministeren
'an attack by the minister'

Giorgi and Longobardi (1989) claim that referential adjectives (relational in our terminology) always express an external semantic function. And it is a fact that relational adjectives derived from nouns semantically capable of being agents in combination with a deverbal head noun are always interpreted as agents:

(16) Tysklands besættelse
'the occupation of/by Germany'

(17) Den tyske besættelse
'the German occupation'

While the first example is ambiguous as to the thematic role of the genitive NP, only an agentive interpretation is possible in the second example. In Danish, however, a relational adjective does not necessarily manifest an external semantic function as can be seen from the following example, where the external argument is realised outside the compound and the relational adjective maintains an instrumental relation to the head noun of the compound:

(18) FN's økonomiske sanktioner mod Irak
'the United Nations' economic sanctions against Iraq'

A very frequent semantic function realised by relational adjectives is the "concerning" relation as in the above-mentioned Danish *-mæssig* derivation of adjectives. Giorgi and Longobardi (1989) claim that this semantic relation occupies a position in between external and internal

semantic functions. But since Danish allows for compounds where the relational adjective is interpretable as an internal argument of the head noun, as shown above (example (18)), it seems unnecessary to postulate the existence of an intermediate status of arguments.

Thus we claim that relational adjectives can function as internal arguments. Interestingly, however, the *-mæssig* adjectives usually occur in connection with non-deverbal head nouns and deverbal head nouns that have lost their verbal semantics, thus excluding an argumental interpretation.

e) A relational adjective can only be coordinated with another relational adjective

(19) Det tidsmæssige og økonomiske pres
'the time and economic pressure'

(20) * Det hårde og økonomiske pres
'the hard and economic pressure'

To this we may add that relational adjectives behave differently in different languages. In Italian, for instance, they always occur in postnominal position:

(21) L'invasione italiana
'The Italian invasion'

(22) * L'italiana invasione

It must be noted that some denominal adjectives have both a relational and a non-relational reading, e.g. *national*, in the reading 'patriotic' is predicating and shows degreeness, and in the sense 'bound to a nation' is relational.

Having established what a relational adjective is, we can now discuss why we agree with Levi (1978) that a sequence of a relational adjective followed by a noun is a compound.

On the basis of a monolingual analysis it can be established that the combination relational adjective and noun forms a semantic unit, as can be seen from:

(23) *Det hårde og økonomiske pres
'the hard and economic pressure'

(24) Det hårde økonomiske pres
'the hard economic pressure'

where *hårde* 'hard' applies to the combination *økonomisk pres* 'economic pressure'. Predicative adjectives can both function as conjuncts to another predicative adjective as in (25) and as modifiers to an adjective + noun complex, as in (26).

(25) En ny og dygtig elev
'a new and clever pupil'

(26) En ny dygtig elev
'a new clever pupil'

Often a compound and an NP containing a relational adjective are interchangable:

(27) tidspres
'time pressure'

(28) tidsmæssigt pres
lit. temporal pressure

And occasional gaps occur, which means that a relational adjective alternates with the corresponding noun according to the head noun in question:

(29)	økonomisk krise 'economic crisis'	* økonomikrise 'economy crisis'
(30)	økonomiminister 'minister of economy'	* økonomisk minister[4] 'economic minister'
(31)	økonomistyring 'economy planning'	* økonomisk styring 'economic planning'

From a translational point of view a compound in one language may denote the very same concept as an NP with a relational adjective in another language:

(32)	økonomisk krise 'economy crisis'	Wirtschaftskrise
(33)	atomic energy	atomkraft

Treating sequences of relational adjective and noun as compounds, however, has the effect of systematically violating Selkirk's subject-argument generalisation and First Order Projection Condition. As will

[4] *Økonomisk minister* means 'economical minister'.

be argued below these principles are not valid for "traditional" Danish nominal compounds either. Hence, the treatment of the sequence relational adjective plus noun as a compound cannot be rejected on these grounds.

Selkirk's subject-argument generalisation states:

"The SUBJ argument of a lexical item may not be satisfied in compound structure". (Selkirk 1982: 34)

As noted above, however, a relational adjective derived from a noun semantically capable of being an agent (possibly through metonymic use) always realises the external argument:

(34) Den italienske invasion
'The Italian invasion'

If we turn to Danish nominal compounds they too seem to occur with a non-head filling the external argument slot:

(35) regnafbrydelse
'interruption by rain'

(36) ekspertbedømmelse
'expert judgment'

(37) regeringsindgreb
'governmental intervention'

(38) EF-redegørelse
'European Community report'

(39) pigesvømning
lit. girl swimming

Such an analysis is however not unproblematic. In example (35) the head is derived from an ergative construction, thus it can be claimed that we are dealing with an internal argument after all. The compounds in example (36) and (37) can occur with a genitive NP seemingly realising the external argument:

(40) Dr. Strangeloves ekspertbedømmelse
'the expert judgment by Dr. Strangelove'

(41) De konservatives regeringsindgreb
'The governmental intervention by the Conservatives'

In these examples the non-head part of the compound seems to be qualifying rather than agentive. This, however, is only possible in connection with nouns semantically capable of expressing properties. In examples (38) and (39) this is no longer possible and the non-head must be interpreted as realising the external argument. This seems only to be possible with collective nouns or nouns with generic reference which are not considered in Selkirk (1982). Leaving aside the exact interpretation of the compounds in (36) and (37) we claim that the realisation of the external argument in the non-head part of a Danish compound is possible although it is subject to certain restrictions. These restrictions seem to be in accordance with Di Sciullo and Williams' Atomicity Condition (cf. Di Sciullo & Williams 1987), stating that words are basically generic. Thus, according to this principle, the non-head in a compound can never have specific reference, but counter-examples can be mentioned, e.g. *computerskærm* can be a 'computer screen' as well as the screen of a specific computer. We shall not, however, pursue the issue any further here. In conclusion, Selkirk's subject-argument generalisation is not an argument against treating the combination of a relational adjective and a noun as a compound.

Selkirk's First Order Projection Condition states:

"All non-SUBJ arguments of a lexical category Xi must be satisfied within the first order projection of Xi", which means that all internal arguments of the head of a compound must be satisfied within the compound immediately dominating the head". (Selkirk 1982: 37)

This does not hold for the combination relational adjective plus noun:

(42) Den tyske besættelse af Danmark
'The German occupation of Denmark'

Danmark is an internal argument of *besættelse* 'occupation' but - considering *tyske besættelse* 'German occupation' a compound - it is nevertheless realised outside the compound in postnominal position. Further examples can be added:

(43) Et ministerielt angreb på oppositionen
'A ministerial attack on the opposition'

The realisation of an internal argument outside the first order projection of the head is however also possible in connection with noun + noun compounds:

(44) EF-redegørelse for sagen
'Community report on the matter'

Even in English:

(45) staff exploitation of privileges

Once again, although relational adjectives do not conform to a general rule supposedly valid for compounds in general, this deviance does not in itself suffice to raise doubts about their status as parts of nominal compounds, since other compound types can be found which do not obey the rule either. It is rather the generality of the First Order Projection Condition, as well as of the subject-argument generalisation previously discussed, which may be called into question. To conclude, it seems to us that the evidence in favour of a treatment of relational adjectives as semantically equivalent to the corresponding nouns is clearly stronger than any counter-arguments. Therefore we have chosen to include sequences of relational adjectives followed by nouns in our typology of Danish nominal compounds.

2. *A Typology of Danish nominal compounds.*

In order to illustrate and exemplify the foregoing discussion we present here a typology of Danish nominal compounds. As already mentioned, we have based our observations on multilingual corpora, from which suites of examples in various European languages were extracted. The examples have been organised into a typology based on the following parameters: the morphological source of the head (deverbal or simple noun) and the deep-syntactic relation holding between the head and the non-head of the compound. The former distinction is also known as the distinction between primary and synthetic compounds (cf. Scalise 1986: 90). It should however be noted that we stick firmly to the morphological criterion, contrary to Selkirk (1982) who considers verbal compounds only those forms where the non-head occupies an (internal) argument slot in the thematic grid of the verb. We call a compound with a deverbal head a verbal compound, irrespective of whether the non-head is argumental or not. As for the relation between typology and structure in compounds, the typology only makes two structural claims:

1. A compound is headed
2. A compound is a binary branching structure

It does not limit itself to two-element compounds as compounding rules may be applied recursively maintaining the binary structure, nor does it presuppose any directionality in the recursive application of compounding rules. Right- as well as leftbranching compounds fit into the scheme. It is also neutral with respect to right- or leftheadedness of compounds, even though Danish endocentric compounds always seem to be rightheaded, also in compound formation of verb + particle (which is one of the examples of leftheaded English compounds):

(46) løbe ud => en udløber
'run out' 'an offshoot'

(47) ringe op => en opringning
'call up' 'a call'

As for the deep syntactic relationship, the following labels have been used: Arg1 specifies the agent or the logical subject of a predication or the entity modifying a property-denoting simple noun (e.g. *antennediameter* 'diameter of the aerial'). Arg2 is assigned to the theme or the logical object of a predication, while obl denotes the prepositional object of a predicate (e.g. *forsyne med* 'provide with') and mod indicates a non-valency bound constituent.[5]

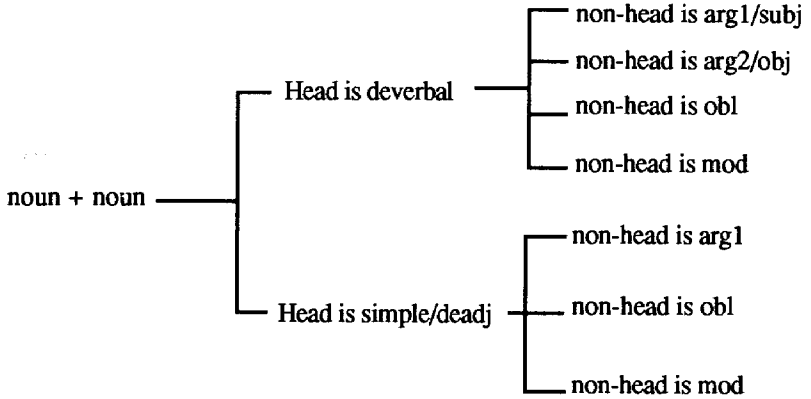

Figure 1. - *A typology of Danish nominal compounds.*

Exemplification:

	Noun + Noun	Adj + Noun
Arg1 + deverbal noun:	lægeerklæring 'doctor's certification'	ministerielt angreb 'ministerial attack'
arg2 + deverbal noun:	arbejdsdeling 'division of work'	legemlig udvikling 'physical development'
obl + deverbal noun:	programforsyning 'supply of programs'	økonomisk sanktion 'economic sanction'
mod + deverbal noun:	forskningsarbejde 'research activity'	skattemæssig afskrivning 'tax write-off'

[5] Our theory of deep syntactic roles is based on the EUROTRA case-frame assignment. Cf. Copeland *et al.* (1991a).

arg1 + simple noun:	antennediameter 'diameter of the aerial'	kulturelt niveau 'cultural level'
obl + simple noun:	anvendelsesmulighed 'application possibility '	eksportmæssig mulighed 'export possibility'
mod + simple noun:	kontorsystem 'office system'	økonomisk krise 'economy crisis'

A point that we have not touched upon until now is the assignment of deep syntactic function to the non-head part of a verbal compound according to the typology. Looking at our corpus this will not necessarily present a major problem as certain patterns are clearly recognisable. For instance, deverbal compounds derived from transitive verbs will usually occur with a non-head part filling the arg2 slot. In ambiguous cases we foresee the use of preference rules based upon heuristic observations such as described in Bennett & Paggio (1993).

3. Semantic analysis of non-argumental compounds: a challenge.

As the typology above shows, we distinguish between argumental and non-argumental compounds. In compounds of the former kind, the right preposition in the Italian equivalent can be generated by looking at the valency requirements of the head-noun, e.g.

(48) a. anvendelsesmulighed => possibilità di applicazione
'application possibility'

b. arbejdsdeling => divisione del lavoro
'division of work'

In both examples, the preposition *di* would be generated on the basis of the valency frame of the head noun as the correct preposition for the object.

However, distinguishing automatically argumental from non-argumental compounds and assigning the right syntactic label to the non-head component is in practice not unproblematic. Furthermore, the translational problem is not completely solved by finding the right preposition, as the mapping of the component nouns onto the correct forms in the target language may also present the system with embarassing choices. The English translation of *deling* in (48b) is symptomatic of this problem, as the same word may well correspond to *share* instead of *division* in a different context.

Although we recognise the seriousness of both issues, we have chosen in this study to focus on the analysis and translation of a different class

of nominal compounds, the treatment of which perhaps constitutes an even greater challenge.

By analysing our corpus manually, we realised in fact that the portion of compounds where the non-head could be said to function as either an external or internal argument of the head only constituted a minority of the cases, namely 116 out of 403 type instances.

The remaining 287 forms belong to the class that we call non-argumental, because the non-head part is not an argument of the head.[6] Note that the head can be either a deverbal or a simple noun, e.g.

(49) a. satellitkommunikation
'satellite communication'
b. forskningsemne
'research topic'

In the literature, compounds of this type are also referred to as 'nonverbal compounds' (Selkirk 1982), or compounds derived by 'predicate deletion' (Levi 1978).

Additional names can be found, but we have chosen to mention these two authors because they stand for diametrically opposite approaches to the issue of non-argumental compounds. Selkirk (1982: 25), in line with Jespersen (1942), states that "...for non-verbal compounds, the range of possible semantic relations between the head and nonhead is so broad and ill defined as to defy any attempt to characterize all or even a majority of the cases".

Contrariwise, Levi (1978) claims that the number of semantic relations is limited, and postulates an actual semantic classification of non-argumental compounds. The position we shall defend in this paper is that such a semantic classification may be possible for a restricted domain, and may help achieve a better, although not flawless translation of non-argumental compounds.

We do not, however, make any claims of universality for the classification used.[7] Before we proceed to describe the actual strategy adopted, a more detailed discussion of the translational difficulties we are faced with seems in order.

[6] The distinction between argumental and non-argumental compounds is not totally unproblematic. Warren (1978), for instance, notes a difficulty arising from this distinction, namely that compounds displaying the same semantic content end up in different categories (e.g. sugar bowl vs. sugar container). However, she also admits that, in the majority of cases, argumental compounds do not fit the same semantic classes she uses to interpret non-argumental forms.

[7] An interesting exercise would be to compare the semantic model we have adopted, which is, with few slight changes, the same used in Levi (1978), with similar descriptions, e.g. the classification in Warren (1978). A superficial inspection already shows that most of the classes occur in both systems, although their definition and names may vary.

3.1. Translational problems.

One of the main problems related to the translation of compounds is that they are not always translated by compounds. Many nominal compounds from Germanic languages are translated by syntactic structures (often noun + prep + noun) in Romance languages. As Bennett (1993: 78) rightly points out, however, "here we encounter definitional problems, since it is not clear how many (if any) of such Romance sequences are compounds themselves". To facilitate the mapping between Germanic compounds and Romance syntactic structures, he proposes a syntactic representation of compounds, to which we subscribe. An example of such a representation would be as shown in fig. 2. Note that a dummy preposition is inserted to facilitate the generation of a prepositional phrase in a Romance language.

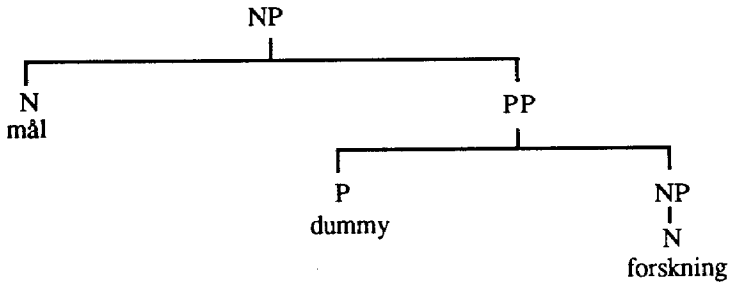

Figure 2. - *Representation of non-argumental compound (forskningsmål 'research objective').*

Having set up a general representational model for non-argumental compounds, let us consider the translational problems encountered in our corpus in more detail. Our corpus shows that at least the following realisations of a Danish non-argumental nominal compound are possible in Italian:

N + N

(50) basisbånd 'base band' => banda base

N + Adj

(51) jordoverflade 'earth surface' => superficie terrestre

N + P + N

(52) adgangsmetode => metodo d'accesso
'method of access'

If the Italian translation is an N + P + N sequence, a number of different prepositions can be used: *di / per / da* ... Additional problems regard definiteness and number of the non-head part:

(53) informationsteknologi => tecnologia dell'informazione (definite)
'information technology'

(54) reparationsteknik => tecnica di riparazione (indefinite)
'reparation technique'

(55) kabelnet => rete di cavi (plural)
'cable network'

(56) kobbergitter => grata di rame (singular)
'copper bars'

However, we shall ignore problems of definiteness and number, as the objective of our study was devising a method to achieve the correct syntactic realisation, as well as the generation of the right preposition where a preposition is necessary.

3.2. A semantic typology.

Having the ultimate goal of solving, or at least alleviating, such translational problems, we attempted a semantic classification of the non-argumental compounds contained in our corpus according to the typology described in Levi (1978). This typology is rooted in a transformational treatment of compounds, according to which nominal compounds (or rather complex nominals if we follow Levi's terminology) are divided into two main groups on the basis of the underlying deep structure. To account for the different surface realisations, Levi defines two main transformation rules, called 'predicate deletion' and 'predicate nominalisation'. The former would be responsible for the non-argumental compound type, and the latter for the argumental type. Thus, only 'predicate deletion' is relevant to our context. According to Levi's theory, the deep structure of complex nominals to which this obligatory transformation applies contains one of a set of semantic predicates for which she provides a list.

Although we have no intention of pleading for a transformational account of nominal compounds (cf. Selkirk (1982) for a disputation of the transformational view of compounding), we think the typology deriving from the list of semantic predicates is an interesting descriptive model. Clearly, if one does not subscribe to the transformational

framework Levi's model originates from, the actual categories this model is based on must be given a different interpretation. In our view, they can be seen as labels for the semantic relations that hold between head and non-head, without having to correspond to concrete nodes in a postulated deep structure. Having clarified this, the semantic typology is shown in fig. 3, with examples from Levi's material as well as from our corpus.

CLASSES	Levi's examples	Our examples
CAUSE	a. tear gas b. drug death	b. koordineringsproblem
HAVE	a. salt lake b. lemon peel	a. kanaludstyr b. forskningsmål
MAKE	a. honey bee b. nervous system	b. antennesystem
USE	pressure cooker	elektronikcentral
BE	head noun	nøglefaktor
IN	city folk	(head is place): aktivitetssektor (head is time): leveringstid
FOR	nose drops	adgangsmetode
FROM	olive oil	forskningsresultat
ABOUT	tax law	satellitprojekt

Figure 3. - *Semantic classes for non-argumental compounds.*

Note that for a few classes an 'a' and a 'b' example are given. In all the 'a' examples, the head of the compound is also the agent of the semantic relation, whereas in the 'b' examples the syntactic modifier is the agent. This is especially relevant for the category HAVE, so we shall refer to the 'a' subgroup as $HAVE_1$ and the 'b' subgroup as $HAVE_2$.

From a monolingual perspective, we found the descriptive adequacy of the typology satisfactory as it could accommodate all the compounds in our list, provided we adjusted the definition of a single semantic class (IN). The table in fig. 4 shows the number of examples for each class.

Class	n°
for	72
in	26
be	19
have	15
about	11
make	9
from	5
cause	3
use	1
Total	161

Figure 4. - *Distribution of semantic classes in the corpus.*

These figures partly confirm Levi's predictions regarding the productivity of each type, as shown by the diagram in fig. 5.

$have_1$, cause, make, from	-
use, be, about	
for, in, $have_2$	+

Figure 5. - *Productivity of semantic classes according to Levi (1978).*

Having assigned a semantic class to all our examples, we looked for regularities in the relation between the Italian translations and the various semantic groups. The diagram in fig. 6 shows how different syntactic structures in the Italian equivalent forms distribute over the various semantic classes. Although the picture may seem rather heterogeneous, a number of generalisations can indeed be made:

a) *di* is by far the most used preposition

(57) modtagerretning => canale di ricezione
'reception channel'

(58) antennesystem => sistema di antenne
'aerial system'

b) a translation rule can be set up for compounds in the BE class, namely:

BE => N + Adj / N + N

Non-head syntax	*n°*	*Class*
di + N	8	$have_2$
	3	cause
	6	make
	23	in
	60	for
	3	from
	10	about
a + N	5	$have_1$
	1	lex.
per + N	5	for
	1	lex.
su + N	1	about
Rel adj	6	be
	5	for
	3	in
	2	from
	2	$have_2$
	1	use
N	7	be

Figure 6. - *Distribution of translational variants over semantic classes.*

(59) naboposition => posizione vicina
'neighbouring position'

(60) basisbånd => banda base
'base band'

c) similarly, for the $HAVE_1$ class:

$HAVE_1$ => N + *a* + N

(61) basisbåndudstyr => apparecchiature a banda base
'base band equipment'

Before trying to state a conclusion on the basis of these generalisations, however, let us consider the issue of computability.

3.3. Computation of the semantic classes.

Assuming for a moment the usefulness of computing the semantic relation the non-head plays in a non-argumental nominal compound, we must naturally ask ourselves how such a computation can be performed. A tempting hypothesis is that it can be done on the basis of the semantic features of the single nouns. To verify it, we coded all the nouns appearing as compound components in the corpus according to the DISEM system. DISEM is a system of semantic features developed at Eurotra-DK to solve the problems of lexical ambiguity and ambiguity of attachment (cf. Boje & Schøsler 1992). It basically consists of a semantic hierarchy tuned to the text domain represented by the Eurotra corpora. The hierarchy is used in such a way that nouns are coded with one of the terminal values, and semantic selectional restrictions are coded by using any of the values in the tree. A preference mechanism performs the computation needed for disambiguation on the basis of the best possible combinations of restrictions and terminal values.

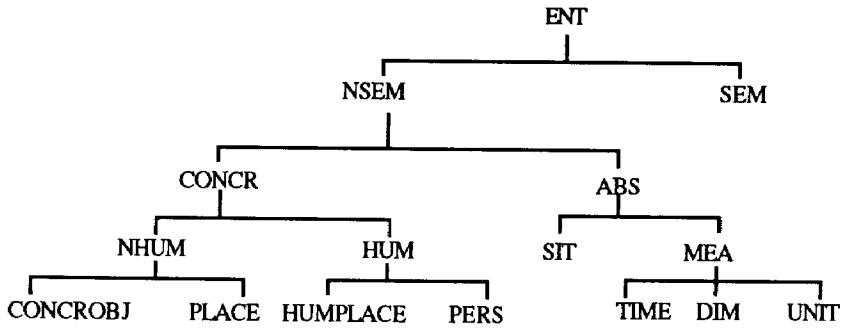

Figure 7. - *The DISEM system.*

In our case, the hierarchy had to serve a slightly different purpose as we wanted to combine the semantic descriptions of the two nouns[8] in a compound to compute the semantic relation played by the non-head.

First of all, we had to modify the hierarchy slightly to accommodate a few distinctions that were important to our purpose. The resulting tree

[8] Based on the arguments discussed in section 1.1.5, we assume that in the interface structure of an MT system, relational adjectives will be assigned noun category.

is shown in fig. 7. As for Levi's typology, we do not here make any claims about the universality of this semantic hierarchy. Our intent is to point at the fact that semantic analysis of compounds may be useful in an MT system, rather than proving the general validity of the actual semantic categories we have used. Therefore, we did not provide a detailed description of the various categories in Levi's model, nor do we intend to discuss the labels in the DISEM system (for a detailed account, the reader is referred to Boje & Schøsler 1992). However, a short list of examples belonging to the terminal categories relevant to our corpus is given below:

CONCROBJ(ject): *kanal* 'channel', *transistor* 'transistor'

PLACE: *kontor* 'office', *station* 'station', *laboratorium* 'laboratory'

PERS(on): *bruger* 'user', *gruppe* 'group'

SIT(uational): *aktivitet* 'activity', *frekvens* 'frequency', *teknologi* 'technology'

TIME: *tid* 'time'

DIM(ension): *hastighed* 'speed', *effekt* 'effect'

SEM(iotic): *program* 'programme'

On the basis of this semantic typing, we could write a number of rules, namely:

a) non-head is SIT(uational) and deverbal,
head is not PLACE & not TIME
=> semclass = FOR

(62) a. afprøvningssystem
'testing system'
b. behandlingsteknik
'treatment technique'

b) non-head is PLACE,
head is not PLACE & not TIME
=> semclass = FOR/FROM

(63) a. kontorsystem
'office system'
b. kontorprodukt
'office product'

c) non-head is DIM(ension),
head is CONCROBJ(ject)
=> semclass = $HAVE_1$

(64) a. lavhastighedskanal
'low-speed channel'
b. høj-effekt transponder
'high-effect transponder'

d) head is PLACE/TIME
=> semclass = IN

(65) a. industrisektor
'industrial sector'
b. leveringstid
'time of delivery'

If we look at these rules together with the translationally relevant generalisations made above, we can restrict the number of semantic classes worth computing to a much smaller set, namely:

a) FOR [where non-head is PLACE] => N + *per* + N

(63) b. kontorsystem => sistema per uffici

b) $HAVE_1$ => N + *a* + N

(64) a. lavhastighedskanal => canale a bassa velocità

To these, a default rule can be added:

c) default => N + *di* + N

(58) antennesystem => sistema di antenne

This may seem a rather disappointing conclusion compared with the ambitious hypothesis we started with.

However, when we look at the percentage of correctness we can achieve in the translation of our corpus by using either all of the three rules or just the default, the results become rather remarkable, as shown by the table in fig. 8.

Degree of accurateness	
Rules used	*Correct results*
Default	74.3 %
Default + a + b	88.8 %

Figure 8. - *Statistics of results.*

The percentage of correct results would be even higher than 88.8 % if we could identify all the examples belonging to the BE class, neither of which is to be translated by the default rule. Our failure in this respect can be due to one of two different factors: either the BE class is too heterogeneous or DISEM is not fine-grained enough to capture the relevant semantic facts. To conclude, the data show that, at least for a relatively homogeneous semantic domain, non-argumental nominal compounds are not as resistant to semantic analysis as is often claimed, and that, at least from a theoretical point of view, semantic analysis can help in establishing translationally relevant generalisations.

4. Conclusion.

Although our original objective was to conduct an empirical study on the translation of Danish nominal compounds into Italian, a clarification of the assumptions made during this study has led us to a discussion of a number of theoretical issues.

Firstly, the necessity for a clear, formal definition of what a nominal compound is, has shown that the traditional view of Danish compounds as typographical units is inadequate. Translational evidence as well as language specific considerations show in fact that a compound can also consist of a relational adjective followed by a noun. This insight, although generally accepted regarding English compounding, is fairly new as regards Danish grammar.

Furthermore, we have shown that if we accept to treat relational adjectives as compound components, this means calling into question a number of general principles often mentioned in relation to compounds. In particular, the principles governing the realisation of valency-bound arguments within compound forms seem, at least in Danish, less restrictive than often assumed. On the one hand, in fact, our data show that the non-head in a Danish compound can function as external as well as internal argument of the head, and on the other that internal arguments can be realised outside the first order projection of the head. This is not to say, however, that no restrictions exist: although we have not pursued the matter here, we have pointed at the fact that a

more precise account of the constraints at work must probably take the issue of reference into consideration.

Finally, on the basis of the empirical study mentioned above, we have discussed whether semantic analysis may provide a useful strategy for the translation of Danish non-argumental compounds into a Romance language like Italian, where the compound is often mapped onto a phrase. A rather trivial, but generally unmentioned fact emerging from the data is that non-argumental compounds are by far the most common type. Furthermore, a variety of syntactic and lexical realisations were found in the Italian translations. Therefore, any system that treats compounds in a compositional way must have a strategy for the translation of these cases. The conclusion we have drawn from the data is that, again contrary to the widely accepted position, semantic analysis seems to be both possible and useful. It remains to be proved, however, whether it would be cost-effective in a concrete implementation.

Other important issues related to the compositional treatment of compounds have not been discussed. Among these, the problem of how to generate the correct number and definiteness values of the non-head when the compound is translated with a phrase certainly demands further investigation. Although infallible rules are probably impossible to formulate in this respect, an analysis based on large data may, as in the case of the semantics of non-argumental compounds, reveal some useful tendencies.

Address of the Authors:

Patrizia Paggio
Center for Sprogteknologi
Njalsgade 80
DK-2300 Copenhagen S
e-mail: patrizia@cst.ku.dk

Bjarne Ørsnes
Institut for Almen og Anvendt Sprogvidenskab
University of Copenhagen
Njalsgade 80
DK-2300 Copenhagen S
e-mail: bjarne@cphling.dk

References

Aniadou, S. & J. McNaught (1990), "Treatment of compounds in a transfer-based machine translation", The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, University of Texas at Austin: 57-63.

Bauer, L. (1978), *The Grammar of Nominal Compounding*, Odense, Odense University Press.

Bauer, L. (1988), *Introducing Linguistic Morphology*, Bristol, Edinburgh University Press.

Bennett, P. (1993), "The interaction of syntax and morphology in machine translation", in Van Eynde, F., ed., *Linguistic Issues in MT*, Oxford: 72-104.

Bennett, P. & P. Paggio, eds. (1993), *Preference in EUROTRA*, Studies in Machine Translation and Natural Language Processing, Vol. 3, CEC, Luxembourg.

Boje F. & L. Schøsler, eds., (1992), *DISEM, A Semantic MT-Component*, Eurotra-DK Working Papers, Centre for Language Technology, Copenhagen.

Bouillon, P., K. Boesefeldt. & G. Russell, (1992), "Compound nouns in a unification-based MT systems", *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento: 209-215.

Copeland, C., J. Durand, J Krawer & B. Maegaard, eds. (1991a), *The Eurotra Linguistic Specifications*, Studies in Machine Translation and Natural Language Processing, Vol. 1, CEC, Luxembourg.

Copeland, C., J. Durand, J Krawer & B. Maegaard, eds. (1991b), *The Eurotra Formal Specifications*, Studies in Machine Translation and Natural Language Processing, Vol. 2, CEC, Luxembourg.

Di Sciullo, A.M. & E. Williams (1987), *On the Definition of Word*, Cambridge, Mass., MIT Press.

Finin, T.W. (1980), "The semantic interpretation of nominal compounds", *Proceedings of the First Annual Conference of the American Association for Artificial Intelligence*: 310-312.

Giorgi, A. & G. Longobardi (1989), "Typology and noun phrases", *Rivista di Linguistica* 1:115-161.

Hansen, A. (1967), *Moderne Dansk*, Vol. 2, København, Grafisk Forlag.

Jespersen, O. (1942), *A Modern English Grammar on Historical Principles*, Vol. 4, København, Munksgaard.

Levi, J.N. (1978), *The Syntax and Semantics of Complex Nominals*, New York, Academic Press.

Scalise, S. (1986), *Generative Morphology*, Studies in Generative Grammar, Dordrecht, Foris.

Selkirk, E. (1982), *The Syntax of Words*, Cambridge, Mass., MIT Press.

Warren, B. (1978), *Semantic Patterns of Noun-Noun Compounds*, Gothenburg Studies in English 41, Acta Universitatis Gothenburgensis, Gothenburg.