Treatment of Compounds in a Transfer-based Machine Translation System

S. ANANIADOU & J. MCNAUGHT

Centre for Computational Linguistics
UMIST, PO Box 88
Manchester, UK
effie%ccl.umist.ac.uk@cunyvm.cuny.edu

Contemporary work on compounds in linguistics and NLP is reviewed. Recent research is presented demonstrating how morphology theory is applicable to characterising and representing compounds in a transfer based MT system, and to aiding in the development of strategies, leading to successful handling of productive, compositionally translatable compounds.

1. Introduction

NLP systems have not yielded satisfactory, generally applicable solutions to the processing of compounds, which is a major stumbling block for systems aspiring to deal with real texts (Isabelle, 1984:509). Efforts to record lexicalised compounds are of little avail, as new compounds appear unceasingly in all areas of language use. Any NLP system must therefore be equipped to deal with new compounds formed according to regular, productive compounding processes.

The 9 language EUROTRA project has, over 3 years, been conducting research into the translation of compounds, in a transfer-based approach. Features of the research undertaken, reported on here, were that it concentrated on:

- determining how contemporary linguistic theory could contribute to the characterisation, representation and translation of compounds;
- determining productive, compositional compounds, monolingually;
- establishing mappings between such compounds in different languages, rejecting those where compositionality was not maintained in translation;
- evolving strategies for effecting translation of these compound types.

The results obtained are readily adaptable to models which have a stratificational linguistic framework and are able to emulate feature value percolation.

The research undertaken was reductionist in nature, and led to a set of compositionally translatable productive compound types being isolated. No claim is therefore made for wide coverage of the compounding phenomenon. However, the representation elaborated for compounds and the strategies developed for their analysis, translation and synthesis are pertinent to any approach to the problem area.

The research described deliberately left out of consideration special language (sublanguage) compounding, as there is little formal knowledge about term formation and as it is dangerous to attempt to analyse and translate compound terms on a compositional basis. Terms have accepted, often standardised or mandatory translations to allow unambiguous designation of the concept they refer to, thus compositional translation would in many cases at best yield an unacceptable paraphrase, at worst some misleading form corresponding to a different concept in the target language. EUROTRA is addressing the issue of terminology in a separate research programme. In addition, we shall not enter here into detail concerning the framework of the EUROTRA MT model. Such information is available in, e.g.. Varile & Lau (1988), Allegranza et al. (1990), Raw et al. (1988), Bech & Nygaard (1988).

2. State of the Art - Linguistics

Compounding, especially in English, has been studied extensively in linguistics. Syntactic and semantic criteria have been exploited in attempts to interpret compounds.

Those basing their work on a combination of syntactic and semantic criteria include Adams (1973) and Marchand (1969). Downing (1977) uses uniquely semantic criteria. Lees (1978), Allen (1978), Levi (1978) and Warren (1978) contain relevant discussions.

In the 1980s, important progress has been made in word formation (WF) studies, due to (where compounds are concerned) Lieber (1980, 1983), Selkirk (1982), Di Sciullo & Williams (1987), Paulissen & Zonneveld (1988) and Ralli (1988). Aronoff (1976), although concerned with a different approach to morphology to that of e.g. Selkirk, Lieber and Williams, deliberately excludes discussion of compounding phenomena. A point of considerable debate is whether morphology can be considered a separate component to syntax. There are extreme and compromise views, the pendulum of consensus vacillating over all. A recent contribution is the collection of papers edited by Everaert et al. (1988). Our preference, in a MT context, is for a view which endows morphology with its own rules, representations and principles, without however seeing it as a monolithic, modular block.

Compounding is particularly troublesome, lying in the grey area between morphology and syntax, thus apparently subject to two sets of operations. This is due to the existence in many languages of oneword compounds - clearly in the morphological domain - and multiword compounds (i.e. several text words), which are interpretable as being in the domain of syntax. However, both types fulfill the same function: a oneword compound and a multiword compound can both be interpreted as expressing one major lexical category, as each being a morphological object This is true also for compounds which involve a combination of major lexical categories and function words, e.g. Spanish and French N + Prep + N formations. Languages differ in the distribution of compound types. English, for example, demonstrates one word compounds, multiword compounds and hyphenated compounds. Cross-linguistically, there is rare identity of mapping between languages in orthographic terms: a oneword compound in English does not necessarily imply a mapping to a similar form in Greek.

3. State of the Art - NLP

MT systems have generally failed to tackle the problem of compounds in anything other than an ad hoc fashion. Honourable exceptions are the systems developed by the TAUM group (Isabelle, 1987). Well-known systems such as ARIANE (Guilbaud, 1987) or Mu (Tsujii, 1987) have no linguistically motivated accounts to offer. In monolingual NLP, Finin (1980, 1986), McDonald (1982), Sparck Jones (1983, 1985) and Hoeppner (1982) have concentrated on semantic, knowledge-based treatments. Each author comments on the difficulty of processing compounds. Finin (1986:165) gives a summary of the problems encountered with English nominal compounds even within a model taking account of discourse context in a sublanguage domain. Hoeppner notes, however, that analysis and interpretation of compounds depend as much on knowledge extracted from morphological or syntactic processing as on semantic or extra-linguistic knowledge, and emphasizes the crucial role of WF studies.

Earlier, EUROTRA had tried to set up contrastive mappings for compounds, based on descriptive work which ignored theories of compounding. The results were not satisfactory. Research was then started to:

Monolingually

- give guidance on which elements are possible compounds;
- state which elements are definitely not compounds;
- capture relationships within compounds;
- relate the treatment of compounds both to word grammar and grammar in general.

Multilingually/bilingually

- determine which classes of compound can be translated compositionally;
- give guidance on which classes of compound cannot be so translated;
- establish translational correspondences between classes of compound in different languages;
- establish criteria for determining translational correspondences between compounds and non-compounds and vice versa.

4. Research results

The research carried out in EUROTRA since 1988 has concentrated on compounds which are compositional, productive, translatable and

tractable. *Compositonality* and *Productivity* are associated with morphophonological regularity in WF and with semantic transparency. *Translatable* meant for us that mappings between languages could be reasonably established for EUROTRA purposes. *Tractability* is interpreted as meaning currently available EUROTRA mechanisms could be reasonably successfully employed. We did not investigate the interaction of compounding with derivational morphology, but work is currently under way in this area.

The theoretical linguistic framework we adopted draws on Lieber (1980, 1983), Selkirk (1982) and Di Sciullo & Williams (1987). In particular, the notion of *head* is crucial to our model. We examined endocentric compounds, which are headed objects. The location of the head is language dependent, thus English typically has right headed compounds, whereas Greek demonstrates both right and left headedness, depending on the compound type. The head plays a key role as it enables proper percolation of feature values to dominating nodes and furthermore enables proper assignment of relational information at other stages. Another crucial notion adopted from the literature is that of the *Principle of Syntactic Atomicity* (cf. Di Sciullo & Williams, 1987) which blocks rules of syntax from applying within compounds. Blocking is required particularly in the case of multiword compounds, objects which appear on the surface to be isomorphic to phrasal categories (i.e. to NPs for example). However a closer look at the internal structure of such objects reveals some basic differences between this kind of element and phrasal units. The internal structure of multiword compounds is distinguished from syntactic phrases by being opaque to the application of some basic phrasal syntactic rules - that is, these objects must be considered atomic with respect to syntax. Examples of such rules that do not apply within compounds are given by Di Sciullo & Williams (1987:49-52):

- words are generic in meaning as opposed to phrases, that is, they contain e.g. no reference to time;
- pronominal reference is not allowed in compounds.

There are other instances of syntactic rules that do not apply within compounds.

The strategies that were evolved in the course of the research reported on here apply the notion of syntactic atomicity to rendering compounds atomic at every representational level after their identification up to the EUROTRA IS level. IS is a deep syntactic dependency representation, consisting of governor-argument-modifier structures and featurised morphosyntactic information. It provides a canonical form where arguments are explicitly related to their predicates. It is not a semantic level. Lower levels in EUROTRA include a surface phrasal syntax level indicating syntactic functions and order of constituents (ECS) and a relational syntax level (ERS). ERS mainly manipulates frame information to build relational structures. Nodes at all levels take the form of feature bundles. Both ERS and IS levels are dependency based. As frames at ERS are syntactic frames, representations of compounds must again remain atomic with respect to ERS as frame inheritance and frame satisfaction within compounds do not follow the same principles as in regular syntactic formations. In generation, compounds must also remain atomic at the higher levels, as otherwise paraphrases instead of compounds will be produced.

We do not however adhere globally to strong atomicity: the strength of atomicity varies depending on the language involved. Thus, for Greek, it is necessary on occasion for syntax to look inside an 'atomic' compound as there are rules relating to cliticization and agreement between elements of compounds that must apply. However, only these specific rules of syntax will be allowed to apply and no other. An example case is: 'διαστημικό λεωφορείο' - 'space ship' versus 'τα διαστημικά τους λεωφορεία' - 'their space ships'.

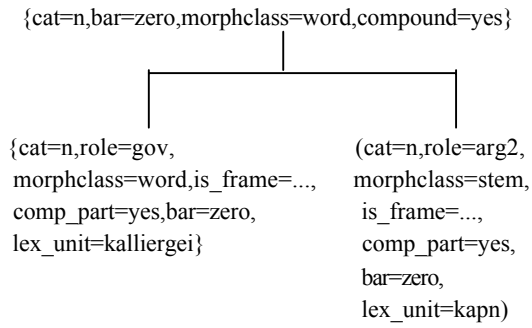
It is important that atomicity should be distinguished from lexicalisation. Lexicalisation takes place only if forms are non-compositional, compositional but non-productive, or compositional but not tractably translatable. It is a *destructive* operation in the sense that no substructure is preserved. Lexicalisation can take place at different levels. A compound could be compositional at a lower level, but lose this characteristic at a higher level, becoming non-compositional with respect to that level, hence necessitating lexicalisation. For example, 'fruitcake' (crazy person) might be analysed

compositionally at lower levels, becoming lexicalised at e.g. the IS level. If a structured object is however rendered atomic, this does not involve any destruction of substructure. Substructure can therefore be carried through a particular level, without other rules 'seeing' inside that object.

For EUROTRA purposes, our proposal for treating compounds at the highest (deep syntactic) level of IS is that:

- Representations for compounds at IS should be structured in the sense that both oneword and multiword compounds should be represented in a hierarchical structure.
- The dominating node of this structure should contain markers (feature values) identifying the whole structure as a compound, i.e. as an X_0 object in X terms. It will thus be rendered atomic at that level.
- Daughter nodes are marked with feature values indicating their special status as parts of a compound. They may furthermore be considered compounds in their own right, if a multi-element compound has been recursively generated, and will then require appropriate marking.
- Nodes are marked with the morphological class of the object they refer to: in our case typically {stem, word}.
- Relevant frame information must appear on the mother node and daughters.
- In addition, information detailing preposition type involved and case values will be necessary, depending on the language involved.
- Argument roles will have to be present {governor, argument, modifier} - the governor indicates the head of a dependency construction.
- Morphosyntactic classes will have to be marked (noun, adjective, etc. - depending on the productive compound types of the language).

Given such information, transfer can translate compounds correctly according to its rules mapping compound types of one language to those of the other. Below is an example of an IS representation for the Greek compound 'καπνοκαλλιέργεια' whose English translation is 'tobacco cultivation'. This is a somewhat idealised and simplified representation in the interests of illustration. Irrelevant detail is omitted and there is also some redundancy again in the interests of exposition. The values of the feature 'role', which yield the basic predicate-argument structure of IS, are here 'gov' (governor) or 'arg2' (deep syntactic object). The feature 'comp_part' indicates whether the element is a part of a compound or not. This representation will be subject to simple transfer by the default transfer rule, which effectively results in translation of lexical units, thus avoiding the need to write specific transfer rules in this case. Note that the Greek linking vowel '-o-' has been dealt with at an earlier stage. Research has shown that indeed many cases of, as previously supposed, complex transfer involving compounds can be successfully handled through simple transfer if compounds receive a structured representation of this type at IS.



The EUROTRA concept of level of representation, where legal structures are specified by an associated generator, implies that (in, say, analysis), each generator is partitioned into at least 2 sets of rules: one dealing with compounding (CRs), the other with non-compounding phenomena (NCRs). Each generator (in the general case) receives its input from a transducer which maps the structures admitted by the language of the generator of the previous level onto the language admitted by the next. Each generator must consolidate incoming data (i.e. build its own structures over the data and validate them) before engaging in further manipulation. Thus when a generator is invoked, CRs must consolidate substructure representing compounds, and mark these as atomic with respect to NCRs. The NCRs must therefore consolidate structures which also include atomic objects representing compounds but which

nevertheless will preserve the morphological substructure associated with compounds. In other words, this implies that there are morphological structures (representations of compounds) that are built early in analysis and that are carried through several levels of representation unaffected, until the deepest level is reached. Technically, CRs are indeed members of the rule-set of a generator and must therefore be seen as generating the legal structures of that generator. Thus for a phrasal syntax generator, CRs belong to the set of phrasal syntax rules even though their task is to construct structured objects that are rendered atomic or opaque with respect to the 'main' NCRs.

Morphology, for us, is thus not confined to a component which applies early in analysis and late in synthesis. Morphological information is, in our view, distributed over all levels. At the deepest one, representations contain morphological objects (compounds). The advantages for transfer are:

- productive, compositionally translatable compounds can be translated via general rules dealing with correspondences between patterns in each language;
- as compounds preserve their morphological structure throughout, paraphrase translations are not generated;
- overgeneration in translation is avoided.

For English, Selkirk (1982) and Lieber (1983) aided us in arriving at our monolingual classification. This work indicates that endocentric, verbal compounds (where there is a deverbal head) are more tractable than other types, in that roles such as Agent, Instrument, etc. can be used and e.g. Lieber's Argument Linking Principle can be applied. Our work on English was parallelled by work by other EUROTRA groups on their own languages.

For transfer purposes, mappings between compound classes were identified in the course of comparative research with other EUROTRA groups. It is to be noted that not all language pairs were covered in this initial work, however these partial results have proved promising. Transfer mappings to English that were identified as being translationally compositional included:

Greek to English

1. $N_{nom} + N_{genitive} \rightarrow N+N$

   ζώνη ασφαλείας → safety zone
   [head] [modifier] [head]

2. $N_{nom} + N_{nom} \rightarrow N + N$

   ποαδι θαύμα → wonder kid

3. one word compounds $N + N \rightarrow N + N$

   καπνοκαλλιέργεια → tobacco cultivation

Italian to Greek and English

1. $N + N \rightarrow N + N$

   uomo scimmia → ape man
   → πιθηκανθρωπος
   (πιθηκ 'monkey/ape' + άνθρωπος 'man')

2. $V + N \rightarrow N + N$

   apriscatole → can opener
   spazzacamino → chimney sweeper

5. Problems

It should be emphasized that work has concentrated on classifying compound types, monolingually and cross-linguistically, on specifying representations adequate for transfer and on evolving strategies to compute these representations. We have not as yet addressed other areas such as:

- how to establish the scope of compounds (the sequence A + N in English for example is notorious);
- how to determine the correct segmentations (for oneword forms) and parses for compounds;
- how to know when to synthesize a oneword compound in a language which admits oneword and multiword forms.

Resolution of the first and second points above demands greater knowledge than at present available. We can achieve a set of segmentations and a set of parses for a compound, and can constrain our analysis further by use of e.g. argument Unking, however it is extremely difficult to know, in the absence of greater knowledge, which of the remaining parses is the

correct one. Compounds in English consisting of strings of nouns are a case in point, demanding semantic and real world knowledge for their interpretation.

These and other other areas demand investigation. However, the research reported on here was more concerned with establishing theoretical principles for analysing, representing and synthesising compounds, no previous such work having been undertaken in EUROTRA. We therefore operated largely with the assumption that we were working with a certain type of object, rather than addressing e.g. the problem of how to identify such an object in the first place in running text. However, with a theoretically motivated foundation, others can now proceed to look at the other areas mentioned above.

6. Conclusion

The application of recent results from linguistics, together with monolingual and comparative research aimed at determining productive, compositionally translatable compound types and appropriate mappings between languages, has led to the development of representations for compounds that are adequate for transfer and of strategies for arriving at such representations, for effecting transfer and for synthesising target language compounds. Whereas compounds taken as a whole have until now proved largely intractable in MT in not admitting theoretically motivated generalisable solutions, the results obtained in this research show that certain classes of compounds can indeed be handled (and moreover by simple transfer): those that are productive and compositionally translatable. This offers at least a partial solution grounded on a theoretical basis which nevertheless covers large numbers of instances of compounds.

The work reported on here is addressed in greater detail in Ananiadou (forthcoming-a) and Ananiadou (forthcoming-b).

Acknowledgements and Disclaimer

The research reported on here was undertaken in the framework of the EUROTRA Machine Translation Project, co-sponsored by the UK Department of Trade and Industry/Information Engineering Directorate and the Commission of the European Communities. We are grateful to A. Ralli (EUROTRA-Greece), who participated in this research, for data relating to Greek and discussions with her in working groups. The views of the authors of this paper are not necessarily those of the EUROTRA Project Management.

References

Adams, V. (1973) *An Introduction to Modern English Word Formation.* Longman, London.

Allegranza, V., Bennett, P.A., Durand, J., van Eynde, F., Humphreys, L. Schmidt, P. & Steiner, E. (1990) *Linguistics for MT: The Eurotra Linguistics Specifications.* Eurotra Papers in Machine Translation and Natural Language Processing. Volume 1. DGXIII, CEC, Luxembourg.

Allen, M.R. (1978) Morphological Investigations. PhD. Dissertation, University of Connecticut.

Ananiadou, S. (forthcoming-a) *Automatic Term Recognition.* Edinburgh University Press, Edinburgh.

Ananiadou, S. (forthcoming-b) Treatment of compounds in EUROTRA. In Ananiadou, S. & Verheul, C. (forthcoming) *Morphology in EUROTRA*. Eurotra Papers in Machine Translation and Natural Language Processing. Volume 3. DGXIII, CEC, Luxembourg.

Aronoff, M. (1976) *Word Formation in Generative Grammar.* MIT Press, Cambridge, Mass.

Bech, A. & Nygaard, A. (1988) The E-framework: a formalism for natural language processing. Proceedings of COLING-88, 36-39.

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

Downing, P. (1977) On the creation and use of English compound nouns. *Language* 53:4, 810-842.

Everaert, M., Everts, A., Huybregts, R. & Tromelen, M. (eds) (1988) *Morphology and Modularity.* Foris, Dordrecht

Finin, T. (1980) The Semantic Interpretation of Compound Nominals. Technical Report T-96, Coordinated Science Lab. Univ. of Illinois.

Finin, T.W. (1986) Constraining the interpretation of nominal compounds in a limited context In Grishman, R. & Kittredge, R. (1986)

Analyzing Language in Restricted Domains. Lawrence Erlbaum, Hillsdale, NJ. 163-174.

Guilbaud, J-Ph. (1987) Principles and results of a German to French MT system at Grenoble University (GETA). In King, M. (ed) (1987), 278-318.

Hoeppner, W. (1982) A multilayered approach to the handling of word formation. Proceedings of COLING-82, 133-138.

Isabelle, P. (1984) Another look at nominal compounds. Proceedings of COLING- 84, 509-516.

Isabelle, P. (1987) Machine Translation at the TAUM Group. In King, M. (ed) (1987), 247-277.

King, M. (ed) (1987) *Machine Translation Today.* Edinburgh University Press.

Lees, R. (1963) *The Grammar of English Nominalizations.* Mouton, The Hague.

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

Lieber, R. (1980) On the Organization of the Lexicon. PhD. Dissertation, MIT.

Lieber, R. (1983) Argument linking and compounds in English. *Linguistic Inquiry* 14, 251-285.

Marchand, H. (1969) *The Categories and Types of Present-Day English Word Formation. 2nd Edition.* C.H. Beck'sche Verlags-buchhandlung, Munich.

McDonald, D.B. (1982) Understanding Noun Compounds. Report CMU-CS-82-102. Department of Computer Science, Carnegie-Mellon University.

Paulissen, D. & Zonneveld, W. (1988) Compound verbs and the adequacy of Lexical Morphology. In Everaert, M. et al. (eds) (1988), 281-302.

Ralli, A. (1988) Verbal compounds in Modem Greek. Proceedings of the 9th Meeting of the Department of Linguistics, Thessaloniki (in Greek).

Raw, A., Vandecapelle, B. & van Eynde, F. (1988) Eurotra: an overview. *Interface* 3:1, 3-32.

Rich, E., Barnett, J., Wittenburg, K. & Wroblewski, D. (1987) Ambiguity Pro-crastination. Proceedings of AAAI-87, 571-576.

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

Sparck Jones, K. (1983) So what about parsing compound nouns? In Sparck Jones, K. & Wilks, Y.A. (eds) (1983) *Automatic Natural Language Parsing.* Ellis Horwood, Chichester. 164-168.

Sparck Jones, K. (1985) Compound noun interpretation problems. In Fallside, F. & Woods, W.A. (eds) (1985) *Computer Speech Processing.* Prentice-Hall International, London. 363-381.

Tsujii, J-I. (1987) The current stage of the Mu-Project, Proceedings of the Machine Translation Summit, September 1987. JEIDA, Tokyo. 122-127.

Varile, G.B. & Lau, P. (1988) EUROTRA: Practical experience with a multilingual machine translation system under development. Proceedings of the Second ACL Conference on Applied Natural Language Processing, 160-167.

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