Pius ten Hacken, Universität Basel, Institut für Informatik

On the Definition of Compounding

ABSTRACT: In this paper, a language-independent definition of compounding is proposed, based on the relation between the elements of a compound and the behaviour of compounds in semantics and discourse. From the definition, language-independent tests are derived, for the recognition of compounding constructions in individual languages. They are applied to a number of obvious and less clear-cut examples. The definition and the tests are intended to be part of a taxonomic system for linguistic expressions, supporting the lexicographer in his judgements.

1. Compounding and the Dictionary

The purpose of a dictionary is to list lexical units and convey information about them as required by its users. A lexical unit is an item that cannot be analyzed compositionally, i.e. the meaning and form of a lexical unit cannot be predicted on the basis of rules. Compounds are combinations of two elements, created by a productive process. The productivity of the process prevents an exhaustive listing. In a Natural Language Processing system, exhaustive coverage is required, and it can only be achieved by a system of rules. Compounding is also a source of new lexical units. When a compound acquires a meaning that cannot be predicted on the basis of rules and the meaning of their component parts, it becomes a lexical unit, to be listed in the dictionary.

In this paper, I will propose a definition of compounding, characterizing it as a phenomenon that can be realized in various constructions. Tests derived from the definition determine whether a particular construction belongs to compounding, to other parts of morphology, or to syntax, and whether a particular item is a lexical unit or a regular compound. The definition is language-independent, because otherwise decisions which constructions belong to compounding are just arbitrary postulates.

The requirement of language-independence implies a certain level of abstraction of the criteria that can be used as part of a definition. The most widespread criterion in the literature is based on stress. In a structure $[X\ Y]_Z$, a single primary stress on X would indicate that Z is a compound. As noted by Bloomfield (1933) already, in English numerous problematic examples exist, where an item has a single-stressed variant alongside a double-stressed one, e.g. *ice cream*. In a language like French, the criterion makes no sense at all, because stress rules in French have a different nature.

Language-independent conditions cannot be imposed on the form of the elements of a compound, because their realization is determined to a large extent by idiosyncratic properties of a language, or even of a particular construction within the language. Since our question is whether a given construction is an instance of compounding, we do not make much headway by determining these formal idiosyncrasies. Instead, we will look at the relation between the elements of a compound, and the restriction the compound status entails for their syntactic and semantic behaviour in a sentence.

2. The Relation between the Elements of a Compound

The main classes of compounds distinguished by Sanskrit grammarians are compounds where the sense of the last element is the main one, compounds where the sense of a different word is the main one, and compounds where neither part is subordinate to the other (Ballantyne 1849). English examples of these types are *bookshop, wetback, bittersweet.* Bloomfield (1933) calls them determinative, exocentric, and copulative compounds, respectively. If we are to use the relation between the elements in the definition of compounding, we have to use three definitions. The three classes are too far apart for a generalization over them to contain any substantial restrictions on class membership. Although accepting that the latter two classes are productive, contrary to what has been claimed by e.g. Lees (1960) for copulative and Allen (1978) for exocentric compounds, I will concentrate on determinative compounding here.

Determinative compounds consist of a head and a modifier. Syntactically, the head can be recognized because it shares with the compound itself a number of features, including syntactic category, gender for nouns, and inflectional categories like number. In this respect, two tendencies can be observed in recent linguistic literature, the ever stricter interpretation of headedness as right-headedness, and its progressive extension in scope. They culminate in Di Sciullo & Williams (1987) taking their Righthand Head Rule as *the* language-independent defining criterion for morphological objects. In Italian *nave passegeri* ('passenger ship'), however, *nave* is feminine and singular like the compound, whereas *passegeri* is masculine and plural. Rejecting it as a compound, while accepting as such its English counterpart, is as artificial a distinction as analyzing *ice cream* as ambiguous between a compound and a synonymous phrase because of stress.

A rigorous characterization of the semantics of the relation between the head and the modifier of a compound has been attempted by Levi (1978). She distinguishes two sources for the predicate expressing the relations, either it is contained in the head, or it is taken from a small set of predicates associated with the phenomenon of compounding. As she works within a generative semantic framework, a head may contain a predicate either overtly, e.g. *truck driver*, or invisibly at the surface, e.g. *steal* in *car thief*. Problems for her theory arise especially when the head does not contain a predicate. The characterization of the relation by a fixed set of predicates is more precise than the vacuous *related to*, but still it makes an impression of a ready-to-wear suit, when fitting the relation in *mountain range* into *make (passive)*. Moreover, a sizeable ambiguity of analysis is created (12-fold), not always corresponding to ambiguity of meaning. Thus, *party members* is analyzed as ambiguous between a *have (passive)* and *in* relation, among others. All these problems cannot be confined to compounds where the head does not contain a predicate, as illustrated by *pressure cooker*.

From the relative success of the components of Levi's theory, it can be concluded that a characterization of the relation in a compound on the basis of the head is to be preferred over an approach where this relation is linked to the phenomenon of compounding. It can be extended to heads not containing a predicate, if each head is associated with a number of relations, instead of a single one. Which of the relations applies depends on the modifier. Thus, *mill* permits different relations in *windmill* and *coffee mill*. As observed by Allen (1978), the relations associated with a particular head are often hierarchically organized. Thus, Dutch *fabriek* ('factory'), is usually modified by an indication of what is

produced, e.g. *autofabriek* ('car factory'), but for *vrouwenfabriek* (lit. 'women factory'), world knowledge demands a different interpretation, e.g. 'factory employing only women'.

3. Compounds as Units

Syntactically, compounds tend to behave as a closed unit, in the sense that it is in general not possible to separate the elements or to modify a single element of a compound by means of a word not belonging to the compound. However, coordination and a number of other low-level syntactic processes can operate on two compounds with the same head, as in *love and horror stories*. As to modification, it is sometimes difficult to assess without further criteria, whether an element is part of a compound or not, cf. *open-air museum*. Therefore, it is difficult to formulate rigid conditions on cohesiveness of a compound and non-modifiability of its elements.

Postal (1969) claims that pronominal reference to an element in a compound is impossible. At least as far as reference to the non-head is concerned, the resulting sentences are at best language puns, e.g. **Harry was looking for a book$_i$ rack, but he only found racks for very small ones$_i$*. If the non-head is a common noun, it is generic (cf. Levi 1978). From these properties, the following generalization can be deduced: The non-head of a compound is not eligible independently for semantic or high-level syntactic processes. It is only visible to low-level processes like coordination.

Proper nouns as non-heads of compounds seem to be a counterexample. They are obviously not generic, and pronominal reference to them is possible, e.g. *Some Haydn$_i$ symphonies are very much alike. Apparently, he$_i$ sometimes lacked inspiration*. First names on their own, however, are impossible altogether, as in **Bill admirer*. On the basis of this supplementary evidence, the generalization can be reformulated as follows: The non-head of a compound does not interact with discourse to get an interpretation. Either it has a fully specified interpretation on its own (proper nouns), or its reference remains underspecified (generic). First names are usually underspecified, but cannot be generic, hence they cannot be non-heads of compounds at all.

4. An Operational Definition

From the discussion of the preceding sections, the following definition of rule-governed (determinative) compounding can be deduced:

A compound is a structure $[X\ Y]_Z$ or $[Y\ X]_Z$, such that:

- The reference of Z is a subset of the reference of Y;
- If S is a possible way of specifying Y, the reference of Z is determined by the range of S's that are compatible with the semantics of X;
- X does not have independent access to the discourse.

The first condition establishes headedness, without fixing its direction. The second condition characterizes the relation between the head and the non-head. It diverges from the Variable R Condition formulated by Allen (1978) in the following respects: it does not presuppose right-headedness; the formalization used by Allen introduces several unex-

plained terms, so that it is less precise than the one given here; and Allen does not use her condition as a defining criterion. The third condition is the conclusion reached in section 3.

From this definition of the phenomenon of compounding, tests can be derived to identify language-specific compounding constructions. First, I will show how the definition excludes a number of obvious non-compounds, a pre-condition for the validity of the definition. In the next section, a number of less clear cases will be considered. The first test is a direct mapping of the structure and the first two conditions:

> *Structure Mapping.*
> If Z is the alleged compound, impose a structure [X Y]Z or [Y X]Z, so that Z is a (kind of) Y, related to X in any of several ways. If it is not possible, Z is not a compound.

The structure mapping test presupposes two elements, with the meaning of a stem. Affixation is excluded as illustrated by *requirement* (not a *ment*) and *ex-president* (not a *president* related to *ex*). The headedness requirement excludes exocentric and copulative compounds (cf. *wetback, northwest*), and, since headedness in syntax has a different meaning, most syntactic combinations (cf. *John disappeared, for John, the table*). For determinative compounds having internally structured elements, the test allows to determine the structure, establishing *[[concert hall] director]* and *[gas [cigarette lighter]]* as more plausible analyses than alternatives involving a *hall director* and *gas cigarettes*.

Two classes of problems remain after the application of the structure mapping test. On the one hand, some syntactic constructions are not excluded. *John's book* exhibits the range of possible relations typical for compounds, 'the book John wrote/owns/published etc.', but we would not like to call it a compound. On the other hand, the boundary between compounding and the lexicon has not been marked very clearly. The phrase 'in any of several ways' in the test is meant to render the range of possible relations, but in case of doubt it offers little support. For these problem cases, two new tests will be introduced, the pronominal reference test and the coordination test.

In the case of *John's book*, the solution to the problem is simple. It cannot be a compound, because *John*, as a first name, cannot occur in non-head position of a compound. In other cases, we need the pronominal reference test, that can be formulated as follows:

> *Pronominal reference.*
> Construct a discourse with the alleged compound Z in one sentence, and a pronoun unambiguously referring to the non-head of Z in the next sentence. If the discourse is correct, Z is not a compound.

An application of the pronominal reference test is *John was not satisfied with his children$_i$'s school. He thought they$_i$ did not learn enough.* This discourse shows that *(his) children's school* is not a compound.

The problem of the demarcation of the boundary between regular and lexicalized compounds is illustrated by *banana republic*. Although it has obviously been lexicalized, the result of the structure mapping test, *A banana republic is a (kind of) republic, related to bananas in any of several ways*, only yields a vague suspicion that *in any of several ways* is inappropriate. Exploiting some properties of coordination, the following test is meant to solve this problem:

Coordination.
If [X Y]Z or [Y X]Z is the alleged compound, and Y is the head, construct a phrase where X is coordinated with an X', such that X and X' refer to disjoint sets of referents, and have the same hyperonym. If the phrase is not acceptable, Z is not a compound.

A test sentence proving that *banana republic* is lexicalized, is **Central America is full of banana and orange republics.* Because of the similarity between *banana* and *orange*, the difference between *banana republic* and *orange republic*, underlying the impossibility of the conjunction, must be explained by one of them having been lexicalized.

5. Application of the Tests

So far, we have only looked at straightforward cases of compounds and non-compounds, in order to show how the definition yields the correct answers when applied to them. In this section, some less obvious cases will be discussed. As a heuristic principle, we use the hypothesis that if something is a compound, its translation in another language might well be a compound, too. In view of the definition, caution is due in two respects. First, the translation has to allow the same range of relations between the two elements, and not only render the most common one. Second, the translation has to behave as a unit on a par with the compound it renders.

The most natural English translation of Dutch *vrouwenfabriek* mentioned above, is *women's factory*. World knowledge eliminates the interpretation where *women* are the product, but many possible relations remain: that they are the employees, or the owners, or that the factory's products are for women, etc. Although similar in form to *John's book* and *his children's school* mentioned above, *women's factory* behaves as a compound with respect to the pronominal reference test, e.g. *John owns a women$_i$'s factory. *They$_i$ receive bad payment.* This implies that there are two constructions with a genitive morpheme, one of them a compounding construction. They have a different constituent structure, [N's N]$_N$ for the compound, [NP's N]$_{NP}$ for the non-compound. The difference can be observed in the behaviour with respect to determiners and adjectives. In [NP's N]$_{NP}$, the genitive NP is the determiner of the NP it is part of, cf. **a John's book.* The genitive NP may itself contain a determiner, but it need not agree with the N modified by the genitive NP, cf. *[these children's] school* vs. *a [women's factory].* The structure [N's N] is cohesive in the sense that an adjective will precede it, rather than separating the two nouns, e.g. *a new women's factory.* In [NP's N], the adjective will be inserted between the determiner (NP's) and the head noun, as usual for NPs, e.g. *his children's new school.*

It is a well-known fact from translation practice, that it is hardly possible to translate a relational adjective (RA), e.g. *developmental*, in isolation, if the target language does not have a corresponding RA. Thus, the best Dutch translation of *developmental problems* is the compound *ontwikkelingsproblemen.* In an obviously non-lexicalized example like *oceanic civilization*, the range of possible relations between *ocean* and *civilization* is characteristic of compounds. *Ocean* may refer to the place, the principal god, the source of wealth, etc., of the civilization. The pronominal reference test seems to support a compound analysis, as in **The ocean$_i$ic civilization flourished as long as it$_i$ provided enough fish.* However, the adjectival status of *oceanic* is sufficient to explain this result, cf. **She had always been joy$_i$ful,*

but it$_i$ disappeared when her business collapsed. Similarly, *ocean* in *oceanic civilization* is generic, but so is *joy* in *joyful*. Levi (1978) has shown that RA's can be coordinated with nouns in non-head position of compounds, cf. *oceanic and forest civilizations*, but not with common adjectives, e.g. **oceanic and impressive civilizations*. Proper names are allowed as a basis for RA's, e.g. *Italian*, but first names only as far as they do not need interaction with the discourse to get a reference, e.g. *Elizabethan*. Thus, RA + N combinations share many properties with N + N compounds. Although some of these properties can also be explained independently, none of them contradicts compoundhood, and the remaining evidence is sufficient to conclude that RA + N combinations are compounds.

The next question to be answered is how to distinghuish RA's from common adjectives, such as *old*. Often cited tests are modification by *very* (*very old* vs. **very oceanic*), and predicative use. According to Levi (1978), the latter is possible for RA's to some extent, but not in contexts like a *civilization which is old / *oceanic*. Still, these tests only exclude part of the common adjectives. Problem cases include *alleged fraud*, and *absolute nonsense*. The adjectives do not allow grading or predicative use, like RA's, but they are not related to nouns in a similar way. Therefore, they fail the structure mapping test, even in its most permissive form, where *urban* in *urban practice* is replaced by *city* in the test. A practical test to draw the borderline correctly is deriving adverbs or nouns from the adjectives. This is not possible for RA's, e.g. **oceanically civilized, *urbanly practical*, but unproblematic for other adjectives, e.g. *allegedly fraudulous, absolutely nonsensical*.

The differences between RA's and common adjectives can be reflected in the dictionary entries for them, by describing them as RA variants of the corresponding nouns. Many dictionaries define them with a formula like "of or relating to", or spell out some of the most common relations in the definition. Thus, the Collins English Dictionary gives as the first two readings of *lunar*: "1. Of or relating to the moon. 2. Occurring on, used on, or designed to land on the surface of the moon." In the light of the preceding discussion, it is preferable to replace these definitions by the description "Relational adjective of *moon*".

In the beginning of this section, it has been stressed that translation can only be a heuristic principle, not a criterion for the identification of compounds. This is illustrated by the following example. The Dutch translation of *stone wall* is usually assumed to be *stenen muur*. However, contrary to *stone wall*, *stenen muur* cannot be used to refer to the walls built in Switzerland to protect roads from falling stones, independent of the material the wall consists of. Rather, *stenen* has a single meaning, 'consisting of stone'. The conclusion that *stenen muur* is not a compound, as opposed to *stone wall*, is supported by coordination: *stone and snow wall* vs. **stenen en sneeuwmuur*.

6. Conclusion

In this paper, the following definition of determinative compounding has been given:

A compound is a structure $[X\ Y]_Z$ or $[Y\ X]_Z$, such that:

- The reference of Z is a subset of the reference of Y;
- If S is a possible way of specifying Y, the reference of Z is determined by the range of S's that are compatible with the semantics of X;
- X does not have independent access to the discourse.

From this definition three language-independent tests have been derived, the structure mapping test, the pronominal reference test, and the coordination test. They serve to identify individual compounding constructions in a non-arbitrary way. For the identification of instances of a compounding construction, it has been shown in some examples how construction-specific properties can be used to formulate additional tests. Consistency in the judgement of borderline cases between rule-governed compounding and lexicalized compounds can be achieved by the application of the coordination test.

The definition is intended to be part of a taxonomic system of definitions distinguishing linguistic expressions in various classes and delimiting these classes intensionally. This taxonomic system will be used in Word Manager, an NLP-system for the specification, use, and maintenance of morphological dictionaries (see Domenig & ten Hacken 1992). It can also be used in more traditional applications of lexicography, increasing consistency by its support of the lexicographer's intuitive judgements. In practical use, the number of tests required to classify a single item will be small. Even for borderline cases, the lexicographer will rarely have to apply more than two tests.

Bibliography

Allen, Margaret (1978), *Morphological Investigations*, Unpublished Ph.D. Dissertation, University of Connecticut.

Ballantyne, James (1849), *The Laghukaumudi of Varadaraja*, Motilal Banarsidass, Delhi (reprint 1967).

Bloomfield, Leonard (1933), *Language*, George Allen & Unwin, London etc.

Domenig, Marc & ten Hacken, Pius (1992), *Word Manager: A System for Morphological Dictionaries*, Olms, Hildesheim.

Lees, Robert (1960), *The Grammar of English Nominalizations*, Mouton, Den Haag.

Levi, Judith (1978), *The syntax and semantics of complex nominals*, Academic Press, New York.

Postal, Paul (1969) 'Anaphoric Islands', in Binnick et al. *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*, pp. 205-239.

Di Sciullo, Anna Maria & Williams, Edwin (1987), *On the Definition of Word*, MIT Press, Cambridge (Mass.) etc.

KEYWORDS: compounding, taxonomy of linguistic expressions, relational adjectives