Where do "Soccer Moms" Come From? :

Cognitive Constraints on Noun-Noun Compounding in English

Mark T. Keane & Fintan Costello

Dept. of Computer Science,
University of Dublin, Trinity College,
Dublin 2, IRELAND
mark.keane, fintan.costello@cs.tcd.ie

Abstract

Every year new noun-noun combinations enter the English language and become common parlance; compounds like "notebook computer" and "soccer mom". But, why is one pair of words chosen rather than another pair ? For example,why do we not use "patio-tile computer" and "sports mother" ? Clearly, many factors influence the process. We concentrate on the cognitive factor of informativeness; namely, that a novel combination should convey its meaning unambiguously. Costello & Keane (1996) have shown that some classes of concept promote ambiguity (or polysemy) in novel noun-noun compounds; artifact and superordinate terms promote polysemy whereas natural-kind and basic-level terms do not. Here we show that the topology of these conceptual classes in a large corpus of familiar compounds indicates that they constrain the compounds that appear in a language.

Introduction

Each year new nominal compounds enter the English language and extend the everyday vocabulary we have to describe our world; recent additions being terms like *notebook computer* (a small portable computer) and *soccer mom* (a type of voter courted by both candidates during the US presidential campaign[1]; see Tulloch, 1992, for other examples). Indeed, Cannon (1987) reports that up to 55% of entries in a corpus of new words are compounds of existing words. But, what factors govern the appearance of such compounds, why is one combination of words chosen over another and what facilitates the acceptance and continued use of a compound by a language community ?

Perhaps the most fundamental constraint is the need to refer to some new object or aspect of the world. If a new consumer product is created, like a palmtop computer, then people need a name for it. Aside from this basic constraint, there must also be a whole host of cultural, social, linguistic and cognitive factors that play a role in determining exactly which combinations of words are used. In our examination of noun-noun compounds, we concentrate on one particularly important factor; namely, the informativeness of the combination (e.g., Grice, 1975). A new compound should convey its meaning unambiguously in a wide range of contexts. As we shall see, meeting this requirement may be harder than it first seems because a novel noun-noun phrase will often evoke several alternative interpretations, each of which combine the phrase's constituent concepts in different ways (see Costello & Keane, 1996).

In this paper, we examine the novel hypothesis that the appearance of compounds in a language is constrained by the tendency of specific classes of nouns (e.g., natural kinds and artifacts) to combine unambiguously. We test this hypothesis by collecting a large corpus of commonly-used, familiar noun-noun compounds and determining the topology of this set of compounds with respect to specific conceptual categories. In the following sections, we present some previous views on the phenonemon and present the bases for the hypothesis examined in our empirical study.

On the Growth of Language Through Compounding

Several cognitive and linguistic factors have been implicated in the appearance of concept combinations in language. For instance, Bauer (1983) has proposed some etymological and phonological constraints on the production of compound phrases, observing that compounding adjectives are mainly monosyllabic words of Germanic origin. Others have suggested that semantics play a role, with compounds being created only if their constituents have a generic, habitual or permanent relationship (e.g., Downing, 1977; Gleitman & Gleitman,

[1] *Soccer mom* was recently voted "word of the year" by the American Dialect Society.

1970). To quote the Gleitmans (1970:96) "not every man who removes the garbage is a garbage-man. Only a man who occupationally, customarily, eternally removes the garbage is a garbage-man". Downing (1977) also notes that compounds are less likely to be acceptable if they describe a category which already has an existing name; for instance, one subject responded to the compound *church janitor* with "Might as well be *factory janitor*. Can't have a new name for everything a janitor cleans" (Downing, 1977:837). Finally, Ryder (1994) suggests that compounds are often created by analogy either to existing compounds (as in the sequence *watergate*, *iran-gate*, *S&L-gate*) or to the use of a particular word in a group of compounds (as in compounds of the form "sea-X", such as *seahorse*, *sealion*, or *seacow*; see also Shoben, 1993).

We believe that all of these proposals contain some truth, but that they are not the whole story. For example, the compound *cameo appearance* is not monosyllabic or germanic, does not describe a permanent relation and is not derived by analogy to some other existing phrase. In the next subsection, we outline a different perspective in which we propose that the appearance of compounds is dependent on the potential ambiguity of those compounds.

Ambiguity & Conceptual Classes

New compounds clearly have a primary aim to communicate some new conceptual content. Furthermore, they must convey this meaning in a wide variety of different contexts, where both the creator of the expression and where the specific referent of it may be absent. These requirements place very strong constraints on what combinations will succeed. They require a new compound to be as sematically umambiguous as possible; a requirement which may be a tall order given the inherent polysemy of such combinations.

It is well known that novel noun-noun compounds can be quite polysemous, suggesting multiple meanings (see e.g. Murphy, 1988; Kay & Zimmer, 1976). For example, a *pencil bed* could be a narrow bed, a container for pencils, a bed that is pencil-shaped and so on. Costello & Keane (1996) have systematically studied the polysemy of a large corpus of novel noun-noun combinations and found that on average they have about two meanings, with a range of between 1 and 6 meanings. It is clear that context will sometimes help to reduce this polysemy by selecting one meaning over another (Murphy, 1990). If you are told that "the pencil bed is in the bedroom upstairs" you are likely to assume that a narrow bed is upstairs, whereas if you are told that "the pencil bed is in the middle of the exam hall" you are more likely to think that it is some receptacle for pencils. However, not all contexts will necessarily disambiguate a novel combination; if you are told "he moved the pencil bed last week" either of the above two meanings could still hold (Mulligan, 1997).

Costello & Keane (1996) have also shown that different classes of concept have a significant effect on the inherent polysemy of a novel combination. They found that combinations containing artifacts are much more polysemous than ones containing natural-kinds. They provided evidence to show that this effect is due, in part, to the presence of functional models in artifacts which with their multiple roles can suggest many meanings (e.g., an elephant gun can be used by an elephant to shoot things or used by someone else to shoot elephants). The effect may also be due to, in part, to the broader scope of roles in artifact concepts (see Wisniewski & Gentner, 1991). Costello & Keane also found that combinations containing superordinate terms were more polysemous than ones containing basic-level terms (e.g., *street vehicle* versus *street bicycle*), based on the wider range of different concepts that could be used to instantiate the former over the latter. For example, a *street vehicl*e could be a bicycle for city use, a car for city use or the name of a skateboard, but a *street bicycle* will tend to be simply a bicycle that has some street-specific aspect to it. Costello & Keane (1996) have proposed a constraint theory of combination to explain these results and have modelled this theory computationally (see Costello, 1996, for details).

In these polysemy studies, the effect of concept class on polysemy was found to center on the *head position* of a compound phrase (i.e., the 2nd word of the combination) as opposed to the *modifier position* (the 1st word in the combination). If these same factors influence the appearance of compounds in a language then familiar noun-noun compounds should reflect these same constraints. However, the pattern of results should be the mirror image of the above polysemy results: the effects should be most visible in the modifier position rather than the head position and center on concept classes that reduce polysemy rather than increase it (i.e., the natural-kind and basic-level classes).

In most compound phrases, the head word indicates the general class of the intended referent and the modifier picks the referent out of the set of instances in that class (94% of new compounds in Cannon's, 1987, corpus have this endocentric pattern). According to our informativeness hypothesis, the selection of a head or modifier word should lead to a preference for polysemy-reducing classes of concepts. However, in the case of the head word, this preference may conflict with the need to identify the general class of the intended referent. In contrast, the modifier word is usually not constrained by the need to match the general categeory of the intended referent (an *apple tree* is a type of tree, not a type of apple). Therefore, the influence of concept class should be most apparent in the modifier word used.

With respect to the artifact/natural-kind distinction this implies there should be a greater tendency to use natural-kinds in the modifier position than in the head position, because natural-kinds suppress polysemy and in the modifier position they will have their greatest scope for reducing polysemy. In contrast, there should be a tendency against the use of artifacts in the modifier position, relative to their usage in the head position; as a combination with a artifact modifier will promote polysemy. With respect to the superordinate/basic-level distinction there should be a greater tendency to use basic-level terms in the modifier position than in the head position, again because the effect such terms have on reducing polysemy should be most visible in the modifier position. Similarly, as in the case of artifact terms, combinations should avoid using superordinates in the modifier position, relative to their usage in the head position. If these factors have no effect on compounds then the occurrence of these concept classes should be uniform in the modifier and head positions of familiar combinations.

It may be noted that these predictions are based on relative comparisons of occurrences of terms in the modifier/head positions of compounds, rather than absolute predictions about levels of occurence. That is, we do not predict that many more natural-kinds will be used than artifacts in the modifier position (even though this may be plausible), because we do not know *a priori* the baseline occurrence of terms from these different classes in the language. For example, people might just use more natural-kinds because they know and use more natural-kinds in everyday life (though this seems to be truer of artifacts). By making relative predictions, we control for the possible effects of the baseline availability of terms in the language and, indeed, baseline rates for new objects in the world that have to be named.

The Topology of Familiar Noun-Noun CompoundS

To test the above hypotheses, we collected a large corpus of familiar noun-noun combinations. In an early study, we asked subjects to spend two weeks generating a list of familiar combinations used in everyday life. In the present study, we adopted a more controlled procedure in which subjects were given 30 minutes to generate familiar combinations in the laboratory. In both studies, raters then classified the resulting combinations according to the conceptual class of the words used (i.e., artifact, natural-kind, superordinate, basic-level or other). Interestingly, the results of both studies were in close agreement. Here, we report the latter study in detail.

Method

Subjects & Procedure. Sixteen undergraduates at Trinity College, University of Dublin were paid to take part in the experiment. Each subject was asked to spend 30 minutes generating familiar compound phrases.

Materials. The materials consisted of a single instruction sheet. The instructions explained that compound phrases are phrases made up of two nouns, and asked subjects to "generate as many of these noun-noun combinations as you can think of. They should be commonplace ones -- that is, ones used in everyday life that are familiar to you.".

Scoring. All of the compound phrases produced were collated and completely lexicalised items[2], and phrases containing adjectives or verbs were removed. Multiple occurrences of phrases were also excluded. The resulting corpus of 1,459 unique phrases formed the basis of the materials used for the rating task.

Two independent raters were paid to categorise all the compound phrases produced. For each phrase the raters had to determine the class of the constituent concepts of the phrase on the artifact/natural-kind and superordinate/basic-level dimensions. The rating task took a total of 8 hours, divided up into two 4-hour sessions. The rating task was clearly a difficult one in terms of the judgements that had to be made, the number of items that had to be rated, and the length of time it took to complete. Perhaps as a result of this, rater error was high and the resulting levels of agreement on the classifiaction of a given combination were not as good as one would like (between 62% and 46%). Note that chance alone would produce an agreement level of 11% (because there are 9 possible classifications of each compound). However, it should be noted that separate analyses of each rater's classifications produced a pattern of results identical to the agreed-item analyses reported here.

Results

In general, the results reflect our predictions about the topology of familiar noun-noun combinations, based on their hypothesised role of communicating their meaning unambiguously. First, the percentage of combinations that used natural-kinds as modifiers was significantly higher than those using natural-kinds as heads. Second, the percentage of combinations using basic-level terms as modifiers was higher than those using basic-level terms as heads. Third, in the case of artifacts and superordinate terms there was a decrease in their occurrence in the modifer position relative to the head position.

Artifact / Natural-kind Dimension. Raters agreed on the classification of 62.5% (911) of familiar compounds involving the artifact/natural-kind dimension (see Table 3). As expected, the percentage of natural-kind

[2] Operationally defined as compounds in which the constituents were not separated (as in *marshmallow* or *bootleg*). These were excluded because any analysis of the constituent concepts of such lexicalised terms will rely on intuition to decide what their constituents actually are.

words in the modifier position (40%) was reliably higher than that of natural-kinds in the head position (16%; $Chi^2(1) = 44.3$, $p < .01$). Similarly, the frequency of artifacts in the modifier position (45%) was reliably lower than the frequency of artifacts in the head position (68% ; $Chi^2(1) = 21.6$, $p < .01$). So, as predicted, in the compounds reported by subjects, natural kinds tended to be used more often as modifiers and artifacts tended to be avoided as modifiers.

Table 1: Frequency of artifacts and natural-kinds in compounds agreed by both raters

Modifier	*Head*			
	Artifact	*Natural kind*	*Other*	
Artifact	282	46	78	**45% (406)**
Natural kind	233	82	48	**40% (363)**
Other	100	21	21	**15% (142)**
	68% (615)	**16% (149)**	**16% (147)**	**911**

Superordinate / Basic-level Dimension. Raters agreed on the classification of only 675 (46.3%) of phrases on the superordinate/basic-level dimension. The first rater classified 37 phrases (2.5%) as containing an "other" concept. This low level of "other" classifications is to be expected, since this dimension is highly inclusive: almost all concepts are either superordinate or basic-level concepts[3]. The second rater, by contrast, classified 524 phrases (35.9%) as containing a concept in the "other" category. This high level of "other" classifications suggests that the second rater did not properly understand the superordinate/basic-level classification task. We, thus, base our analysis here on the first rater's classifications alone (see Table 4).

As expected, the frequency of basic-level words in the modifier position (97%) was reliably higher than the frequency of basic-level words in the head position (79%; $Chi^2(1) = 7.37$, $p < .01$). Similarly, the frequency of superordinates in the modifier position (7%) was reliably lower than the frequency of superordinates in the head position (19%; $Chi^2(1) = 43.1$, $p < .01$).

Table 2: Frequency of superordinate and basic-level terms in compounds

Modifier	*Head*			
	Super	*Basic*	*Other*	
Super	27	71	1	**7% (99)**
Basic	243	1081	23	**92% (1347)**
Other	8	4	1	**1% (13)**
	19% (278)	**79% (1156)**	**2% (25)**	**1459**

It is notable that the frequency of basic-level terms was high in both positions: perhaps because basic-level terms are more *lexically available*, they easily come to mind and are better remembered than other terms (Murphy & Smith, 1982; Rosch, et al., 1976). This contrasts with the case of natural-kind/artifact terms, in which the predicted preference for natural-kinds was shown despite the fact that natural-kinds have a relatively low-level overall occurence. To put it another way, even though baseline rates of availability seem to differ for these different categories, the preferences for one category over another can still be found in the modifer-head comparisons we carried out.

General Discussion

The present study shows that the appearance of noun-noun compounds in the English language is influenced by conceptual constraints aimed at reducing the polysemy of such compounds. In particular, it shows that natural kinds are more likely to be used in the modifer position of a noun-noun combination (relative to their use in the head) because of their tendency to reduce the polysemy of such combinations. It also shows that more basic-level terms are likely to be used in the modifer position of compounds (relative to their use in the head) because of their polysemy-reduction tendencies. In contrast, both artifact and superordinate terms seem to be echewed as candidate words for the modifier position (relative to their use in the head position). These results establish a close link between the cognitive mechanism of conceptual combination and the growth of language through the coining of compound phrases. Indeed, we would hope that cognitive influences on the appearance of compounds found here may be fruitfully applied to elucidate a number of unanswered questions in that field (for

[3] Subordinate concepts are usually named by two-word compounds themselves (e.g. *mountain bike*) and thus would not appear as constituents in the set of compounds we examine here.

example, the question of how the best name for a new category is selected).

References

Bauer, L. (1983). *English word formation*. Cambridge: Cambridge University Press.

Cannon, G. (1987). *Historical change and English word-formation*. New York: Lang.

Costello, F. (1996). *Noun-noun conceptual combination: the polysemy of compound phrases*. PhD Thesis University of Dublin, Dublin, Ireland.

Costello, F. & Keane, M. T. (1996). *Polysemy in conceptual combination*. University of Dublin, Trinity College, Technical Report TCD-CS-96-19.

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

Gleitman, L. R. & Gleitman, H. (1970). *Phrase and paraphrase: Some innovative uses of language.* New York: Norton.

Grice, H.P. (1975). Logic and conversation. In P. Cole and J.L. Morgan (Eds.)*, Syntax and semantics (vol 3): Speech acts*. New York: Academic Press.

Kay, P. & Zimmer, K. (1976). On the semantics of compounds and genitives in English. *Sixth California Linguistics Association Proceedings.* San Diego, CA: Campile Press.

Murphy, G. L. (1988). Comprehending complex concepts. *Cognitive Science*, *12*, 529-562.

Murphy, G. L. (1990). Noun phrase interpretation and conceptual combination. *Journal of Memory and Language, 29*, 259-288.

Murphy, G.L. & Smith, E. (1982). Basic-level superiority in picture categorisation. *Journal of Verbal Learning & Verbal Behaviour*, *21*, 1-20.

Mulligan, M. (1997). *Context & Concept Combination*. University of Dublin, Trinity College, Technical Report TCD-CS-97-04.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382-439.

Ryder, M. E. (1994). *Ordered chaos: The interpretation of English noun-noun compounds*. University of California Publications in Linguistics, *123*. Berkeley, CA: University of California Press.

Shoben, E. J. (1993). Non-predicating conceptual combinations. *The Psychology of Learning and Motivation, 29*, 391-409.

Tulloch, S. (1992) *The oxford dictionary of new words*. Oxford: Oxford University Press.