Designing and Evaluating a Semantic Annotation Scheme for Compound Nouns

Diarmuid Ó Séaghdha[1]

Abstract

There is no standard set of semantic relations for classifying noun-noun compounds. This paper describes the development of a new annotation scheme which fulfils a number of desirable criteria. A rigorous dual-annotator experiment indicates that reasonably good agreement can be achieved but that the task remains a very difficult one. Analysis of the annotators' disagreements suggests which categories are most problematic and identifies specific cases for which the annotation guidelines could be further refined. Nonetheless there is a very long tail of disagreement patterns which render infeasible the production of fully exhaustive guidelines.

1 Introduction

Noun-noun compounds are sequences of two or more nouns that function as single lexical items, e.g. *fish knife*, *laptop computer*, *tree house*. Compounding is a very common and productive process in English and other languages, and the semantics of compounds has long been a topic of interest in descriptive, philosophical, psychological and computational studies of language. A wide range of semantic relations can hold between the entities referred to by a compound, and a recurring research question is how to produce a classification of these relations. Several authors including (Jespersen, 1942) and (Downing, 1977) have argued influentially that an exhaustive taxonomy cannot be produced as the number of relational concepts that can underlie a compound is potentially infinite. Yet even these authors recognise that while many compound meanings are lexicalised (e.g. *monkey business*) or highly context-dependent (e.g. *plate length*[2]), very many others are characterised by general categories such as identity, location and possession. Whether this is an inherent property of compounding or a fact about the way we conceptualise interactions between entities in the world, it suggests that developing a classification scheme of sufficient coverage and usefulness is a feasible goal.

The work described in this paper was carried out during the design of experiments on automatic interpretation of compounds using methods of statistical natural language processing (Ó Séaghdha and Copestake, 2007). It was initially motivated by the observation that there is a proliferation of classification schemes used by other researchers in computational experiments yet reported measurements of inter-annotator agreement are universally low and there is little discussion of schemes' relative merits and failings or of why one should be preferred to others. The case of

1 Computer Laboratory, University of Cambridge
e-mail: do242@cl.cam.ac.uk

2 "What your hair is when it drags in your food" (Downing, 1977, 828)

compound semantics contrasts with other tasks such as word-sense tagging, where there are at least standard sets of categories for classification, e.g. as WordNet (Fellbaum, 1998). The primary question investigated here is whether good agreement can be achieved through a rigorous procedure of annotation scheme development. The applied focus of the resulting scheme has in some places affected design decisions; this will be noted and explained below where appropriate. Nonetheless it is hoped that this work will also be of theoretical interest, both for its characterisation of compound semantics and for its more general contribution to the still underdeveloped field of semantic annotation design.

2 Desiderata for a semantic annotation scheme

In deciding on a classification scheme for compound relations, we are trying to pin down aspects of human conceptualisation that cannot be described using clear-cut observable distinctions, e.g. syntactic patterns. It is important not to choose a classification of relations on the sole basis of introspective intuition, as there is no guarantee that two subjects will share the same intuitions and it does not give us a basis to select one scheme among many. That said, the literature on "best practice" for semantic annotation schemes is rather sparse. The task shares some of the nature of ontology building and semantic field analysis, for which some design guidelines have been given by (Hovy, 2005) and (Wilson and Thomas, 1997) respectively, and the discussion in this section has much in common with the latter authors' proposals.

Faced with the need to select an appropriate classification scheme for compound relations, a number of desirable criteria were identified. They should be relevant for all semantic annotation studies. Most have an *a priori* theoretical motivation but they are also informed by the experience of developing our annotation scheme and became clear in the course of the development process:

1. **Coverage: The inventory of informative categories should account for as much data as possible.** The schemes of (Levi, 1978) and (Lauer, 1995) do not assign semantic relations to compounds whose head is a nominalised verb and whose modifier is an argument of that verb, leading to the unintuitive situation where *history professor* is assigned a semantic relation and *history teacher* is not. Lauer's scheme, which identifies semantic relations with prepositional paraphrases, also excludes equative compounds such as *woman driver* as they cannot be paraphrased prepositionally.
2. **Coherence: The category boundaries should be clear and categories should describe a coherent concept.** If categories are vague or overlapping then consistent annotation will be very difficult. Detailed annotation guidelines are invaluable for the clarification of category boundaries, but cannot save a scheme with bad conceptual design.
3. **Balance: The class distribution should not be overly skewed or sparse.** A central motivation for this criterion is the goal of creating a dataset for machine learning experiments – skewed class distributions cause particular problems for statistical classifiers (Zhang and Oles, 2001; Weiss and Provost, 2003). From a descriptive point of view, it may not be problematic that a single category out of five accounts for almost half of all data

(Nastase and Szpakowicz, 2003) or that the three most frequent categories out of eight account for three quarters (Lauer, 1995), as this may indeed be a fair reflection of the phenomenon of interest. It is more worrying when categories are posited which occur very rarely in corpus data, or where categories exist at very different levels of granularity. This typically occurs with large inventories of relations, as in (Nastase and Szpakowicz, 2003)'s fine-grained relations and those of (Girju *et al*., 2005).

4. **Generalisation: The concepts underlying the categories should generalise to other linguistic phenomena.** The regularities we hope to identify in compound relations or similar phenomena are assumed to reflect more general regularities in human semantic processing. Such regularities have been studied extensively by researchers in cognitive linguistics, and a categorisation scheme can be defended on the basis that it is consistent with and supported by those researchers' findings.

5. **Ease of Annotation: There should be detailed annotation guidelines which make the annotation process as simple as possible.** Of course, a coherent set of categories should be easier to annotate with than an incoherent set.

6. **Utility: The categories should provide useful semantic information.** The usefulness of a classification scheme is a subjective matter, and depends on how the annotated data will be applied. However, we can impose minimal criteria for utility. Each label in the scheme should be unambiguous and should carry truly semantic information. Hence Lauer's prepositional categories do not meet this requirement, as prepositions themselves can be ambiguous; the category OF can be assigned to *music school* (*school of music*), *computation theory* (*theory of computation*) and *church bell* (*bell of the church*) but these compounds all encode very different relations. A further concern affecting utility is the selection of granularity level, which must be fine enough for the intended application yet coarse enough to facilitate non-trivial generalisations about the data.

It is clear that there will be tension among these desiderata. Striving for balance may detrimentally affect coherence if unrelated concepts are conflated. A more "surfacy" set of categories may be easier to annotate but provide less useful semantic information. We can only achieve a best-fit solution.

How can these criteria be used to judge an annotation scheme? Generalisation and utility can only be argued subjectively and with theoretical evidence, but the others can be evaluated empirically through annotation experiments. Coverage can be directly measured from an annotated corpus as the proportion of data that is assigned a "useful" relation, i.e. one other than OTHER, UNKNOWN, *etc*. Balance can also be measured directly. Ease of annotation can be estimated through inter-annotator agreement between multiple annotators. Problems with coherence can be identified by analysis of inter-annotator disagreements. A definitive comparison of multiple schemes would require annotation of a single corpus with every scheme, but in practice this is rarely done.

3 Developing a new annotation scheme

The set of nine compound relations (BE, HAVE, IN, ABOUT, FOR, MAKE, CAUSE, USE, FROM) proposed by (Levi, 1978) were taken as an initial classification scheme. Levi's proposals are informed by linguistic theory and by empirical observations, and they intuitively seem to comprise the right kind of semantic relations. In attempting to annotate trial data with this scheme, however, a number of problems were identified that necessitated major revisions:

1. The CAUSE relation is extremely infrequent, with only two unambiguous examples (*blaze victim* and *staff cost*) identified in a sample of 300 compounds.
2. MAKE is also a scarce relation (9 occurrences in 300). More seriously, most if not all examples given by Levi for this relation can also be analysed as expressing other relations (for example, *sap tree* is also HAVE, *music box* is also FOR and *sugar cube* is also BE).
3. Nominalisations are analysed with a separate set of relations (SUBJ and OBJ). This is due to the assumptions of Levi's linguistic theory and not desirable under our approach.
4. More generally, Levi does not provide detailed guidelines for the application of her categories, and is not concerned with avoiding overlapping or vague category boundaries.

The annotation scheme was refined over the course of six months through a series of annotation trials followed by analysis of disagreements and changes in the scheme. Extensive guidelines were developed to clarify the application of the categories and the boundaries between them.[3] The most serious and pervasive problem encountered was that most compounds can be assigned multiple semantic relations even when their meanings are clear, but only one category per compound is permitted by our desired experimental design. For example, *car factory* is plausibly *a factory for producing cars* (FOR), *a factory that causes cars to be created* (CAUSE), *a factory in which cars are produced* (IN) and *a factory from which cars originate* (FROM). An *office chair* can be *a chair typically used/found in an office* (IN), *a chair for use in an office* (FOR) and *a chair belonging to an office* (HAVE). This phenomenon is problematic not just for Levi's scheme, but also for most other relation inventories described in the literature. To surmount this problem, the guidelines were refined to guide category selection in cases of doubt and the set of categories was modified. The MAKE, CAUSE and USE relations were replaced by two more general relations ACTOR and INST(rument) which apply to all compounds describing an event or situation in which the constituents are participants. These new relations therefore also account for most nominalised compounds and many compounds typically analysed as FOR. A consequence of this change was that FOR itself became redundant and was removed. This may seem surprising, given that PURPOSE is a traditionally uncontroversial entry in compound taxonomies and it is of course the case that many compounds denote the purpose of an item. However, most purpose-expressing compounds also seem to qualify for other relations: *dining room* and *kitchen knife* have strong locative senses, *cheese knife* and *welding iron* are good candidates for INST and *mining engineer* and *stamp collector* seem more naturally analysed as ACTOR. It was therefore decided that the purposive aspect of such compounds is not in opposition to what might be called their core semantics. Rather, it is simply a fact that a compound may have a particular semantics because that

3 The final version of these guidelines is available online at `http://www.cl.cam.ac.uk/~do242/guidelines.pdf`

semantics captures a salient characteristic of the compound's referent, and this might be due to intentionality, habituality, contrast with other instances of the head noun denotatum, or some other kind of "classificatory appropriateness" in the sense of (Zimmer 1971).

The development process resulted in six main "semantic" relations, three "miscellaneous" categories for compounds where the semantic relations do not apply, and two categories for sequences that are not valid compounds but have been identified as such by the automatic corpus compilation process (see Section 4.1). Each category is described in the guidelines by one or more rules which set out the various cases in which that category applies. The semantic categories and summaries of the rules are as follows:

1. **BE**
   - Rule 1.1 Identity (*learner driver*, *elm tree*)
   - Rule 1.2 Substance-Form (*stone obelisk*, *plastic box*)
   - Rule 1.3 Similarity (*father figure*, *pie chart*)
2. **HAVE**
   - Rule 2.1 Possession (*customer account*, *street name*)
   - Rule 2.2 Condition-Experiencer (*polio sufferer*, *cat instinct*)
   - Rule 2.3 Property-Object (*water volume*, *human kindness*)
   - Rule 2.4 Part-Whole (*car door*, *chicken curry*, *human blood*)
   - Rule 2.5 Group-Member (*stamp collection*, *infantry soldier*)
3. **IN**
   - Rule 3.1 Spatially located object (*forest hut*, *shoe box*)
   - Rule 3.2 Spatially located event (*dining room*, *hospital visit*)
   - Rule 3.3 Temporally located object (*night watchman*, *coffee morning*)
   - Rule 3.4 Temporally located event (*future event*, *midnight mass*)
4. **ACTOR** (most prominent participant role is sentient)
   - Rule 4.1 Participant-Event (*student demonstration*, *government interference*)
   - Rule 4.2 Participant-Participant (*honey bee*, *taxi driver*, *expressionist poem*)
5. **INST** (most prominent participant role is non-sentient)
   - Rule 5.1 Participant-Event (*skimming stone*, *machine translation*)
   - Rule 5.2 Participant-Participant (*rice cooker*, *tear gas*, *petrol motor*)
6. **ABOUT**
   - Rule 6.1 Topic-Object (*fairy tale*, *tax law*, *exclamation mark*)

Rule 6.2 Topic-Collection (*history exhibition*, *war archive*)

Rule 6.3 Focus-Mental Activity (*crime investigation*, *holiday plan*)

Rule 6.4 Commodity-Charge (*share price*, *income tax*)

The other categories are as follows:

7. **REL:** The compound does not belong to any of the above categories but seems to be produced by a productive pattern and is not lexicalised (*carbon dioxide*, *Penguin Books*)
8. **LEX:** The compound is lexicalised; replacing either of the constituents by semantically similar words does not give a semantically similar compound (*turf accountant*, *monkey business*, *home secretary*).
9. **UNKNOWN:** The meaning of the compound is not clear.
10. **MISTAG:** Either of the constituents were wrongly tagged as common nouns by the part-of-speech tagger (*London town*, *blazing fire*).
11. **NONCOMPOUND:** The extracted sequence, while correctly tagged, is not a valid 2-noun compound because of its context in the corpus ([*real tennis*] *club*, [*Liberal Democrat*] *candidate*)

Table 1 gives the distribution of these categories in the sample of 2,000 compounds described in Section 4.1, as annotated by the current author (due to constraints on time and resources it was not possible to have the entire set annotated by two annotators). Revisiting the desiderata of Section 2, it can be seen that the distribution of the six semantic relations is relatively balanced with no sparse categories and that their coverage is good (92.03% of syntactically valid compounds). The coherence of the categories and ease of annotation are tested by the annotation experiment described below. The generalisation criterion is satisfied as many of the guidelines are based on general linguistic principles such as animacy, substitutability and count/mass and event/object distinctions. In particular, the definition of the HAVE relation is based on the accounts of possession in (Taylor, 1996) and of part-whole relations in (Cruse, 1986). The ACTOR and INST categories are underpinned by a notion of underlying event that is compatible with frame semantic approaches to noun-noun compounds (Ryder, 1994; Coulson, 2001). The coarse semantic role hierarchy used to identify the more prominent of two mentioned participants in an underlying event and thus to distinguish between ACTOR and INST is inspired by (Talmy, 2000). Finally, the categories are useful in that they provide true semantic information, not ambiguous paraphrases.

Relation	Distribution
BE	191 (9.55%)
HAVE	199 (9.95%)
IN	308 (15.40%)
ACTOR	236 (11.80%)
INST	266 (13.30%)
ABOUT	243 (12.15%)
REL	81 (4.05%)
LEX	35 (1.75%)
UNKNOWN	9 (0.45%)
MISTAG	220 (11.00%)
NONCOMPOUND	212 (10.60%)

Table 1: Sample Class Frequencies

4 Experimental methodology

4.1 Data

A simple heuristic was used to compile a corpus of two-noun candidate compounds from the 90-million word written component of the British National Corpus (Burnard, 1995). The corpus was first lemmatised and tagged for parts of speech with RASP (Briscoe *et al*., 2006). We extracted every sequence of two common nouns which consist solely of alphabetic characters and are not adjacent to another common noun. Similar techniques were used by (Lauer, 1995) and (Lapata and Lascarides, 2003). This produced a corpus of almost 1.6 million tokens with 430,555 types.[4] A sample of 2,000 type-distinct compound tokens was randomly selected for use in the annotation experiments.

This heuristic can admit false positives for a number of reasons: tagging errors in the candidate sequence or in the adjacent words, "bracketing" issues whereby the modifier is itself a compound with noun head and non-noun modifier (e.g. [*real tennis*] *club*), and adjacency of nouns for reasons other than compounding (including lists and reduced relative clauses). (Lapata and Lascarides, 2003) report accuracy of 70.3% on identifying valid compounds in the BNC; the figures in Table 1 suggest that our heuristic, which is stricter due to the exclusion of all sequences containing non-

4 Even allowing for extraction error, this suggests that close to 3% of all words in the BNC are constituents of a noun-noun compound.

alphabetic characters, does better at about 78.4%. Lapata and Lascarides also describe how compound idenfication can be improved through statistical measures, but this has not been investigated in this work as the simpler heuristic seems sufficient.

4.2 Annotators

Two annotators were used – the present author (Annotator 1) and an annotator experienced in lexicography but without any special knowledge of compounds or any role in the development of the annotation scheme (Annotator 2). The distance of the second annotator from the development phase is important as her judgements should be based only on the text of the annotation guidelines and a small amount of clarificatory email correspondance, not on shared knowledge that might have emerged during development but not explicitly included in the guidelines. This adds rigour to claims of reproducibility regarding our agreement results.

4.3 Procedure

Each compound was presented alongside the sentence in which it was found in the corpus. Each annotator labelled it with the appropriate semantic category, the rule licensing that label in the annotation guidelines, and the order of compound constituents with regard to the argument slots in that rule. The following is a representative example:

```
483883: air disaster

IN,2,2.1.3.2

In the country 's fifth air disaster in four months , the
China Southern Airlines plane crashed as it approached to
land at the city of Guilin
|In_II| |the_AT| |country_NN1| |'s+_$| |fifth_MD| |air-
disaster_QNN1| |in_II| |four_MC| |month+s_NNT2| |,_,|
|the_AT| |China_NP1| |Southern_JJ| |Airline+s_NN2|
|plane_NN1| |crash+ed_VVN| |as_CSA| |it_PPH1|
|approach+ed_VVD| |to_TO| |land_VV0| |at_II| |the_AT|
|city_NN1| |of_IO| |Guilin_NN1|
```

Here the annotation states that the category is IN, it is *a disaster in the air* not *air in a disaster* and that the licensing rule is 2.1.3.2 *N1/N2 is an event or activity spatially located in N2/N1*.

Two trial batches of 100 compounds each were annotated to familiarise the second annotator with the guidelines and to confirm that adequate agreement could be reached without further revisions. The first trial resulted in agreement of 52% and the second in agreement of 73%. The result of the second trial, corresponding to a Kappa beyond-chance agreement estimate (Cohen, 1960) of 0.693, was very impressive and it was decided to proceed to a larger-scale task. 500 compounds not used in the trial runs were drawn from the 2,000-item set and annotated. As the data

contained many rare and technical terms (97 occur just once in the BNC), the annotators were permitted to make use of resources including Google, the Oxford English Dictionary and Wikipedia so that the task would not be compromised by an inability to understand the data.

5 Analysis

5.1 Agreement

Agreement on the 500-item test set was 66.2%. corresponding to a Kappa score of 0.62. This is lower than the result of the second trial annotation, but may be a more accurate estimate of the "true" population Kappa score due to the larger sample size. On the other hand the larger task size may have led to a decrease in agreement, as the test set annotation had to be done over the course of multiple days and inconsistencies may have resulted – the second annotator has endorsed this suggestion.

The granularity of the agreement analysis can be refined by considering the directionality and rule information included in the annotations. Agreement on category and directionality (order of the compound constituents with regard to the arguments listed in the rule) is similar to agreement on categories alone at 64% (Kappa = 0.606). Agreement on rules licensing category assignment is lower at 58.8% (Kappa = 0.562) but it should be borne in mind that the guidelines were not developed with the intention of maximising the distinctions between rules in the same category.

Unlike most other studies of compound annotation, this annotation task requires the annotator to distinguish syntactically valid compounds from non-compounds and lexicalised compounds from non-lexicalised ones in addition to assigning semantic relations to non-anomalous data items. To get a rough estimate of agreement on the six "semantic" categories that would be used in the classification experiments of (Ó Séaghdha and Copestake, 2007) and to aid comparison with studies that use cleaner pre-filtered data, an analysis was carried out using only those items which both annotators had labelled with one of those categories. This left 343 items with agreement of 73.6% and Kappa = 0.683. Of course, this is not a perfect estimate of agreement on these categories as it excludes items which one annotator labelled with a semantic category and the other did not but may have done if the other ``non-semantic'' categories were not available.

5.2 Causes of disagreement

It is interesting to investigate which categories caused the most disagreement, and which category-category boundaries were least clear. One simple way of identifying category-specific differences between the annotators is to compare the number of items each annotator assigned to each category; this may indicate whether one annotator has a stronger preference for a given category than the other annotator has, but it does not tell us about actual agreement. One-against-all agreement scores and the corresponding Kappa values can highlight agreement problems concerning a single category C by measuring agreement on the binary task of classifying the data as either C or not-C (i.e., as belonging to one of the other categories). These measures are given for the test data in Table 2. The most striking disparities in the per-category counts show a bias for INST on the part of Annotator 1 and a bias for ABOUT on the part of Annotator 2; there are also smaller

Category	Ann. 1	Ann. 2	Agreement	Kappa
BE	52	63	0.926	0.637
HAVE	59	77	0.888	0.525
IN	69	66	0.930	0.700
INST	73	42	0.902	0.523
ACTOR	52	48	0.948	0.711
ABOUT	55	83	0.908	0.616
REL	19	20	0.930	0.066
LEX	11	8	0.974	0.303
UNKNOWN	3	3	0.988	-0.006
MISTAG	57	52	0.966	0.825
NONCOMPOUND	50	38	0.964	0.776

Table 2: Per-category assignments for each annotator and one-against-all agreement measures

biases regarding BE, HAVE and NONCOMPOUND. The raw one-against-all agreement figures are universally high. This is not surprising as when there are many categories with a relatively balanced distribution, for any category C the majority of items will be clear-cut cases of the non-C category. More informative are the one-against-all Kappa values, which show agreement above 0.7 for IN, ACTOR, MISTAG and NONCOMPOUND, agreement close to 0.5 for HAVE and INST, and extremely low agreement (below 0.1) for REL and UNKNOWN.

Studying agreement between pairs of categories can explain which kinds of compounds are most difficult to label and can suggest where the annotation guidelines are in need of further refinement. Standardised Pearson Residuals (Haberman, 1973) were used to give a chance-corrected estimate of between-category agreement. These residuals are defined on a confusion matrix or contingency table of assignments and the residual e_{ij} for two categories i and j is given by

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{[\hat{p}_{i+}\hat{p}_{+j}(1-\hat{p}_{i+})(1-\hat{p}_{+j})]^{1/2}}$$

where n_{ij} is the observed value of cell ij and $\hat{p}_{i+}$, $\hat{p}_{+j}$ are row and column marginal probabilities estimated from the data. Intuitively, this residual compares the proportion of data items assigned by Annotator 1 to category i and by Annotator 2 to category j with the expected proportion given Annotator 1's overall proportion of assignments to i and Annotator 2's overall proportion of assignments to j, normalised by a variance term. The resulting table of residuals is therefore not symmetric, $e_{ij} \neq e_{ji}$. In the context of an annotation experiment it is expected that the observed data will diverge strongly from independence, giving large positive values on the same-category diagonals and negative off-diagonal values. Problematic boundaries can be identified where this

pattern is not observed.

Residuals for the experimental results are given in Table 3. There are clear problems with REL, LEX and UNKNOWN, precisely because the borders of these categories are very difficult to pin down. In the case of UNKNOWN disagreement is unavoidable as different annotators will bring different background knowledge to the task and some annotators may be more willing than others to assign a possible relation in doubtful cases. The only off-diagonal positive residual among the six semantic relations is between INST and ABOUT. Inspection of the data shows that this is due to a set of items such as *gas alarm* which can justifiably be interpreted as both *an alarm activated by the presence of gas* (INST) and *an alarm signalling the presence of gas* (ABOUT). In these cases Annotator 1 tended to assign INST and Annotator 2 tended to assign ABOUT. The low one-against-all Kappa score for HAVE seems to arise mainly from an interaction with REL; many of the problematic items here are borderline properties such as *pay rate* and *resource level*. Adding further examples to the annotation guidelines should clarify these cases. On the other hand, many disagreements fall into other patterns that are not common enough to show up in this analysis and thus constitute a "long tail" for which the provision of exhaustive guidelines is not practically feasible.

A different perspective on observed disagreements can be obtained through a general analysis of the reasons why annotators give different labels to a data item. In some cases, one annotator makes a mistake; in others, the annotation guidelines are unclear; in others, there is genuine disagreement about the meaning of the compound. The distribution of these factors can inform us of the genuine upper bound that can be achieved even with a perfect annotation scheme and error-free annotators and of the degree to which agreement could be improved by further refining the guidelines. To this end, a classification of disagreement types was produced and all disagreements in the annotated test corpus were attributed one of these types. In many cases the reason for disagreement was clear from the data; if not, it was identified by consultation among the annotators. The classification used and distribution of types were as follows:

1. True disagreement about the referent of the compound (10.06%). Examples are *peat boy*, which one annotator understood as *a boy who works with or sells peat* and the other understood as *a boy buried in peat*, and *school management*, which was understood both as *the personnel who manage the school* and as *the activity of managing the school*. It is possible that the number of these disagreements could be reduced by providing more context to the annotators, but these disagreements cannot be avoided completely.

2. Agreement about the compound referent, but disagreement about the relation between the nouns (20.12%). This often results from disagreement about the meaning of one of the compound's constituents; a *video phone* may be interpreted as *a phone that plays video (information)* (INST) or as *a phone that is also a video (player)* (BE), though both interpretations allow the compound to denote the same set of devices.[5] Likewise

5 There are many phenomena in natural language which exhibit clear ambiguity but do not usually lead to misunderstandings or breakdown in dialogue. Similar observations have been made about syntactic sentence structure by (Poesio, 1996) and (Sampson and Babarczy, 2006) and about "sloppy" anaphoric reference (Poesio *et al*., 2006).

sponsorship cash can be *cash gained through sponsorship* (INST) or *sponsorship in the form of cash* (BE). Annotation practice for some recurring compound classes of this type could be stipulated in the guidelines, but it is probably impossible to produce an exhaustive listing that would eliminate all disagreements.

3. Disagreement about part of speech or bracketing, whereby both analyses are plausible (11.83%). Examples are *mass death* (*mass* could be adjective or noun) and *new technology applications* (*applications of new technology* or *technology applications which are new*). These disagreements are unavoidable where noisy data is used.

4. Mistakes: one annotation clearly contradicts the guidelines and no reasonable explanation can be given for the annotation (8.88%). Examples found in the test data are *cat owner* (annotated as ACTOR, should be HAVE), *credit facility* (annotated as ABOUT, should be INST) and *pearl brooch* (annotated as BE, in context this is *mother of pearl brooch* and should be NONCOMPOUND). As might have been expected, the majority of mistakes were made by the annotator with less experience of the annotation scheme (Annotator 2).

5. Vague guidelines: there is probably agreement on the meaning of the compound but uncertainty about category boundaries leads to disagreement (20.71%). Many of these cases lie on the INST/ABOUT borderline discussed above. Others relate to vagueness in the distinction between common and proper nouns; one annotator labelled both *Christmas cake* and *Palace player* (*Palace* = Crystal Palace football club) as MISTAG while the other assigned IN and REL respectively, and the guidelines did not specify the correct annotation.

	BE	HAVE	IN	ACTOR	INST	ABOUT	REL	LEX	UNK	MIS	NON
BE	14.32	-1.63	-2.54	-2.48	-1.78	-2.22	-1.56	**0.20**	-0.59	-1.64	-1.63
HAVE	-1.02	11.87	-1.14	-1.72	-1.98	-3.28	**1.87**	-1.04	-0.64	-2.79	-2.35
IN	-3.01	-0.58	15.66	-2.04	-2.71	-2.95	**0.16**	-0.11	-0.70	-3.05	-2.57
ACTOR	-1.57	-1.63	-2.54	15.92	-2.31	-2.61	**0.69**	-0.97	**3.20**	-2.60	-2.18
INST	-0.84	-1.14	-1.36	-3.01	12.27	**0.64**	-0.59	-0.17	-0.72	-2.32	-2.65
ABOUT	-2.98	-2.56	-2.64	-1.59	-2.38	14.16	-0.88	**0.14**	-0.61	-2.68	-1.18
REL	-0.98	**0.05**	-1.73	-1.45	**1.18**	**3.05**	1.48	**1.30**	-0.35	-0.75	-1.27
LEX	**0.56**	**0.26**	-0.41	-1.09	-1.02	-0.68	**0.87**	6.86	-0.26	-0.14	-0.96
UNK	-0.66	**0.86**	-0.68	-0.57	**1.56**	-0.78	**2.60**	-0.22	-0.13	-0.59	-0.50
MIS	-1.77	-3.03	-2.71	-1.18	-1.92	-3.58	-0.92	-1.02	**1.20**	18.47	-2.30
NON	-1.93	-1.94	-2.91	-1.42	-1.18	-1.32	-0.76	-0.95	-0.58	-2.54	17.55

Table 3: Standardised Pearson Residuals for the annotated test set; off-diagonal positive values are in bold

6. There is no evidence of disagreement about the compound's meaning but at least one annotator has assigned one of the categories REL, LEX and UNKNOWN (28.4%). As observed above, these categories are especially problematic. As they apply when no other category seems appropriate, some disagreements of this type could be reduced by clarifying the boundaries of the other categories. For example, disagreement about *football enthusiast* (one annotator has ACTOR, the other REL) and about *pay rate* (HAVE versus REL) might be avoided by improving the definitions of ACTOR and HAVE respectively. On the other hand, it is harder to solve the problem of distinguishing lexicalised compounds from non-lexicalised. The substitutability criterion used in the guidelines for LEX functions well much of the time, but different annotators will have different intuitions about substitutability and disagreements may be inevitable. Examples found in the test data include *platform game*, *rugby league* and *trace element*. As previously noted, the UNKNOWN category will always be likely to cause disagreements, though the overall number of assignments to this category might be reduced by the provision of more context.

It has been argued that for part of speech annotation (Babarczy *et al*., 2006) and for syntactic annotation of sentences (Sampson and Babarczy, 2006), the abilities of annotators to follow guidelines contribute more to annotation disagreements than imprecision in those guidelines does. Those studies use a highly-refined exhaustive set of annotation guidelines and expert annotators, so their results will be more conclusive than ones drawn from the current study. However, the breakdown of disagreement types presented here does suggest that even with a rigorously developed annotation scheme the division of responsibility is less clear in the case of compound semantics. If we attribute all cases of true disagreement and all mistakes (categories 1–4) to annotator issues, 50.86% of disagreements can be thus accounted for. Perhaps some of these could be resolved by expanding the guidelines and providing more context around the compound to the annotators. However, there are only a few obvious cases where a change in the guidelines would make a significant difference to the agreement rate. All category 5 cases are due to the annotation guidelines. It is less clear how to analyse category 6 cases. In many, the annotators may agree on the compound semantics but be unclear whether or not it fits into one of the six semantic categories, or whether or not it is lexicalised. This suggests that the problem lies with the guidelines, but beyond certain common disagreement types, it will be difficult to solve. The conclusion drawn from this analysis is that it may not be practically feasible to develop an annotation scheme for compound relations with the same precision as has been achieved for syntactic annotation tasks.

6 Discussion

This work appears to be the first reported study of annotating compounds in context. This aspect is important, as in-context interpretation is closer to the way compounds are used and understood in the real world, and compound meanings are often context-dependent. It is not clear whether in-context or out-of-context interpretation is easier, but they are indeed distinct tasks. Out-of-context interpretation relies on a compound having a single most frequent meaning and where this holds agreement should be higher. In-context interpretation allows even improbable interpretations to be considered (a *fish knife* could be *a knife that looks like a fish*) and where the intended meaning is

not fully explicit in the context annotators may vary in their willingness to discard the most frequent meaning on the basis of partial evidence.

Some authors of compound annotation schemes and compound datasets do not describe any inter-annotator agreement measurements, notably (Lauer, 1995) and (Nastase and Szpakowicz, 2003). Other authors have given out-of-context agreement figures for corpus data. (Kim and Baldwin, 2005) report an experiment using 2,169 compounds taken from newspaper text and the categories of Nastase and Szpakowicz. Their annotators could assign multiple labels in case of doubt and were judged to agree on an item if their annotations had any label in common. This less stringent measure yielded agreement of 52.31%. (Girju *et al*., 2005) report agreement for annotation using both Lauer's 8 prepositional labels (Kappa = 0.8) and their own 35 semantic relations (Kappa = 0.58). These figures are difficult to interpret as annotators were again allowed assign multiple labels (for the prepositions this occurred in "almost all" cases) and the multiply-labelled items were excluded from the calculation of Kappa. This entails discarding the items which are hardest to classify and thus most likely to cause disagreement. It is clear that the agreement results reported in this paper compare favourably with other results in the literature; one significant factor in this success is the rigorous development of the annotation schemes and guidelines.

(Girju, 2006) has recently reported impressive agreement (Kappa = 0.67) on a compound annotation task, but differences in experimental design preclude direct comparison. The data used in this experiment was taken from a multilingual dictionary and thus might be expected to contain more familiar terms than a balanced corpus containing many technical items and context-dependent usages. Compounds judged to be lexicalised were discarded and there was no noise in the data as it was not extracted from a corpus. Furthermore, each compound was presented alongside its translation in four Romance languages. Compounding is relatively rare in these languages and English compounds often have periphrastic translations that disambiguate their meaning – this was in fact the primary motivation for the multilingual experiment. On the other hand, the annotation involved a larger set of semantic categories than the six used in this work and the annotation task will therefore have been more difficult in one aspect; the author lists 22 categories, though only 10 occur in more than 2% of her data.

It is clear from the results reported here and by other authors that the compound annotation task is a very difficult one. Why is this the case? A general problem in semantic annotation of text is that the annotator does not have access to all the information available to the author and his/her intended audience. Interpreting referring expressions in dialogue has been shown to be much harder for overhearers than for participants (Schober and Clark, 1989). In technical or specialist genres, an annotator may lack much of the background knowledge required to arrive at a full or correct interpretation. Even where the source of the data is written and intended for a general readership, it is not practical to read a large portion of the source text as may be necessary for accurate interpretation. This difficulty is exacerbated in the case of compounds, which are often regarded as compressed descriptions of their referents (Downing, 1977). To decompress the semantics of a compound, the hearer must share certain knowledge with the speaker either through mutual world knowledge or through common ground established in the preceding text. The use of compounds thus reflects the tendency of speakers to use shorter referring expressions as a discourse develops

(Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986; Master, 1993) and the tendency to reduce redundant syntactic structures and maintain a constant information density (Levy and Jaeger, 2007). Much of the difficulty in annotation thus arises from the very nature of compounds and compound usage.

7 Conclusion

This paper has described a novel annotation scheme for compound relations accompanied by detailed, publicly available guidelines. In addition, a number of general criteria for evaluating semantic annotation schemes has been presented. The relatively good inter-annotator agreement figures confirm the value of a rigorous development process and of satisfying these desiderata. Howver, the annotation task remains difficult and seems unavoidably so given the long tail of disagreement patterns and the manner in compounds are processed by speakers and hearers in discourse. Further revision of the guidelines should be helpful in resolving certain frequent disagreements but will ultimately yield diminishing returns. It may be a general property of all semantic annotation that exhaustivity is beyond reach.

References

Babarczy, A., J. Carroll and G. Sampson (2006) Definitional, personal, and mechanical constraints on part of speech annotation performance. *Journal of Natural Language Engineering* 12(1): 77–90.

Briscoe, T., J. Carroll and R. Watson. (2006) The second release of the RASP system. *Proceedings of the ACL-06 Interactive Presentation Sessions*.

Burnard, L. (1995) Users' Guide for the British National Corpus. British National Corpus Consortium, Oxford University Computing Service.

Clark, H. H. and D. Wilkes-Gibbs (1986) Referring as a collaborative process. *Cognition* 22(1): 1–39.

Coulson, S. (2001) Semantic Leaps: Frame Shifting and Conceptual Blending in Meaning Construction. Cambridge: Cambridge University Press.

Cruse, D. A. (1986) Lexical Semantics. Cambridge: Cambridge University Press.

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

Fellbaum, C. (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Girju, R. (2006) Out-of-context noun phrase semantic interpretation with cross-linguistic evidence. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*.

Girju, R., D. Moldovan, M. Tatu and D. Antohe (2005) On the semantics of noun compounds. *Computer Speech and Language* 19(4): 479–496.

Haberman, S. J. (1973) The analysis of residuals in cross-classified tables. *Biometrics* 29(1): 205–220.

Hovy, E. (2005) Methodologies for the reliable construction of ontological knowledge. *Proceedings of the 13th Annual Conference on Conceptual Structures*.

Jespersen, O. (1942) A Modern English Grammar on Historical Principles, Part VI: Morphology. Copenhagen: Ejnar Munksgaard.

Kim, S. N. and T. Baldwin (2005) Automatic interpretation of noun compounds using WordNet similarity. *Proceedings of the 2nd International Joint Conference on Natural Language Processing*.

Krauss, R. M. and S. Weinheimer. (1964). Changes in the length of reference phrases as a function of social interaction : A preliminary study. *Psychonomic Science* 1: 113–114.

Lapata, M. and A. Lascarides (2003) Detecting novel compounds: The role of distributional evidence. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Lauer, M. (1995) Designing Statistical Language Learners: Experiments on Compound Nouns. PhD thesis, Macquarie University.

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

Levy, R. and T. F. Jaeger (2007) Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt and T. Hoffman (eds.), *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press.

Master, P. (1993) On-line nominal compound formation in an experimental pidgin. *Journal of Pragmatics* 20(4): 359–375.

Nastase, V. and S. Szpakowicz (2003) Exploring noun-modifier semantic relations. *Proceedings of the 5th International Workshop on Computational Semantics*.

Ó Séaghdha, D. and A. Copestake (2007) Co-occurrence contexts for noun compound interpretation. *Proceedings of the ACL-07 Workshop A Broader Perspective on Multiword Expressions*.

Poesio, M. (1996) Semantic ambiguity and perceived ambiguity. In K. van Deemter and S. Peters (eds.), *Ambiguity and Underspecification*. Stanford, CA: CSLI Publications.

Poesio, M., P. Sturt, R. Artstein and R. Filik (2006) Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes* 42(2): 157–175.

Ryder, M. E. (1994). Ordered Chaos: The Interpretation of English Noun-Noun Compounds. Berkeley, CA: University of California Press.

Sampson, G. and A. Babarczy (2006) Definitional and human constraints on structural annotation of English. *Proceedings of the 2nd Conference on Quantitative Investigations in Theoretical Linguistics*.

Schober, M. F. and H. H. Clark (1989) Understanding by addressees and overhearers. *Cognitive Psychology* 21: 211–232.

Talmy, L. (2000) The semantics of causation. In *Toward a Cognitive Semantics, vol. 1: Concept Structuring Systems*. Cambridge, MA: MIT Press.

Taylor, John R. 1996. Possessives in English: An Exploration in Cognitive Grammar. Oxford: Oxford University Press.

Weiss, G. M. and F. Provost (2003) Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research* 19: 315–354.

Wilson, A. and J. Thomas (1997) Semantic annotation. In R. Garside, G. Leech and A. McEnery (eds.), *Corpus Annotation*. London: Longman.

Zhang, T. and F. J. Oles (2001) Text categorization based on regularized linear classification methods. *Information Retrieval* 4(1): 5–31.

Zimmer, K. E. (1971) Some general observations about nominal compounds. *Stanford Working Papers on Linguistic Universals* 5: C1–C21.