Dynamic Creation of Analogically-Motivated Terms and Categories in Lexical Ontologies

Tony Veale,
School of Computer Science and Informatics, University College Dublin, Ireland.

1. Introduction

Ontologies are, for the most part, static organizations of categories and relations that attempt to model some aspect of the world. But such organizations can be strained to the breaking point when creative actions in this sub-world necessitate a dynamic response in the corresponding ontological structures. An important case in point is when, through some purposeful activity, an agent dynamically creates a new category of entity on the fly that must then be accommodated within the ontology (e.g., see Way, 1991; Veale, 2003a, 2005). This contrasts with the creation of new senses for existing ontological terms (as, for example, described in Pustejovsky, 1991). Such "ad-hoc" categories, as described by Barsalou (1983), are typically created in response to a specific goal or task, and may thus be considered task-specific or goal-specific constructs. Examples of ad-hoc categories include "things to take on a camping trip", "useful wedding presents", "substances from which sugar can be extracted", "objects that would make unusual murder weapons", and so on. Ad-hoc categories do not correspond to the existing hierarchical categories of an ontology, and their members are rarely clustered in the same localized area of the ontology. Rather, the members of a given ad-hoc category may be drawn from many different established categories. Ad-

hoc categories thus constitute a horizontal rather than vertical slice of an ontology, cutting across conventional hierarchical structures, and as such, the creation of ad-hoc categories may intuitively be seen as a form of "lateral thinking" (de Bono, 1994).

The cross-cutting nature of ad-hoc categories means that they cannot easily be lexicalized with any of the labels associated with existing hierarchical categories; in fact, as seen above, the lexical label given to an ad-hoc category can be quite a mouthful, as a new multi-word expression must be constructed to capture the functional rather than taxonomic nature of the category. Nonetheless, some ad-hoc categories can be given compact labels that may subsequently find use as meaningful collocations in their own right and which merit their own individual listings in the lexicon. The creativity inherent in the construction of ad-hoc categories can thus apply at two different levels of representation, involving the creation of not just new ontological categories but of new lexical entries as well.

Of course, these levels of representation become one and the same when the ontology in question is a lexical ontology (see Veale *et al.* 2004, Hayes and Veale, 2005), that is, an ontology that concerns itself only with those units of meaning, called lexical concepts, that correspond to specific words or compound terms in a language. In this case, the creation of a new ontological category will correspond to the creation of a new lexical concept, forcing all new categories to assume a lexical label that serves a useful indexing role in the lexicon. For example, a new category like French-Food will serve to cluster together the various foodstuffs, dishes and wines that can be considered French into a single ontological category. At present, a number of large-

scale lexical ontologies are available to support research in this area, such as WordNet, a comprehensive electronic thesaurus of English whose design reflects psycholinguistic insights into the structure of the mental lexicon. Thus, nouns, verbs and adjectives are represented by WordNet in different ways, and each lexical concept partakes in one or more relations to other concepts. HowNet (see Dong 1998; Carpuat *et al.* 2002) is an equivalent Chinese ontology (with English translations) in which each lexical concept is associated with a propositional semantic structure. Both WordNet and HowNet are more properly described as weak ontologies since they exhibit neither the relational richness nor formal precision of the structures normally called ontologies by philosophers. In contrast, the Cyc ontology of Lenat and Guha (1991) is a rich axiomatic system explicitly designed to support logical inference and knowledge-based problem-solving, but which also maps word-forms onto a selection of underlying logical forms, both atomic and formulaic. However, though Cyc deserves to be called a strong ontology, it is not a lexical ontology, since natural language is not a motivating factor in its design; consequently, Cyc contains a high proportion of unlexicalized concepts and new concepts are not required to have corresponding lexical forms. Philosophers of language often contrast the role of the dictionary and the encyclopaedia when considering the knowledge demands of language; a lexical ontology aims to capture key aspects of both the dictionary and the encyclopaedia, and so constitutes the ideal framework in which to explore the mechanics of term creation.

Since ad-hoc categories are goal-specific, different goal contexts might give rise to different kinds of ad-hoc categories, suggesting that ad-hoc categories are best studied

from the perspective of a specific cognitive goal or task. One such important task to which lexical ontologies have been directed is the construction and interpretation of lexical analogies. Analogy has been identified as a reasoning mechanism at several different levels of linguistic operation, from native speaker intuitions about pronunciation (see Baron, 1977) to intuitions about morphological inflection (see Trask, 1996) to intuitions about semantic relatedness (see Rumelhart and Abrahamson, 1973). At a lexico-conceptual level, analogies such as *Fructose is to Fruit as Lactose is to Milk* (e.g., see Veale 2003, 2004, 2005) exhibit creativity not only in their production, since they constitute novel linguistic artefacts, but also in their interpretation, which frequently requires the dynamic construction of new ad-hoc categories. Consider the joke given in Freud (1905), and analysed as an analogy in Attardo *et al.* (2002):

"A wife is like an umbrella. Sometimes one takes a cab"

Attardo *et al*. (2002) provide the missing concept, Prostitute, to complete the analogy:

wife : prostitute :: umbrella : cab

To understand the analogy (and thus the joke) the listener must recognize that wives are personal lovers, while prostitutes are hired lovers; and that umbrellas are personal resources, while cabs are hired resources. This recognition necessitates the creation of the ad-hoc categories Personal-Lover and Personal-Resource (both sub-types of Personal-Belonging, to take a Victorian view of marriage and wives), as well as Personal-Resource and Hired-Resource. The burden of creativity is not borne solely by the creator of the joke, as the listener must also carry much of this burden through

the creation of new categories that mirror the mind-set of the joker. Now, categories like Hired-Lover are highly goal-specific, and may not persist beyond the immediate context of the analogy that gives rise to them. However, if an ad-hoc category demonstrates some long-lasting value, its lexical label may also persist, to the point that it becomes a unit of common currency in the language and a permanent entry in a speaker's lexical ontology.

In this paper we consider how a particular kind of lexicalized ad-hoc category is created in a lexical ontology when the motivating task is that of analogical reasoning. We argue that the interpretation of lexical analogies often necessitates the creation of new conceptual categories that in turn necessitate the creation of new lexical items. These lexical terms may be as short-lived as the analogies themselves, but a corpus analysis can be used to reveal those terms that have sufficient durability to merit a place in the lexical ontology. For our current purposes, we ground our investigation in the context of WordNet, and explore a variety of ways in which analogy can be used to drive the creation of lexical innovations that do not already exist within the WordNet lexicon.

We are careful to note that, in the context of a lexical ontology, the term "analogical categorization" is an ambiguous one. It can mean either the creation of new categories, like Hired-Lover and Personal-Resource, to resolve a particular analogy, or it can denote the use of analogy as an explicit term-creation mechanism, in much the same way that analogy can be used to suggest spelling, morphology and pronunciation. Since each reading denotes a process that is meaningfully performed

within a lexical ontology, the ambiguity is a benign one that reflects the multipurpose nature of lexical analogy. So for the sake of completeness, we consider both of the foregoing uses in this paper. In sections two and three we explore the use of analogy as an explicit and quite deliberate mechanism of term creation, while in section four we consider how new ad-hoc concepts and their corresponding lexicalizations can arise as by-products from the interpretation of lexical analogies. More specifically, section two considers how analogy can be used to produce new words, while sections three and four explore the role of analogy in the creation of new compound terms. In section five we then consider the application of these ideas to the WordNet lexical ontology, which allows us to empirically evaluate their effectiveness in the context of a real analogical retrieval task. We then conclude with some remarks on the limitations of these ideas in section six.

2. Analogy as a Mechanism of Word Creation

To begin, one might well consider how analogy is implicated in the determination of morphological inflections, even in the face of stronger and more accepted linguistic principles. Indeed, as noted by Trask (1996), analogies can be so persuasive that any fallacious conclusions that can be drawn may seem even more natural than those of a first principles analysis. For instance, the lexical analogy in (1) is compelling even though the *+us→+i* pluralization rule is valid only for words of Latin origin, as exemplified by radius/radii and succubus/succubi.

(1) cactus : cacti :: octopus : octopi

The correct inflection, following the Greek origins of the word octopus, is "octopodes", yet this is far less favoured by English speakers, to the extent that the automatic spelling corrector provided by Microsoft Word deems "octopi" to be valid and "octopodes" to be a misspelling. This usefulness of analogy in dealing with irregular plurals even extends to the treatment of regular verbs, where a compelling analogy can make even a regular verb like "dive" seem irregular. The analogy *drive:drove:dive:dove* will rightly strike some readers as invalid, yet many Eastern American speakers strongly prefer "dove" to "dived" (Trask, ibid). In contrast, the analogy *teach:taught::catch:caught* seems a valid one, though as Trask notes, "catched" is actually the historically favoured past tense of catch. Again, however, analogy prevails to the extent that most spelling checkers will flag "catched" as a misspelling (no doubt due to the fact that spelling checkers are based on a corpus analysis of how language is actually used, rather than how it should be used).

Morpheme-level analogies can do more than suggest inflection patterns, and can even be used to derive new words and meanings that frequently exhibit a high degree of lexical creativity. Consider an example of analogy-based derivational morphology:

(2) astronomy : astronomer :: gastronomy : gastronomer

Neither WordNet nor the spelling checker for Microsoft Word recognize "gastronomer" as a valid word, though a web-search reveals that it is a real word with the semantics that one would expect from the analogy (i.e., a specialist in gastronomy). The analogy in (2) is semantically sound since Astronomy and Astronomer are strongly related concepts, but as one allows the analogy to veer towards the speculative, and to rely as much on sound similarity as semantic similarity, one can

achieve even more innovative results, as in (3).

(3) astronomy : astronaut :: gastronomy : gastronaut

The relation between astronomy and astronaut is a good deal more tenuous than that between astronomy and astronomer, but a relation does exist (one observes the stars, the other explores the stars, in name at least). Indeed, one can argue that Astronaut is itself analogically derived from Argonaut. A gastronaut might thus denote anything from an adventurous gastronomer to a food tourist; at the very least, we know that a gastronaut is a person, with some of the signal characteristics of an astronaut (bravery, perhaps), who takes his directions from the field of gastronomy.

If it seems that the process of lexical creativity can be as strongly influenced by phonetic concerns as semantic concerns, this should not be too surprising a conclusion. New words survive and thrive for a whole host of reasons, but an important factor in their survival is euphony: natural sounding words are more likely to secure a lasting place in the lexicon than those that are difficult to pronounce. Analogies with existing words can transplant the euphony of an original form onto a newly minted neologism only if phonetic similarity is also allowed to influence the mapping. Indeed, the most innovative creations may give so much prominence to phonetic similarity that an analogy may lack a credible semantic basis. Nonetheless, as demonstrated in (4), terms predicated on a false analogy can still be seen as lexically innovative:

(4) astronomy : astrodome :: gastronomy : gastrodome

Of course, there is no real semantic connection between astronomy and astrodome. Nonetheless, the word “gastrodome” can be seen as a deliberate malapropism that amply suggests a place where gastronomy is performed, and perhaps even celebrated.

The analogy works, despite its lack of semantic grounding, because "dome" is itself a word denoting a large enclosed space where people congregate, like an arena or stadium. A second, implied analogy can be used to tease out its precise meaning:

(5) astrodome : stadium :: gastrodome : restaurant

That is, just as an astrodome is a large, impressive stadium, a gastrodome is a large impressive restaurant (where restaurant is itself implied by the morphological conjunction of gastronomy and place in "gastrodome"). Ultimately, "Gastrodome" is preferable to the neologisms "Gastroarena" and "Gastrostadium" in part because the largely phonetic analogy ensures that it is a euphonious combination of morphemes, each of which can be considered in isolation to provide a compositional meaning to the neologism as a whole.

3. Analogy as a Mechanism of Term Creation

Moving from the level of morphemes to that of words, analogy again reveals itself as a powerful force in the creation of compound terms. Consider the analogy of (6):

(6) Greek-Alphabet : Hebrew-Alphabet :: Greek-Deity : Hebrew-Deity

The analogy captures a basic symmetry both in the way concepts can be differentiated and how such differentiations are lexically expressed as compound terms. Both "Greek" and "Hebrew" denote a cultural amalgam of people, language and belief, so it makes sense to conclude that if "Greek" can be used to culturally differentiate a particular concept, then so can "Hebrew". In fact, WordNet contains only three of the four compound terms in the above analogy: "Greek-Alphabet", "Hebrew-Alphabet"

and “Greek-Deity”. The lexical concept “Hebrew-Deity” is not listed as a WordNet entry, most likely because it is deemed to have little indexing value; while there are many Greek deities listed in WordNet that would structurally benefit from the clustering offered by the hypernym Greek-Deity, only one deity, Jehovah, is listed as having a Hebrew origin. Yet the concept seems logically well-formed, and a usage analysis (using the World Wide Web as a corpus) reveals that the term “Hebrew deity” has relatively widespread acceptance. The analogy suggests then that, on the basis of the similarity between Greek and Hebrew, WordNet should incorporate the lexical concept Hebrew-Deity. Were it to do so, its treatment of deities would become more systematic and balanced, with each proper deity (such as “Zeus”, “Mars” and “Jehovah”) instantiating a compound category that denoted its cultural basis.

Nonetheless, a simple proportional analogy like (6) may seem a weak basis on which to predict the existence of a new term. Consider the analogies of (7) and (8):

(7) Roman-Alphabet : Greek-Alphabet :: Roman-Empire : Greek-Empire

(8) Roman-Alphabet : Hebrew-Alphabet :: Roman-Empire : Hebrew-Empire

Here we seem to be predicating the possession of an empire on the existence of a unique alphabet, but an alphabet alone does not an empire make. The analogy of (7) holds true, since there is a historical entity called the “Greek Empire”, but the term “Hebrew Empire” can only be used metaphorically, perhaps to refer to the Jewish diaspora. However, this is not to say that analogy cannot be of use here, for what (7) and (8) fail to reveal is the variety of different analogies that support (7), and the comparative dearth of analogies that support (8). In addition to alphabets, the Greeks and the Romans both possessed their own mythologies, architectures, religions and deities. This semantic isomorphism suggests that if Rome possessed its own empire, it

is at least meaningful to consider the possibility of a Greek empire also. The analogy in (7) is therefore strengthened by the lexico-conceptual fit between the concepts Roman and Greek, while the analogy in (8) is much weaker because of the lack of a coherent fit. This "fit" is not a measure of ontological closeness, but a measure of the overlap between the set of affordances possessed by both concepts. We can loosely estimate this set of affordances by observing the lexical behavior of each term and how it relates to others. These observations will require us to define a set of basic term composition and decomposition operators along the following lines:

$\mathbf{U_M}$ **{X}** : *Usage as modifier*: return a set of all compound terms such that the modifier of each is a member of the set {X}.

E.g., $\mathbf{U_M}$ {Greek, Roman} = {Roman-deity, Greek-deity, ...}

$\mathbf{U_H}$ **{X}** : *Usage as head*: return a set of all compounds such that the head of each is a member of the set {X}.

E.g., $\mathbf{U_H}$ {Greek, Roman} = {Ancient-Greek, Times-Roman, ...}

M {X} : *get modifiers*: return the set of all modifiers of all compounds in {X}

E.g., **M** {Ancient-Greek, Times-Roman, ...} = {Ancient, Times, ...}

H{X} : *get heads*: return the set of all heads of all compounds in {X}

E.g., **H** {Ancient-Greek, Times-Roman, ...} = {Greek, Roman, ...}

C ({X}, {Y}) : *combination*: return the set of all possible compound terms whose modifier is in {X} and whose head is in {Y}

E.g., **C**({Greek, Hebrew}, {Alphabet, Deity}) = {Greek-Alphabet, Greek-Deity, …}

These operators allow us to dissect existing compound terms into their component parts (modifier and head), retrieve compound terms with a particular sub-component (modifier or head), and create novel combinations of these sub-components (modifiers crossed with heads). We can thus estimate the fitness of a novel compound X-Y in terms of the set of known compounds that support it, as follows (where ***L*** here denotes the set of all lexical items in the lexicon, i.e., all known terms):

$$\textit{support-set}(\text{X-Y}) = \mathbf{C}(\mathbf{M}(\mathbf{U}_{\mathbf{H}}\{\text{Y}\}), \mathbf{H}(\mathbf{U}_{\mathbf{M}}\{\text{X}\})) \setminus \boldsymbol{L}$$

For example, consider the support set for the novel compound Hebrew-Deity from (6):

support-set(Hebrew-Deity)

= C(**M**({Greek-Deity, Roman-Deity, Semitic-Deity, ...}),
H({Hebrew-Alphabet, Hebrew-Calendar, …})) \ ***L***

= **C**({Greek, Roman, Semitic, Hindu, Celtic, Norse, …},
{Alphabet, Calendar, Lesson}) \ ***L***

= {Greek-Alphabet, Roman-Alphabet,
Roman-Calendar, Hindu-Calendar}

In effect, this set of four existing compounds represents the lexico-conceptual cross-product of the lexical concepts Hebrew and Deity. The larger the cross-product, the greater the potential interaction – and the greater the lexical fit – between both terms.

One might think it strange that a lexical concept like Hindu-Calendar should support a term like Hebrew-Deity, but the intuition at work here is that deities and calendars appear to be differentiated in the same kind of way (e.g., culturally) and thus possess many of the same affordances.

Working backward from this formulation of fitness, we can formulate a generation mechanism for producing new compound terms from old, one that implicitly incorporates the notion of lexical analogy. Consider that when a compound like Hebrew-Deity is generated from an analogy involving Hebrew-Alphabet, the head term is effectively modulated from Alphabet to Deity (see Veale *et al*. 2004; Hayes and Veale, 2005). Head modulation thus offers an alternate perspective on the generation process. Imagine that $new_{\mathbf{H}}$(X-Y) is a function that derives, via head modulation, a set of novel compound terms from an existing term X-Y. Using the operators above, $new_{\mathbf{H}}$ can be formulated as follows:

$$new_{\mathbf{H}}(\text{X-Y}) = \mathbf{C}(\{\text{X}\}, \mathbf{H}(\mathbf{U}_{\mathbf{M}}(\mathbf{M}(\mathbf{U}_{\mathbf{H}}\{\text{Y}\})))\backslash\{\text{Y}\}) \setminus \boldsymbol{L}$$

For example,

$new_{\mathbf{H}}$ (Muslim-Calendar)

$= \mathbf{C}(\{\text{Muslim}\}, \mathbf{H}(\mathbf{U}_{\mathbf{M}}(\mathbf{M}\{\text{Hebrew-Calendar, Hindu-Calendar ...}\}))) \setminus \boldsymbol{L}$

$= \mathbf{C}(\{\text{Muslim}\}, \mathbf{H}(\mathbf{U}_{\mathbf{M}}\{\text{Hebrew, Roman, Hindu, ...}\})) \setminus \boldsymbol{L}$

$= \mathbf{C}(\{\text{Muslim}\}, \mathbf{H}(\{\text{Hebrew-Alphabet, Roman-Deity, ...}\})) \setminus \boldsymbol{L}$

$= \mathbf{C}(\{\text{Muslim}\}, \{\text{Alphabet, Deity, Empire, ...}\}) \setminus \boldsymbol{L}$

$= \{\text{Muslim-Alphabet, Muslim-Deity, ...}\}$

The resulting set of speculative compounds must now be evaluated for lexico-

conceptual fitness, using the *support-set* measure described earlier. At this point we expect those compounds with the greatest fit to be the best candidates for lexical innovation and subsequent admission to the lexicon. Before taking this final step, which could potentially corrupt the lexicon, we can apply a further fitness filter by demanding that each new term be present a given number of times in a given corpus (such as the WWW). Veale *et al*. (2004) report experimental findings which suggest that the probability of finding a newly generated term in a corpus such as the WWW increases with the size of the support set for that term. The larger the support set, then, the safer it is to conclude that a lexical innovation is in fact meaningful.

If new compounds can be generated by modulating the head component of existing terms, it follows that generation can also proceed via a process of modifier modulation, whereby the modifier component of an existing term is modulated according to an implicit analogy. We can formulate modifier modulation as follows:

$$new_{\mathbf{M}}(\text{X-Y}) = \mathbf{C}(\mathbf{M}(\mathbf{U_H}(\mathbf{H}(\mathbf{U_M}\{\text{X}\})))\backslash\{\text{X}\}, \{\text{Y}\}) \setminus \boldsymbol{L}$$

For example,

$new_{\mathbf{M}}$ (Hebrew-Alphabet)

$= \mathbf{C}(\mathbf{M}(\mathbf{U_H}(\mathbf{H}(\{\text{Hebrew-Lesson, Hebrew-Calendar}\})),$

$\{\text{Alphabet}\}) \setminus \boldsymbol{L}$

$= \mathbf{C}(\mathbf{M}(\mathbf{U_H}(\{\text{Lesson, Calendar}\}), \{\text{Alphabet}\}) \setminus \boldsymbol{L}$

$= \mathbf{C}(\mathbf{M}(\{\text{German-Lesson, …, Muslim-Calendar…}\}),$

$\{\text{Alphabet}\}) \setminus \boldsymbol{L}$

$= \mathbf{C}(\{\text{German, French, Muslim, …}\}, \{\text{Alphabet}\}) \setminus \boldsymbol{L}$

$= \{\text{German-Alphabet, …, Muslim-Alphabet,...}\}$

So we speculatively create the compounds German-Alphabet because of an implicit

analogy between Hebrew-Lesson and German-Lesson, and Muslim-Alphabet because of the implicit analogy between Hebrew-Calendar and Muslim-Calendar.

Alternately, we can conceive of a generation scheme in which both the modifier and head of an existing term can be modulated simultaneously, as formulated below:

$$new_{\mathbf{MH}}\text{ (X-Y)} = \mathbf{C}(\mathbf{M}(\mathbf{U_H}\ \{Y\})\backslash\{X\}, \mathbf{H}(\mathbf{U_M}\ \{X\})\backslash\{Y\}) \setminus \boldsymbol{L}$$

This highly speculative formulation generates the cross-product of all modifiers that can apply to Y with all heads that can be modified by X. Again, many untenable combinations will be produced, but following Veale (2004), we can expect that those with a sizeable support set will be meaningful.

3.1. Experimental Support

This belief is further supported by an experiment in which 100,000 novel compounds were chosen at random from the set of all compounds that can be created via the modulation of existing WordNet compounds. These new compounds, created using the formulations of $new_{\mathbf{MH}}$, $new_{\mathbf{M}}$ and $new_{\mathbf{H}}$ given above, are grouped into different categories according to the size of their support sets; for example, compounds with a support set of 5 other compounds are organized under the category *group-5*, and so on. We can thus estimate the probability that a compound with a support set of size *n* will be validated on the WWW as that fraction of those elements of *group-n* that are so validated. In fact, there is a significant positive correlation (0.4) between *n* and the probability that an element of *group-n* will be validated via web-search. This correlation remains stable whether modifier modulation ($new_{\mathbf{M}}$), head modulation ($new_{\mathbf{H}}$) or simultaneous modulation ($new_{\mathbf{MH}}$) is used to generate the test data.

3.2. Phonetic Analogies Revisited

We are now in a better position to consider the phonetically-inspired morpheme-level analogies of section 2. In each case we can view a multi-morphemic word as a compound term, composed of morphemes rather than words, such that one of these morphemes is modulated by an implicit analogy. Consider again the analogy of (2):

(2') *astro-onomy : astro-onomer :: gastro-onomy : gastro-onomer*

Now we can view "gastronomer" as a product of head modulation, where the head morpheme "-onomy" is transformed into the morpheme "-onomer" on the basis of an analogy with astronomy:astronomer. A similar process occurs in (3'), with the additional phonetic similarity between "astro-" and "gastro-" ensuring that the modulation, which is morphologically sound, also produces a euphonious result.

(3') *astro-onomy : astro-onaut :: gastro-onomy : gastro-onaut*

Note that the modulation perspective saves us from having to rationalize a relationship between astronomy and astronaut, allowing us instead to view them as words that share a common modifier "astro-". This in turn allows us to exploit analogies like that of (4') where no such semantic relationship exists:

(4') *astro-onomy : astro-dome :: gastro-onomy : gastro-dome*

The terms "gastronaut" and "gastrodome" each have a singleton support set, corresponding to a single analogy of (3') and (4') respectively. In lieu of substantial support, however, these analogues are grounded by a phonetic similarity to their supports and this provides the requisite credibility for the new terms. That is, the

similarity between "gastro-" and "astro-" is itself a support for the new terms. Note also that order is important, yielding a bootstrapping effect as new terms are incrementally accepted into the lexicon. For instance, if (4') is processed after (2') and (3'), the support for "gastrodome" can be determined as follows.

support-set(gastro-dome)

= **C**(**M**({astro-dome}),

H({gastro-onomy, gastro-onomer, …})) \ ***L***

= **C**({astro-}, {-onomy, -onomer, -onaut}) \ ***L***

= {astronomy, astronomer, astronaut}

4. Ad-Hoc Categories as Analogical By-Products

Thus far we have considered the deliberate and explicit creation of new terms and categories whose existence is predicated on an implicit analogy (i.e., where the analogy is implicit in the workings of $new_{\mathbf{M}}$, $new_{\mathbf{H}}$ and $new_{\mathbf{MH}}$). We now consider the situation where a new lexical concept is created implicitly, as a by-product of the interpretation of an explicit analogy. For instance, consider the analogies in (9):

(9a) *Zeus : Greek :: Jupiter : Roman*

(9b) *Zeus : Greek :: ??? : Roman*

(9c) *"Zeus is the Greek Jupiter"*

The analogy of (9a) establishes an explicit mapping between Zeus and Jupiter and between Greek and Roman, suggesting that Zeus is the Greek equivalent of Jupiter. The variant in (9b) employs an elliptical form of analogy commonly found on **scholastic aptitude tests**, and requires us to provide the missing information; in effect,

it equates to the question “What or Who is the Roman Zeus”? In contrast, the variant of (9c) assumes a compressed natural language form that can also be considered a metaphoric expression (e.g., see Hutton, 1982).

The implicit relation common to (9a), (9b) and (9c) appears to be ??“deity of”: Zeus is a deity of the Greeks, while Jupiter is a deity of the Romans. However, consider the longer form of this analogy in (10):

(10) *Zeus is to Greek as*

a. *Skanda is to Hindu*

b. *Thor is to Norse*

c. *Jupiter is to Roman*

d. *Brigit is to Celtic*

e. *Donar is to Teutonic*

Each of the candidate pairings in (10) can be seen as instantiating the “deity of” relationship, so a more specialized relationship is clearly at work here. In fact, the correct relationship is “supreme deity of”, since this is the only conceptual relationship for the stem pairing that picks out just one of the five possible candidates. Now, WordNet contains the concept Deity, so one can imagine constructing the relationship “deity of” from this concept in a relatively straightforward fashion. But WordNet does not contain the concept Supreme-Deity, and for good reason: it is not a conventional collocation, and its meaning is simply a compositional function of existing terms. One of two situations must therefore hold: either the concept already exists but is not lexicalized; or else neither the concept nor its lexicalization exists prior to the analogy. In either case, we can reasonably assume that the lexical term “supreme deity” is constructed especially to resolve the analogy.

Not all such analogies require us to construct new lexical concepts. Consider the

analogy in (11), which can be seen as a close conceptual neighbor of (9a):

(11) Ares : Greek :: Mars : Roman

Here it is the relationship "war god of" that connects Ares to Greek and Mars to Roman. In this case, however, WordNet does contain the lexical concept War-God, while its lexicalization "war god" is such a conventional collocation that few would argue that it is constructed especially for the purpose of this analogy. However, this is not to say that the interpretation of (11) should be substantially different from that of (9) or (10). We can still presuppose that for each analogy, the same process is employed to construct a relational category between each concept in each pairing. In the case of (11), this relational category (War-God) will correspond to an existing lexical concept, while in (9) and (10) it will result in a lexical innovation ("supreme deity") that may be added to the lexicon following an assessment of its support set or a corpus analysis.

The construction of these relational categories raises two key questions: first, where do the component parts such as "war", "supreme" and "deity" come from; and second, why are these components, rather than others, selected? The lexicon or lexical ontology presumably plays a central role in resolving these questions, which further begs the question of what theory of the lexicon we should adopt. To remain as agnostic as possible, let us assume a rather simple, feature-theoretic view of the lexicon. Let ***F*** denote a function that maps a lexical concept onto a set of component features. Furthermore, let us assume that these features can be of one of two types. Taxonomic features, denoted with a ↑, are those that indicate the position of a concept in the lexical ontology. Associative features, denoted with a @, are those that predicate descriptive properties of the concept. For instance, consider Zeus again:

$$\boldsymbol{F}(Zeus) = \{\uparrow deity, @Greek, @supreme, @mythology, @Olympus\}$$

Thus, Zeus is a deity that is Greek and supreme, associated with both mythology and Olympus. In contrast, we can define Jupiter as follows:

$$\boldsymbol{F}(Jupiter) = \{\uparrow deity, @Roman, @supreme, @mythology, @rain\}$$

Jupiter is thus a deity that is Roman and supreme, associated with both mythology and rain (in the guise of *Jupiter Pluvius*). This feature-level decomposition suggests a means whereby new categorizations can be created for a given concept. Consider the following formulation of a function ***alt***, which derives a set of alternate categorizations for a concept by constructing alternate compositions of elements in ***F***:

$$\begin{aligned} \boldsymbol{alt}(A) \;=\; \boldsymbol{C}(\; & \{X \mid \mathbf{U_M}\{X\} \neq \{\} \\ & \qquad \wedge\, @X \in \boldsymbol{F}(A) \;\wedge\; \uparrow X \notin \boldsymbol{F}(A)\}, \\ & \{Y \mid\; \mathbf{U_H}\{Y\} \neq \{\} \wedge \uparrow Y \in \boldsymbol{F}(A)\}) \end{aligned}$$

That is, the set of alternate categorizations of A comprises just those compound terms that can be created by combining the associative features of A that have in the past been used as compound modifiers with the taxonomic features of A that have in the past been used as compound heads. The resulting compound terms are thus well-formed with respect to the lexicon and the language that it represents. Note that this formulation of ***alt*** prohibits the hypernymic terms of a concept (like ↑*deity* for Jupiter) from serving as a modifiers in any alternative categorization of the concept, since this is a combination strategy rarely seen among English compounds[1].

[1] Generally speaking, the modifier of a compound term denotes a property of the head (as in "wax paper") or a concept from which a property is transferred to the head (as in "beehive hairdo") or a

Now, a simplistic view of analogy, based on the Aristotelian account (see Hutton, 1982), might attempt to reconcile Zeus and Jupiter by seeking a common taxonomic feature (e.g., ↑deity) in both representations, but as demonstrated by (10), a genus term alone lacks discriminatory power. We need a common category that combines the Aristotelian notions of both genus *and* differentia. Given an analogical pairing A:B, we can construct this category using the function **adhoc**, formulated as follows:

$$\begin{aligned} \textbf{\textit{adhoc}}(\text{A:B}) \quad = \{\text{X-Y} \;\mid\; \text{X-Y} \in\; & (\textbf{\textit{alt}}(\text{A}) \cap \textbf{\textit{alt}}(\text{B})) \\ & \wedge \neg(\exists \text{P} \uparrow\text{P} \in \boldsymbol{F}(\text{A}) \\ & \wedge \uparrow\text{P} \in \boldsymbol{F}(\text{B}) \\ & \wedge \uparrow\text{Y} \in \boldsymbol{F}(\text{P}))\} \end{aligned}$$

Expressed in English, ***adhoc***(A:B) generates a set of compound terms X-Y such that: i) X-Y is an alternative categorization of both A and B; and ii) there is no other shared taxonomic feature of A and B (P, say) that is more specific than Y.

An analogy *A:B::C:D* is well-formed if precisely the same relationship holds between A and B and between C and D. For example, the analogy *Zeus:Hindu::Jupiter:Roman* is malformed because Zeus is not Hindu but Greek. Thus:

$$\begin{aligned} \textbf{\textit{well-formed}}(\text{A:B::C:D}) \quad = \quad & (\exists \text{M} \quad \text{M-B} \in \textbf{\textit{alt}}(\text{A}) \\ & \wedge \quad \text{M-D} \in \textbf{\textit{alt}}(\text{C})) \\ \vee \quad & (\exists \text{H} \quad \text{B-H} \in \textbf{\textit{alt}}(\text{A}) \end{aligned}$$

concept to which the head relates via slot-filling (as in "harpoon gun"). In some hybrid compounds, both the modifier and the head denote a hypernym of the compound, as in "sofa bed", but these are so rare as to be safely precluded from the current analysis. Were we to allow hybrid compounds, the formalism given here would surely over-generate; that is, precision would greatly suffer in exchange for modest gains in recall.

$$\wedge \quad \text{D-H} \in \textbf{\textit{alt}}(\text{C}))$$

$$\vee\ (\exists \text{M}_1\text{M}_2\text{H}_1\text{H}_2 \quad \text{M}_1\text{-H}_1 \in \textbf{\textit{alt}}(\text{A})$$

$$\wedge \quad \text{M}_1\text{-H}_2 \in \textbf{\textit{alt}}(\text{C})$$

$$\wedge \quad \text{M}_2\text{-H}_1 \in \textbf{\textit{alt}}(\text{B})$$

$$\wedge \quad \text{M}_2\text{-H}_2 \in \textbf{\textit{alt}}(\text{B}))$$

The first disjunct covers the situations where B and D are super-ordinates of A and C (as in the analogy *ewe:sheep::hen:chicken* where coherence is given by the relations female-sheep and female-chicken). The second disjunct covers the situations where B and D are features of A and C (as in the analogy *Athena:Greek::Ganesh:Hindu*). The third disjunct, the most complex, covers those situations where B and D are in some sense antonyms of A and C (as in *wife:prostitute::umbrella:cab*). Now, well-formedness does not always imply solvability; for that, there must exist a relationship between A and C that is mirrored between B and D. Thus, given the analogy *A:B::C:D*, we additionally expect that it has a non-empty relational basis:

$$\textbf{\textit{basis}}(\text{A:B::C:D}) = \textbf{\textit{adhoc}}(\text{A:C}) \neq \{\}$$

That is, the pairing A:C in a proportional analogy should share at least one relational category if *A:B::C:D* is to be considered a solvable analogy. As formulated above, ***basis*** may return a set containing a plurality of categories. In the case of analogies like (9a) and (11), it is sufficient that this be a non-empty set. But in the case of long-form analogies like (10), where a stem pairing must be matched with just one other in a group of candidate pairings, it may be possible that multiple candidate pairings share a

non-empty relational basis with the stem pairing. In this case, one must choose the candidate with the strongest relational basis. Since each element returned by ***basis*** is a conceptual category, we can determine the discrimination strength of each category by considering it from an extensional perspective. Given two categories in the relational basis of an analogical pairing, e.g., supreme-deity and Greek-deity, the strongest category is taken to be that which has the smallest extension (and which is thus the most discriminating). The extension of Greek-deity is larger that that of supreme-deity (108 members versus 6 members in WordNet), so we take supreme-deity to be the stronger category on which to ground an interpretation.

What of partial analogies like (9b), which form the basis of both examination questions (where a student must provide the missing information) and metaphoric allusions? In such cases, a suitable analogue must be retrieved to complete the analogy, using the available information as a retrieval cue. We can formulate a retrieval-oriented variant of ***basis*** as follows:

$$\textbf{\textit{basis}}(A{:}B{::}???{:}D) = \{X\text{-}Y \mid \exists C\ @D \in \textbf{\textit{F}}(C) \wedge X\text{-}Y \in \textbf{\textit{adhoc}}(A{:}C)\}$$

If the lexicon is sufficiently indexed, as one might expect in a structured lexical ontology, it should be relatively straightforward to identify C using D as an index.

5. Analogical Retrieval in WordNet

The comprehensive scale of WordNet as a lexical database of English word meanings, with over 100,000 lexical concepts, allows us to put the intuitions and formulations of previous sections to the test. The specific task we propose in this section is that of analogical retrieval (see Veale, 2003b; Veale, 2004): given a lexical concept in one domain, such as "Zeus", and a modifier that denotes another domain, such as "Roman", we seek to retrieve those concepts in the modifier domain that are meaningful analogies for the original head concept. The retrieval task is thus a question-answering task, in which we attempt to find answers for queries such as "Who is the Norse Zeus?" and "Who is the Hindu Athena". For balance, we shall conduct our test in two different domains of knowledge, namely deities and alphabets. The deities domain is quite well represented in WordNet, while structurally, the alphabetic domain is relatively impoverished. We shall demonstrate that the creation of ad-hoc categories that are subsequently admitted to the lexicon can significantly improve the state of these impoverished domains.

We concentrate our efforts then on the noun section of WordNet, which contains over 70,000 taxonomically organized entries. In addition to this taxonomic information, WordNet associates a textual gloss with each entry, much like that offered by a regular dictionary. For example, WordNet associates the following information with the concepts Zeus, Jupiter, Alpha and Aleph:

Zeus: *Taxonomy* = {Greek-deity is-a deity is-a god …}

Gloss = "The supreme god of ancient mythology"

Jupiter: *Taxonomy* = {Roman-deity is-a deity, is-a god…}

Gloss = "(Roman myth) supreme god of Romans"

Alpha: *Taxonomy* = {letter is-a character is-a written-symbol …}

Gloss = "the 1st letter of the Greek alphabet"

Aleph: *Taxonomy* = {letter is-a character is-a written-symbol …}

Gloss = "the 1st letter of the Hebrew alphabet"

Unfortunately, WordNet does not offer an explicitly feature-theoretic description of each lexical concept, such as that provided by our function ***F***. However, we can approximate the corresponding ***F*** for WordNet by assuming that the textual gloss of each concept is, in fact, a bag of associative features; we simply eject any non-content words (such as determiners, prepositions, and so on), and merge the resulting word set with the set of taxonomic parents that is explicitly provided by the WordNet. Thus, from WordNet we derive the following mappings for ***F***:

F(Zeus): = {↑Greek-deity ↑deity @supreme @god @ancient @mythology}

F(Jupiter): = {↑Roman-deity ↑deity @Roman @supreme @god @Romans}

F(Alpha): = {↑letter ↑character @1st @letter @Greek @alphabet}

F(Aleph): = {↑letter ↑character @1st @letter @Hebrew @alphabet}

Applying the function ***adhoc*** to these representations, we obtain the following:

adhoc(Zeus:Jupiter) = {supreme-deity}

adhoc(Alpha:Aleph) = {1st-letter, alphabet-letter}

Note that the ad-hoc concepts god-deity and letter-letter, though seemingly possible from the given values of ***F***, are not created because of the definition of ***alt*** as formulated earlier (i.e., no taxonyms as modifiers). Note also that ***adhoc*** returns two different categories for the pairing of Alpha with Aleph. In this case, based on the extension of both categories, 1st-letter is deemed the stronger of the two. In fact, an extensional analysis reveals that the extension of 1st-letter (with just two members) is a proper subset of that of alphabet-letter (with 49 members), which suggests that 1st-letter is a specialization of the category alphabet-letter.

Figure 1 illustrates the taxonomic structure of the letter domain in WordNet before any letter analogies (of the form *Alpha:Greek::???:Hebrew*) have been interpreted. Note the general paucity of organizational structure here: each letter from each alphabet is forced to share the same super-ordinate category, letter, and no attempt is made to gather letters from different alphabets under separate super-ordinates.

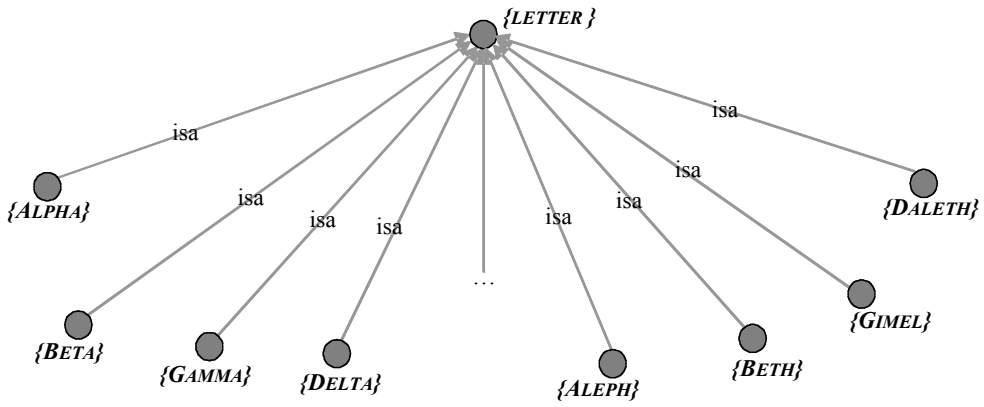

Figure 1: *The structure of the Greek and Hebrew letters domain in WordNet*

This picture changes dramatically once each letter in the Greek alphabet is placed in analogical alignment with its corresponding letter in the Hebrew domain. Note that as

the latter lacks vowels, a strict 1-to-1 alignment is not possible. Figure 2 illustrates the situation once the ***adhoc*** function has been allowed to create new lexical terms to cluster each pairing of letters under an analogically-specific category.

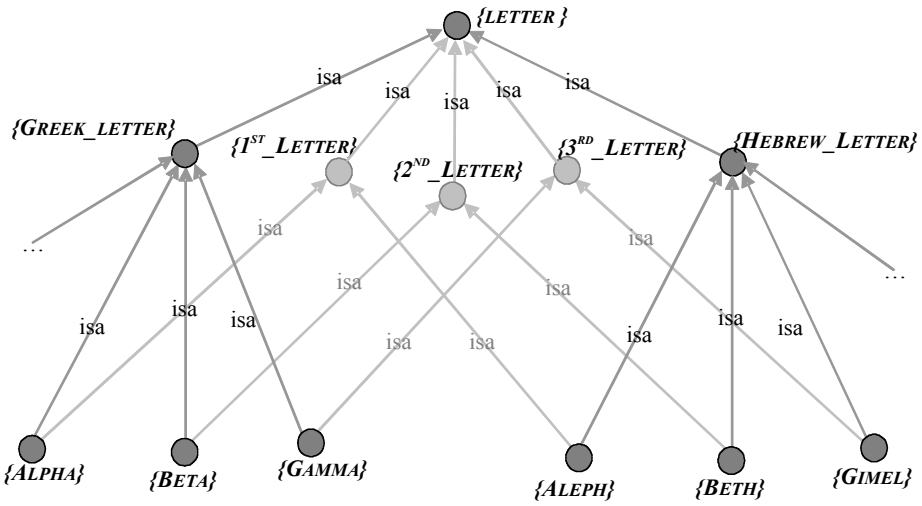

Figure 2: *WordNet supplemented with new adhoc categories like Greek-letter, Hebrew-letter and 1st-letter, created as by-products of analogical retrieval*.

5.1. Evaluation

We first consider the effectiveness of ad-hoc category construction on the precision and recall of analogical retrieval in the WordNet deities domain. Table 1 presents the results of an experiment in which analogical variants are sought for the members of five different families of deity.

Table 1: *Cross-domain variants are sought for each member of five deity pantheons*

Ad-hoc categor	*Greek*	*Roman*	*Hindu*	*Norse*	*Celtic*
supreme-deity	Zeus	Jove	Varuna	Odin	N/A
wisdom-deity	Athena	Minerva	Ganesh	*n/a*	Brigit
beauty-deity, l(	Aphrodite	Venus	Kama	Freyja	Arianrhod
sea-deity*	Poseidon	Neptune	*n/a*	*n/a*	Lir
fertility-deity	Dionysus	Ops	*n/a*	Freyr	Brigit
Queen-deity	Hera	Juno	Aditi	Hela	Ana
war-deity*	Ares	Mars	Skanda	Tyr	Morrigan
hearth-deity	Hestia	Vesta	Agni	*n/a*	Brigit
Moon-deity	Artemis	Diana	Aditi	*n/a*	*n/a*
sun-deity*	Apollo	Apollo	Rahu	*n/a*	Lug

** WordNet contains the concepts Sea-God, War-God and Sun-God*

This experiment thus involves 20 different mapping tasks (i.e., Greek to Roman deities, Hindu to Norse deities, Celtic to Greek deities, etc.). The average precision of analogical retrieval across all tasks is 93%, while the average recall is 61%.

For the letter mapping experiment, an analogous Hebrew letter was retrieved for each Greek letter, and vice versa. The ad-hoc categories created for each retrieval are of the form 1st-letter, 2nd-letter, and so on, and serve to pinpoint a precise analogue whenever one is available (that is, each ad-hoc category has an extension containing precisely two members). The precision for the letter experiment is thus 100% (that is, no retrieval errors are made). Since the Greek alphabet has more letters than the Hebrew alphabet, recall is 100% for the Hebrew to Greek task, but only 96% for the Greek to Hebrew task (since the latter has one less letter than the former).

5.2. Explicit Category Creation in WordNet

Though we have described the process of ad-hoc category creation as an implicit by-product of analogical reasoning, our formulations of ***alt*** and ***adhoc*** nonetheless allow us to exploit analogy as a deliberate mechanism of explicit category and term creation. For every lexical concept A in WordNet, we need simply consider those alternate categorizations (derivable via ***alt***) that are also generated by at least one other concept:

$$\textbf{\textit{adhoc}}(A{:}???) = \{X\text{-}Y \mid X\text{-}Y \in \textbf{\textit{alt}}(A) \wedge \exists B\ A \neq B \wedge X\text{-}Y \in \textbf{\textit{alt}}(B)\}$$

In effect, we are generating alternate categorizations of a given concept that have the analogical potential to relate that concept to at least one other in the ontology. That is, we interest ourselves here only with those alternate categorizations that possess an extension of two or more members, and which might thus make non-trivial additions to the ontological lexicon to serve a genuine organizational purpose. For example, the alternate categorization Greek-Wine constitutes a trivial addition to WordNet, since it serves to index a sole category member, Retsina. In contrast, the categorization Italian-Wine serves to index at least three different members (sweet vermouth, soave and Chianti). By this measure, Italian-Wine serves a useful indexing and clustering role in the ontology and should be retained, while Greek-Wine serves no clustering role and should be discarded[2]. In the food domain alone, WordNet provides definitions for over

[2] While Greek-Wine serves no useful clustering role, inasmuch as it serves to index just one concept, it may yet be seen as a useful addition to the ontology for reasons of symmetry. In an ontology that contains nodes like French-Wine, Italian-Wine and German-Wine, the addition of Greek-Wine, if only to index a single instance, would enhance the systematicity of the ontology (in that the Wine node

200 different terms whose gloss mentions a proper-named country like “Italy”, “Greece” or “Mexico”, so we should expect that the alternate categorizer (as formulated via ***alt***) will pick out these national ties as features to be reified.

Applying the above formulation of ***adhoc*** to the 70,000+ noun concepts in WordNet, we obtain 8564 new and non-trivial compound categories. In total, these 8564 compounds differentiate 2737 different head concepts, suggesting that each head is differentiated in three different ways on average. Overall, the most differentiating modifier is “Mexico”, which serves to differentiate 34 different heads; for example, Mexico-Dish serves to group together Taco, Burrito and Refried-beans. The most differentiated head is “herb”, which is differentiated into 134 sub-categories such as Prickly-Herb, Perennial-Herb, European-Herb, etc.

To consider just a few other domains: sports are differentiated into team sports, net sports, court sports, racket sports and ball sports (surprisingly, but not meaninglessly, Bingo becomes categorized as a Ball-Game); constellations are divided into northern and southern variations; food dishes are differentiated according to their nationalities and their ingredients, e.g., into cheese dishes, meat dishes, chicken dishes, rice dishes, and so on. As noted earlier, letters are differentiated both by culture, giving Greek letters and Hebrew letters, and by relative position, so that “Alpha” is both a 1st_letter and a Greek_letter, while “Aleph” becomes both a 1st_letter and a Hebrew_letter. Likewise, Deity is further differentiated into War_deity, Love_deity, Wine_deity, Sea_deity, Thunder_deity, Fertility_deity, and so on.

would be consistently sub-organized by Nationality). The criteria considered in this section should thus be viewed as heuristics rather than hard constraints.

One can validly ask whether such terms are truly creative, for it seems that we comprehend linguistic creativity here in its broadest sense, that used by Chomsky (1957) to describe the potential of human language to generate (i.e., create) an unlimited number of valid word combinations. It would seem that by our reckoning, then, that any novel combination of words that is syntactically and semantically valid should be considered creative. This criticism would certainly be apropos if the compounds under consideration were either entirely lexical or entirely conceptual. However, these compounds are both lexical *and* conceptual and are created relative to a lexical ontology in which they serve a useful organizational role. Compounds like "Strong-Drink" or "Love-deity" may seem mundane as linguistic artefacts in the context of general language usage, but from the context of a lexical ontology, they represent an insightful partitioning of a given conceptual space. Creativity requires clarity of perception, and the value of this insight can be seen most forcefully in the kinds of analogies that these new categories allow one to construct. For instance, the category Strong-Drink creates a cluster of diverse (but appropriate) bedfellows from espresso (strong coffee) to concentrated orange juice (strong juice) to whiskey (strong liquor). In turn, this cluster provides a firm lexico-conceptual basis for analogies of the kind whiskey:liquor::espresso:coffee.

6. Conclusions

With this paper we have attempted to provide a common formalization – in terms of lexical composition and decomposition operators – for two different perspectives on

the production of new lexical terms and categories. These perspectives are both analogically-motivated, and concern the explicit and implicit use of analogy in the implicit and explicit creation of new terms and categories. As such, it should be clear that we assign to analogy a central role in the mechanism of linguistic creativity.

Both perspectives create compound terms of the same form – simple modifier-head constructions – using lexical precedents to ensure that each term possesses both a linguistic and conceptual validity. However, while the outputs of both processes may look similar, the processes themselves are quite different, and cover different parts of the lexico-conceptual space. The implicit analogy approach, for instance, is only capable of generating compounds that can be reached via modulation from some existing support base in the lexicon. In contrast, the explicit analogy approach works directly with the feature-theoretic representation of concepts in the lexicon, and can generate compounds that, while meaningful, may have an empty support set. Each approach is thus complementary to the other, and both taken together yield a creative reach that is beyond either alone.

However, it is clear that a fusion of both perspectives does not provide full coverage of the lexico-conceptual space, even when this space is limited to that of simple modifier head constructions. Consider the terms "Gastropub" (a public house that serves restaurant-quality food), "metrosexual" (a heterosexual male with female grooming and fashion habits[3]) and "retrosexual" (a back-formation from "metrosexual" that describes the prototypical heterosexual male against which

[3] Intriguingly, the term "metrosexual" was first coined by the British journalist Mark Simpson in 1994 (see Simpson, 1994), but lay dormant for the rest of the decade. The term underwent a resurgence in popular culture when used in a New York Times article in 2003.

metrosexuality is defined as a reaction). These terms each combine a bound morpheme with a free morpheme, and while their structure is easy to analyse, it is extremely difficult to hypothesize an effective generation mechanism that does not simply combine every bound morpheme with every free morpheme in the lexicon. These compounds cannot be predicted either on the basis of existing compounds (via modulation and/or phonetic similarity[4]) or on the basis of conceptual features alone. Rather, since they are created for use in a particular communicative context, it is this context that provides the missing features that would make possible both the prediction of "Gastropub" and "Metrosexual" as valid words, as well as a broad ontological categorization of their meanings (e.g., that gastropub is a kind of public house, or that metrosexual is a type of heterosexual male).

Since portmanteau words like "Gastropub" and "Metrosexual" comprise one of the most interesting varieties of modern lexical innovation (e.g., see Veale and O'Donoghue, 2001), it would be a shame if this were all that one could conclude. Most likely, there exists a middle ground in which these terms might be, if not predictable from lexical structure, then constrained by lexical structure to the extent that the addition of automatic corpus analysis (using the WWW, say) might allow a computational system the ability to harvest such novel terms and categories as they arise in a cultural setting. As such, the analysis framework described here should provide an adequate basis for interpreting novel portmanteau words if such words could be *harvested* automatically in lieu of being *predicted* automatically. One

[4] It may be that "Gastropub" obtains some minor support from its phonetic neighbour "Gastropod" *after* it has been created using other means.

harvesting source that we are currently investigating is Wikipedia[5], an on-line open-source encyclopaedia that is constantly updated and modified by a veritable army of users. The popularity of Wikipedia makes it an ideal source from which to harvest new words as they gain prominence in the language, long before these words earn their place in conventional print dictionaries. For instance, Wikipedia offers a detailed entry for each of "Gastropub", "Metrosexual" and "Retrosexual", provides links between related terms, contains sufficient context to allow an automated system to construct an interpretation (e.g., the "Gastropub" entry mentions both public houses and gastronomy), and in some cases, even provides a pertinent analogy to explain the term (e.g., Wikipedia helpfully points out that the Gastropub, as commonly conceived, is the English equivalent of the French brasserie).

Exploring term creation in the context of resources such as WordNet and Wikipedia, which blur the traditional distinction between dictionary and encyclopaedia, constitutes an ongoing research programme that is predicated on the belief that term creation is a scaleable phenomenon through which one can explore creativity in general. That is, the issues in term creation run the gamut from phonological to conceptual, involving terms that range from the mundane to the humorous to the wildly creative. It is our hope that an understanding of the processes that underlie term creation may thus lead to a deeper understanding of creativity overall, one that can be ultimately be exploited to build computational systems that exhibit genuine linguistic inventiveness.

[5] http://www.wikipedia.org

REFERENCES

Attardo, S, Hempelmann, C. F. and Di Maio, S. (2002). Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor: International Journal of Humor Research,* 15-1, 3-46.

Baron, J. (1977). What we might know about orthographic rules. In S. Dornic (Ed.), *Attention and performance VI*. Hillsdale, NJ: Erlbaum.

Barsalou, L. (1983). Ad-Hoc Categories. *Memory and Cognition*, 11(3), 211-227.

Carpuat, M., Ngai, G. Fung, P., and Church, K.W. (2002). Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet. In the *proceedings of GWC 2002, the 1st Global WordNet conference*, Mysore, India.
Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

de Bono, E. (1994). Parallel Thinking. Viking Press: London.

Dong, Z. (1998). Knowledge Description: What, How and Who? *The Proceedings of the International Symposium on Electronic Dictionaries*, Tokyo, Japan.

Freud, S. (1905) *Der Witz und seine Beziehung zum Unbewußten*. Leipzig, Vienna: Dueticke. (Reprinted Frankfurt am Main: Fischer. English edition,1976. Jokes and their Relation to the Unconscious. Harmondsworth: Penguin.)

Hayes, J. and Veale, T. (2005). Creative discovery in the lexical validation gap. *Journal of Computer speech and Language* 19(4):513-523.

Hutton, J. (1982). *Aristotle's Poetics*. New York, NY: Norton.

Lenat, D. and Guha, R. V. (1991). *Building Large Knowledge-Based Systems*. Reading, Massachusetts: Addison Wesley.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics,* 17(4).

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11).

Rumelhart, D. E. and Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, 5, pp 1 – 28.

Simpson, M. (1994). "Here come the mirror men". *The Independent* Newspaper (London), November 15 edition, p. 22.

Trask, R. L. (1996). *Historical Linguistics*. London: Edward Arnold.

Veale, T. and O'Donoghue, D. (2001). Computation and Blending. *Cognitive Linguistics* 11(3/4), 253-281.

Veale, T. (2003a). Dynamic Type Creation in Metaphor Interpretation and Analogical Reasoning. *In the proceedings of the International Conference on Conceptual Structure, Conceptual structures for Knowledge Creation and Communication*, LNAI 2746. Berlin: Springer Verlag.

Veale, T. (2003b). The Analogical Thesaurus: An Emerging Application at the Juncture of Lexical Metaphor and Information Retrieval. *In the Proceedings of IAAI'03, the 2003 Innovative Applications of Artificial Intelligence conference*. Menlo Park, California: AAAI Press.

Veale, T. (2004). Creative Information Retrieval. *In the proceedings of CICLing'04. Lecture Notes in Computer Science*. Berlin: Springer Verlag.

Veale, T., Seco, N. and Hayes, J. (2004). Creative Discovery in Lexical Ontologies. *In the proceedings of COLING'2004, the 20th International Conference on Computational Linguistics*. Geneva, Switzerland. San Mateo, California: Morgan Kaufmann.

Veale, T. (2005). An Analogy-oriented Type Hierarchy for Linguistic Creativity. *Journal of Knowledge-Based Systems* 19(7):471-479.

Way, E. C. (1991). *Knowledge Representation and Metaphor*. Studies in Cognitive systems. Amsterdam: Kluwer Academic Publishers.