

cogPrime: A database of semantic features for investigations into word compositionality*

J. Kevin Varden

1. Introduction

This paper reports on a line of research into how cognitive abilities impinge upon a compositional semantic analysis of word meaning being conducted as part of the *Special Research Project of the Typological Investigation into Languages & Cultures of the East & West* (LACEW)¹ at Tsukuba University in Japan. The goal of this research is a cross-linguistic database application, tentatively called cogPrime, that can be used to collect and distribute data for investigations into compositional word meaning. Its initial focus is on the base semantic categories and features that are reflected by both cognitive abilities and adjectivals in human language to express characteristics of the physical world—what are referred to herein as cognitive primitives. Issues related to basic human and primate cognitive processes and how they relate to semantic compositionality will therefore also be discussed.

1.1 Searching for meaning

The semantics of language has long intrigued human beings. From the times of Aristotle² and most certainly before, what we mean by what we say, and the processes governing our choice of words, has captured our attention. One way to describe the meaning of the words we use is through comparing them to similar words in sense relations. An extension of this is to elaborate their differences by assigning them semantic features.

In most analyses (e.g. Katz & Fodor 1963³, Jackendoff 1983) semantic features of

* My thanks to Yo Matsumoto and Takashi Yoshida for comments on a draft of this article, and in particular to Yo for helping me get started in this research and clarify my thinking on many issues. All shortcomings of course remain my own.

¹ <http://www.modern.tsukuba.ac.jp/~lace/>

² See McKeon (1941) for discussion and extension of Aristotle's categories.

³ See Bolinger (1969) for a particularly insightful exposure of the weaknesses of Katz & Fodor's (1963) system. My thanks to Yo Matsumoto for pointing this out.

some type are used to delineate word relationships. A more verbose approach is the use of sentential descriptions of a word's meaning components of Wierzbicka (1985; 1996). In either case, the goal is the same—to decompose the meaning of a word into the semantic features or properties that it utilizes. The research this paper describes uses features of the first type for the same purpose—as a means to explicate the compositional nature of lexical meaning.

By initially focusing on the physical properties of static objects and spatial relationships, and incorporating studies from all cognitive research, it is hoped that language-independent observations can be made and generalized from. The initial set of features in the cogPrime database was based on a set of properties and the monomorphemic English adjectives reflecting them; entries for other languages are initially being based on this set of adjectives. The database will be extended to a more extended set of features used by other word classes, in particular verbals and nominals, so that other research into the semantic content of other word classes (e.g. Ikegami 1970; Miller & Johnson-Laird 1976) can be incorporated into the database. In this way the compositional analysis of word meaning can be tested more fully.

It is also hoped that this project will become a collaborative effort, and in the future can provide input to other studies such as the verbal study underway at LACEW.

1.2 Databasing and the internet in research

The utility of text corpora in linguistic research has widely recognized. From the simple word lists discussed in Fries & Traver (1940) to the original Brown corpus⁴ (Francis & Kuchera 1982; Suzuki 1998) to the many large corpora introduced in Church & Mercer (1994), text corpora have provided a means to systematically evaluate language (e.g. Armstrong 1994; Fellbaum 1998; Armstrong et al. 1999).

The WordNet English lexicon project⁵ has been particularly successful, so much so it has inspired similar lexicon projects for other languages (see its website for links). It has also in part inspired this project.⁶ Its large database, well-developed sense relations,

⁴ The Brown corpus is included in the corpus collection available from ICAME::; see <http://nora.hd.uib.no/icame/newcd.htm> for details.

⁵ <http://www.cogsci.princeton.edu/~wn/>

⁶ As noted by George Miller in his forward to Fellbaum (1998), one of the original designs of WordNet was to base the lexicon on semantic features of the type used in Miller & Johnson-

and inclusion of primary source examples make it a highly useful cross-disciplinary reference tool (see the papers in Fellbaum 1998 for examples). As an added plus, it is available as a downloadable application package via the Internet.

At the same time, the Internet has afforded an unprecedented ability to collect and analyze data. The culmination of this network design is possibly the software distributed by the University of California at Berkeley—the Seti program⁷ crunches Search for Extra-terrestrial Intelligence (SETI) data while acting as a screensaver on private users' home computers, and automatically reports its results back to the central server in the background when the user logs on to the Internet for their useful business. And a consortium of physicists are now working on collaborative processing of experimental data via the same internet technology primarily used by the Gnutella file sharing network⁸ for distributing copies of (often bootleg) music—a totally decentralized system allows the end-user to access information anywhere on the system grid without concern for data format or actual file location (Voss 2000).

The derivative World Wide Web (WWW) has achieved its own special status since its release in 1991⁹. Now users with internet access can use web browsers to conduct research, share data, and access and contribute to databases. Several successful examples of use of the web for research collaboration include the Child Language Data Exchange System (CHILDES)¹⁰, the Human Genome Project¹¹ and Raf Alvarado's Mayan Epigraphic Database Project (MED)¹².

The need for large bodies of data with which to test semantic theory of all types was noted in Weinreich (1972), and echoed in Lipka (1997). Unfortunately, it appears that there is still no large database of semantic features for use with testing compositional analyses available on the internet or web with general access.

Laird (1976). That idea was abandoned as undoable by Miller and Johnson-Laird due to implementational difficulties. The database described herein is the first step in an attempt to take up that task.

⁷ <http://setiathome.ssl.berkeley.edu/>

⁸ <http://gnutella.wego.com/>

⁹ <http://public.web.cern.ch/Public/ACHIEVEMENTS/web.html>

¹⁰ <http://childes.psy.cmu.edu/>; <http://jchat.sccs.chukyo-u.ac.jp/CHILDES/>

¹¹ http://www.ornl.gov/TechResources/Human_Genome/

¹² <http://jefferson.village.virginia.edu/med/medwww.html>; see Schwimmer (1996) for details

1.3 The cogPrime database

In an attempt to provide the linguistic community with such a readily accessible body of such data, the cogPrime database¹³ of adjectives and their related semantic features is being developed in research supported by LACEW at Tsukuba University in Japan. The database has been motivated—and hopefully in part justified—by the success of the aforementioned collaborative web projects. By making the database available as both a downloadable application, as WordNet is, and accessible via the WWW, it is hoped that those interested in semantic decomposition can participate in both data collection and the direction that the database structure takes.¹⁴ Through providing the means to search and export data as text files, it is also hoped that the collected data can contribute in some small way to other projects such as the verb use database underway at LACEW.

In particular, the advances in neuroscience of the past few decades make this enterprise seem particularly worth doing. Neuroscience has evolved from a peripheral discipline to a central component of almost all other psychological and physiological disciplines (Kandel & Squire 2000); the literature abounds with results that impinge upon semantics either directly or indirectly. One of the cogPrime goals is to explicate this connection more fully.

It is, of course, impossible to identify exactly which semantic features a speaker possesses in their repertoire. Despite intriguing indications that memory is a function of certain genes controlling patterns of synaptic firing (Lisman & Fallon 1999), we still have no means of demarcating a particular thought, let alone one small component of such a thought. However, among our available resources, the constraints cognition places on our interpretation of the world would seem to be a good point of departure for such an endeavor.

1.4 Background assumptions and scope

This paper will follow P. Bloom (2000), among his many predecessors, that cognitive development is not dependent on language development; that the child has a rich mental life before the onset of language; and that cognitive development would develop in

¹³ <http://www.cogprime.com>

¹⁴ Please address requests for information to info@cogprime.com; all comments, criticisms, etc. should be sent to comments@cogprime.com.

some fashion even in the total absence of linguistic stimuli. This is inherent in L. Bloom's (1993) *principle of relevance*: "Words are learned when they are relevant to what the child has in mind." P. Bloom attributes Fodor (1975) with saying, eloquently, that all language learning is actually *second-language* learning—a labeling of knowledge already present.¹⁵

It is also taken for granted that at least a significant portion of semantic analysis is compositional; i.e., the meaning of the words that we use in our languages is composed of some finite number of 'features', or facts that we know about our world (Bloom 2000, ch. 6; Jackendoff 1996: 545). This and other considerations of a compositional analysis of word meaning will be discussed further in the next section. Discussion of other semantic frameworks such as prototype theory (Rosch 1975a; Armstrong 1983), semantic fields (e.g. Lyons 1977), frames (e.g. Fillmore 1982¹⁶), and generative formalizations (Pustejovsky et al. 1994; Pustejovsky 1995), while crucial to understanding the meaning of words and larger constituents, are placed outside the scope of this paper.

1.5 Paper organization

The organization of the paper is as follows. §2 discusses issues related to cognition and a compositional analysis of word meaning. §3 discusses issues related to semantic features and categories in this analysis, while §4 details the structure of the database and its contents. Finally, §5 summarizes the paper and outlines future directions.

2. Theoretical background

Again, it is being taken that language involves the attachment of labels to cognitive constructs that already exist. Perhaps the language being learned plays a more central role after initial labeling has occurred, but since language acquisition occurs in tandem with—and is necessarily dependent on—cognitive perception and development, it seems appropriate to begin the discussion by taking a look at the cognitive domain.

¹⁵ See, e.g., Fodor 1975: 58-59 ftnt 4.

¹⁶ The FrameNet web site is accessible at: <http://www.icsi.berkeley.edu/~framenet/>; thanks to Takashi Yoshida for pointing this out.

2.1 The world as a cognitive construct

As discussed in Jackendoff (1983), language and language use reflects the projected world, not the physical world. By projected world it is meant the virtual reality construct that we all experience the physical world through (see Jackendoff 1985 for discussion of I-semantics).

This projected world gains support from the fields of music and vision. In both cases, our brain manipulates and organizes incoming sensory data to fit the cognitive system we use to interpret it. In regard to visual phenomena, most everyone is familiar with popular optical illusions. In particular the Gestalt school's black and white image of what can be interpreted as either a vase or as two opposing silhouettes (see Wade 1982 §1.2 for examples and discussion) is a striking example of how we project objects into our conceptual space based on how we interpret incoming visual stimuli (Jackendoff 1983: 24). Although the visual sensory data does not change, the brain can reconstruct either of the two images (the vase, or the two faces) with equal facility, and shift easily between the two—but not view both images at once. When we interpret the image as that of two faces, the vase disappears; when we see the vase, the faces cease to exist. Our brain chooses between the two, and that becomes our perceived reality.

In addition, studies of imagery have indicated that the same basic magnitude judgments are made when conjuring up images as when viewing real objects. For example, participants were asked to image various sounds (e.g. “the barking of a large dog”; “a car horn honking”) and then judge their relative loudness. Their responses followed the same basic response pattern as other studies where participants actually listened to the sounds (Baird & Harder 2000). This held true for judgments about the magnitudes of images with components of light, sound and smell.

Also noted in Jackendoff, what we call ‘music’ does not exist as a systematic means of experiencing sound other than within the listener's mind; Beethoven's 5th is not a physical construct, nor does it ‘exist’ as a collection of blank ink marks on a sheet of staffed paper. In addition, we do not hear frequencies of sound in a linear fashion, as might be produced by someone steadily turning the frequency dial of a pitch generator; when listening to music we hear pitch adjusted for the context the sound is encountered in. Pitch judgments—even among musicians with supposedly perfect pitch—can be confounded by a number of factors, including but not limited to strength of harmonic

components and, surprisingly, cyclic hormonal fluctuations in both sexes (van den Brink 1982).

Perhaps less well-known but stronger evidence comes from vision studies related to the perceptual inversion. As with any single lens system, the images formed from visual stimuli entering the eye are inverted and reversed. Most people have probably seen this in an explanation of a pin-hole box camera. The world, according to the light striking the back of our eyes, is upside-down and backward with respect to our tactile experience.

Of course, we do not 'see' the world as inverted, or reversed. The correction of visual stimuli is hard-wired into the mammalian (and, presumably, animal) visual processing system via the cross-over of neurons on their path along the optic nerve (Beazley et al. 1995). Evidence for the brain's capacity to manipulate the composite incoming visual stimulus, however, comes from the inversion studies whereby participants wear prism eyeglasses that shift, invert or reverse vision. Welch (1978) reports on several such studies where participants have been fitted with prism eyeglasses that totally invert and reverse the incoming sensory data. These subjects go through a period where they see the world as upside-down and reversed, with accompanying nausea and lack of functional facility. In some of these studies, participants have reported no correction of the inverted visual field (see also Linden 1999). However, in several studies, especially those where participants were engaged in physically and mentally demanding activities such as mountain climbing, skiing and riding bicycles in heavy traffic, adaptation by full re-inversion of the perceived world was reported (see Welch 1978: 111 and references therein). After removing the glasses, the participants again went through the disorientation of the world inverting and reversing, and once again their brain had to adjust its construction of the mental projection. These studies of adaptation to altered visual input show quite strongly that the reality we 'see' exists 'in our head'.

One final piece of telling evidence of a mental construction of our projected world lies in its 3-dimensional (3D) nature. Although the back of the eye is concave, resulting in some minute 3D structure to the light-sensing cells, the images it receives are basically two-dimensional. However, our brain either knows innately or learns quite early that the 'objects' that we can sense by touch are distinct, and are separated in the

physical world's spatial plane (Kellman & Spelke 1983; Spelke 1994). The 3D world we experience with our sight and touch is the result.

2.2 Universal cognitive processing

Many semanticists (e.g. Bierwisch 1967; Katz 1972) have maintained that humans have a universally innate mechanism for interpreting the physical environment; this has been upheld by studies of cognitive development and abilities of mammals and human infants (for discussion see Locke 1993 ch. 7 and references therein; Flint 1999; also the *Cognitive Science* vol. 24 issue on primate cognition, esp. Byrne 2000). In one striking example of the neonate's innate sensitivity to social interaction, Meltzoff & Moore (1977) showed that human infants will imitate facial gestures such as tongue protrusion within minutes of being born. Since this occurs with infants who are for the first time encountering an unmasked face, the gestural responses associated with highly social facial expressions such as smiling, tongue protrusion, etc. must be innate. In the pioneering study of Kuhl & Miller (1978), chinchillas were shown to be able to categorically discriminate speech sounds just as humans. Other studies (e.g. Doupe & Kuhl 1999) show that birds, non-primate animals, primates and even insects share this ability to categorically identify their species' sounds and prosody. More recently, both cotton-top tamarin monkeys and human infants were shown to be able to discriminate between Dutch and Japanese sentences if the sentences were played forwards, but not if played backwards—when played backwards, the speech prosody information they were evidently attending to is lost (Ramus et al. 2000). These same cotton-top tamarins—just as human infants—have been shown to be able to discriminate pseudo-word boundaries using only syllable frequency information in the speech stream (Hauser et al. 2001; see the discussion of statistical learning below in §2.6).

Other research has shown that these cognitive abilities are multi-modal; sensory input from several sources often combines to perceive a specific event or perform task (Bushara et al. 2001). Graziano et al. (2000) report that the sensory integration used in proprioception is accomplished by one specific region of the monkey brain; neurons in this region fire fastest when a monkey sees a realistic false monkey arm on the table in a position that matches its own. Assumedly this is a function of the primate brain.

These studies and the many others like them suggest cognitive processing abilities are a results of our evolutionary heritage. As such some of these innate abilities are

shared by primates, some by all mammals, and others possibly by all sensing creatures.

2.3 Universal linguistic processing

What of language? If cognitive processes can be innate, what of linguistic processes?

Innate linguistic universals are often taken for granted by generative linguists; indeed, it is a principle underlying concept of the entire Principles & Parameters framework (Chomsky 1981). Generally, syntactic examples such as the **Coordinate Structure Constraint** (Ross 1967) are used to support the innate nature of grammar (e.g. Newmeyer 1986: 73-74). Perhaps stronger evidence for the universality of basic linguistic processing comes from animal studies. They show that other mammalian species have the ability to categorically perceive sound, just as we do, as discussed above. The evolutionary result of the categorical sound processing appears to be human infants as young as six months having the ability to divide their language's vowel space up into what are known as 'perceptual magnets'—idealized variants of their language's vowels that allow them to deal with the wide-ranging speaker variation that they encounter (Kuhl et al. 1992; Kuhl & Meltzoff 1995). In addition, it has been shown that dolphins are sensitive to syntactic differences of commands they have been taught (Herman et al. 1993; Herman & Uyeyama 1999). It appears more and more that what is special about human language is not the uniqueness of the processes that it uses, but the extent of the facility of and interaction between those processes.

There is also a good deal of evidence that language acquisition is multi-modal, just as general cognitive development is. For example, studies have shown that speakers make active use of lip reading and gestures in discourse. In what is known as the McGurk effect, listeners presented with video of someone pronouncing [ba] accompanied by the utterance [ga] consistently report hearing neither—they blend the visual and aural input together and 'hear' [da] (McGurk & MacDonald 1976). Similar conflation of visual and aural stimuli has been recently extended to infants of 5 mos. by Rosenblum et al. (1997).

Adjustments have also been reported for speakers who were presented with altered but still perceptible feedback from their ongoing whispered speech¹⁷. These speakers

¹⁷ Whispered speech was used to avoid the conduction of lower frequencies by bone that are common in voiced speech.

adjusted the formant structure of their vowels to compensate for the distortion, evidently at the phonemic level (Houde & Jordan 1998). In addition, Gracco et al. (1994) and Kawahara (1993) show that auditory feedback of our own voice is used by the motor speech processing centers during speech production.

These and other studies like them make it clear that language shares many of the cognitive abilities used for other tasks and by lower animals.

2.4 Linguistic and cultural specifics

On the other hand, it cannot be said that all cognitive processes are innate, or that all language use is based solely on physical interaction with the world. The cultural and linguistic component of our cognitive experience is hard to deny.

The embodiment of this effect of language and/or culture on cognitive processes is known as linguistic determinism¹⁸: that our language determines our thought processes to some extent. In its strong version, our language defines our cognitive processing of the environment so that we are aware of only what we have linguistic means of describing; in its weak version, language influences cognitive processing to some degree.

Perhaps the best-known studies involving linguistic and or cultural effect on cognitive processing are color categorization studies inspired by Berlin & Kay's (1969) landmark work. However, although a number of researchers tried to show color naming as reflective of ability to perceive colors, this view could not be maintained. Even the Dani, a stone-age culture whose language contains only two lexicalized color terms, pattern with those from language groups using more complex color terminology in being able to learn and recall novel color terms (Heider [Rosch] 1972; Rosch 1975b; see Dedrick 1998 for discussion).

A more intriguing group of studies showing what appears to be a genuine effect of linguistic background on perception are studies involving tritones and ambiguous notes. A tritone is formed by simultaneously playing two notes that are exactly one-half octave apart (e.g. C and F#). Ambiguous notes are constructed from the same note

¹⁸ This is popularly known as the (Sapir-)Whorf Hypothesis (Sapir 1949; Whorf 1956), but Whorf did not actually formulate any hypothesis in his work. See the Dan Alford's excellent comments at <http://listserv.linguistlist.org/cgi-bin/wa?A2=ind9508D&L=linguist&P=R1818>;

played at multiple octaves simultaneously; e.g. an ambiguous C can be formed by simultaneously playing all the C notes on a keyboard (C1, C2, C3, etc.). Playing ambiguous tones in a tritone relationship in sequence (e.g. an ambiguous C followed by an ambiguous F#) has been used quite successfully to judge listeners perception of pitch. When such a sequence is played, listeners will judge one ambiguous note higher than the other, even though 'higher' or 'lower' pitch has no real meaning when comparing ambiguous tones—the C4 component of the ambiguous C, for example, is higher than the F#3 component of the ambiguous F#, but the C3 component is lower than the F#4. Further, they often will judge one ambiguous tritone as higher than another in one context, and then in another context make the opposite judgment (Deutsch 1987). What is so intriguing is that whether listeners judge one ambiguous tone higher or lower than the other is sensitive to the listener's language experience—listeners from California tended to judge tritone pairs in an opposite manner than listeners from the south of England (Deutsch 1992: 74). This is assumed to be due to their exposure to their language's prosody during language acquisition. Results of these studies with Vietnamese bilinguals in California shows that this early exposure effect holds even when the listeners no longer can speak their first language fluently (Deutsch 2000).

Results of studies on arithmetic ability in Russian-English bilinguals speak stronger to effects of language ability and use on cognitive organization, in this case to cognitive L1/L2 separation. The reaction times of Russian-English bilinguals were tested performing simple arithmetic tasks such as adding two single-digit numbers; their reaction times in initial testing were similar when doing the problems in either language. However, after practicing in one language and then being presented with new problems in both languages, the reaction times for problems done in the language they practiced in were significantly faster than the reaction times in the language they did not practice in. fMRI¹⁹ brain scans indicate this to be due to the heavy use of language-specific processing centers of the brain during this type of simple arithmetic—the numerals are evidently treated as words, and therefore processed by the language centers (Dehaene et al. 1999; but see Brysbaert et al. 1998 for further discussion). The

see also Brysbaert et al. (2000) for newer directions in 'Whorfian' research.

¹⁹ functional Magnetic Resonance Images

fact that performance suffered when attending to a task in the language the participants were not trained in speaks to cognitive separation of the processing stages of the two languages within each participant.

Although the results of these studies are fascinating, the fact remains that incontrovertible experimental evidence showing language defining cognitive ability has proven difficult to provide. The problem in studies involving linguistic determinism effects is that it is seemingly impossible to find two groups of speakers who are different only with respect to their language; language groups invariably carry their cultural baggage along with them making it difficult if not impossible to separate the effects of cultural markedness and habit (i.e. whether or not it is taboo to speak of a certain object or process) from the effects of the languages' structure. It is also extremely difficult to judge cognitive ability based on language use; language use—by definition—is use, not ability.

However, just such a group of subjects has been purported to be found by Peterson & Siegal (1995). A group comprised of 25 young deaf children of hearing parents and one young deaf child of deaf parents, ages 8-13, all residing in Australia, were presented with a task that tests for the ability to reason abstractly, the so-called Sally-Anne task. In this task, a researcher acts out a character named Sally hiding a marble in a box. She then leaves the room, and the researcher uses another character named Anne to re-hide the marble in a basket. The children are then asked where Sally will look for the marble when she comes back in the room. The only children to regularly fail this test—i.e. answer incorrectly that the doll will look in the basket, not in the box—are autistic children and those under four years of age (see Peterson & Siegal 1995 for details; but see Bloom & German 2000 for limitations of the Sally-Anne task).

An intriguing difference shows up in deaf children's responses. The one deaf child of deaf parents in their study patterned with the normally developing hearing children by responding that the doll will look in the box where it hid the marble. However, roughly two-thirds of the deaf children of hearing parents patterned with the young and autistic children; they responded that the doll will look in the basket, since that is where the child saw the marble being moved to. This is despite the fact that all of the children in this study were at least 8 years old, and that by the age of 4 years hearing children have already mastered this task.

This difference is assumed to be due to the hearing parent's general lack of ability to discuss abstract concepts using what limited sign ability they possess, since the children in this study were all rigorously tested and showed otherwise normal-range intelligence and no sign of autism. Deaf parents, on the other hand, being fluent in sign, have no such problems discussing abstract concepts with their young children. This difference in ability among the children to extrapolate abstractly, then, appears to be a highly language-dependent artifact of their linguistic experience—young deaf children of hearing parents have no chance to participate in discussions requiring this abstract reasoning, and so tend not to develop at least this specific capacity for abstract extension. Russell et al. (1998) extended the age range of deaf children tested to 16, and found that a significantly higher proportion of the children above 13 did pass the test, consistent with an analysis that this difference in cognitive ability is a delay in development.

This paper, then, acknowledges a weak version of linguistic determinism, the 'linguistic relativism' discussed in Brysbaert et al. (1998). That is, our cultural and language experience is seen to influence the way that we interpret and categorize incoming sensory data by being the vehicle that draws our attention to certain aspects of that data. One of the stated goals of the database being developed is to provide data to help delineate, in addition to the similarities of languages due to innate abilities, the differences due to cultural and/or linguistic experience.

2.5 Compositional language acquisition

Upon reflection, the whole of language acquisition seems to be rooted in one sort of compositionality or another. One of the child's first tasks as a language learner is to learn to extract component words out of the ongoing speech stream they are exposed to. In addition to experimenting with the sounds and intonational contours of their language, children begin to collect words in order to be able to converse as those about them do. Once enough words have been collected, children can begin to see the patterning in their language that allows them to acquire the syntax of their language through the semantic relations of words, the so-called 'semantic bootstrapping' effect of Pinker (1987). Once their syntax has begun to develop, they also appear to use syntactic position as a means of delineating a word's meaning using 'syntactic bootstrapping' as

per Fisher et al. (1994); see also Pinker (1994) in the same work.

In generative phonology children's acquisition of phonological features—distinct but coordinated articulatory gestures that produce distinct acoustic goals (Clements 1985; Sagey 1986; Clements & Hume 1995)—and their language's system for combining them—is crucial for the development of a fluent speaker. Compositionality is also reflected in their learning of the possible combinations of segments into syllables (Clements & Keyser 1983; Blevins 1995) and words into intonational groups (Selkirk 1995; Vihman 1996).

In acquiring semantics, children also steadily increase their store of morphological components, and learn which of them are productive and which are not (see Clark 1998 for discussion). They must also learn the use of idiomatic speech, which, contrary to most cursory analyses, Nunberg et al. (1994) has shown to be generally compositional in nature. Pitt & Katz (2000) also argue for a fully compositional analysis of what they term *compositional idioms* (e.g. 'plastic flower') that necessarily involve decomposing the lexical items that make up the idiom. And, as noted in Jackendoff (1996), the success of such compositional analyses for word meaning itself is encouraging (e.g. Katz & Fodor 1963; Ikegami 1970; Miller & Johnson-Laird 1976).

2.6 Statistical language acquisition

This is not, of course, to say that all language acquisition is compositional; our language and our cognitive nature in general tends toward habits—learned combinational responses. There is also a good deal of recent research showing that connectionism and statistical learning²⁰ are heavily relied on by the language learner and user, as is discussed in Seidenberg (1997). Indeed, statistical learning is the basis for the whole Optimality Theory enterprise (Prince & Smolensky 1993; Kager 1999; Tesar & Smolensky 2000). Sensitivity to frequency distribution has been demonstrated for human visual discrimination of letters, syllables and words, aural discrimination of phonemes, phoneme sequences, syllables, word boundaries, and prosodic structure in general (Kelly & Martin 1994 and refs. therein). Discrimination of word boundaries by means of syllable frequency in the speech stream has now even been demonstrated for cotton-top tamarin monkeys (Hauser et al. 20001). Further, it is particularly telling that

no generative method of language recognition can yet come close to matching the success of methods based on statistical techniques (see the discussion of stochastic vs. knowledge-based systems in Church & Mercer 1994).

The popular notions of the generative enterprise—in particular, that language learning is best modeled by the infant identifying its grammar from among the myriad possible grammars—needs to be revised and moderated in view of this research. However, although a child evidently breaks into its language via frequency and context of word use, and adults continue to make active use of these processes, an innate process directing the assignment of features to word meaning is just as much a part of the process—once a word has been identified by whatever means, meaning must still be attached to it through some mechanism, either holistic or compositional. Testing a compositional analysis of word meaning therefore remains a valid line of investigation.

3. Semantic features

Having said that a compositional semantic analysis—i.e., an analysis of word meaning based on ‘semantic features’—is justified as part of the semantic enterprise, it is then necessary to define exactly what is meant by ‘semantic feature’.

3.1 Features as cognitive constructs

The term semantic feature is used here as it is in most works, whether explicitly stated or not—a semantic feature is a characteristic of some object or action that we have discerned in our environment, or have synthesized on the basis of our experience. Fodor (1975: 55ff) argues that such a core set of primitive concepts must be innate (i.e. present at birth) for learning to begin. The ability to discern the features of objects we observe appears to be inherent in cognitive capabilities (e.g. the processing of 3D visual input reported in Jansen et al. 2000 and references therein). Indeed, the research discussed in Landau (1994) indicates that objecthood and spatial location are based on separate cognitive abilities, and hence consistently show similar encoding in human languages. Other features appear to be learned during experiencing the world. Schyns & Rodet (1995) and Schyns et al. (in press) discuss formation of features & categories,

²⁰ See <http://www.bcs.rochester.edu/bcs/research/LIS/lis.frameset.html>.

and show that feature-learning is necessary for at least some conditions such as determining the constituent parts of novel shapes (see also Schyns & Murphy 1994).

3.2 Defining features & features categories

The methodological approach initially used here was to base an initial set of primitive semantic features on objects' physical properties. Although the current set of features detailed below in §4.2.4 will surely be modified as research into the literature proceeds, these features have, for the time being, been assigned to the feature categories below, roughly based on sensory modality (cf. the noun 'domains' in Niida 1975: 178ff).

To begin, the physical characteristics of readily available objects in the accessible environment were considered. For each physical property discerned, a semantic feature was assigned; e.g. for a book sitting on a table features were assigned for length, width, thickness, weight, density, color, brightness, etc. Once a list of features was gathered, they were placed into categories according to sensory mode. However, rather than repeat features in more than one modal category as an exhaustive analysis would require (e.g. the length of an object can be sensed either by sight or by touch), the categories themselves gradually evolved to allow what seemed a more intuitive grouping (e.g. length and other dimensional features have been assigned to the category **size**). As noted before, this is consistent with the sensory conflation that permeates much of our perception.

3.3 Concrete features and cognitive primes

The original set of features being used are the subset of those features that identify physical properties of concrete objects, hereafter referred to as **concrete features**. This set of concrete features is thought to be further distilled into a set of **cognitive primes**, features that universally are assigned due to their being a direct consequence of an inherent cognitive processing constraint. As an example, the objecthood of an object has shown to be a part of even very young infants' mental representations (Kellman & Spelke 1983; Spelke 1994).

It is intended that such cognitive primes be identified and marked in the database as progress continues—hence the name of the database application. It is precisely these features that should show the greatest representation cross-linguistically; e.g. the shape

and location terms discussed in Landau (1994).

3.4. State and scale characteristics

It is taken in this research that semantic features of adjectives are generally reflective of cognitive use of relational scales (Bierwisch 1967; Miller & Johnson-Laird 1976: 324ff). That is, an adjective such as *tall* requires access of some scale of height, extending from the earth to what the language user knows is a greater-than-average value for things of that type.

Three categories have been used for the organization of the cogPrime feature set: **state features** and **scalar features**, with scalar features being further divided into **open** scalar features and **bounded** scalar features (cf. Nöth 1997).

3.4.1 State features

The term **state feature** is used herein to denote a feature that is assigned a unitary value as opposed to a gradated one. These features reflect adjective use more than meaning, and are used as meta-language to track adjectival characteristics.

A good example of this is the feature **requiresObject**; a positive value for this feature is used to signify that the adjective in question can only be used to modify concrete objects; e.g. the English words *light* and *heavy* as used in reference to weight. If one excludes such extensional phrases as “a heavy thought”, all adjectives referring to the physical property of weight will necessarily be marked as **requiresObject**. In contrast, adjectives such as *long*, which are used to modify words of temporal duration as well as physical length, will not require this feature.

3.4.2 Scalar features

In contrast, there is often justification for using spectra rather than state features. Many things in our environment are graded—weight, heat, light, etc. can all be discussed in terms of values along a scalar continuum (see Bierwisch 1987; Nöth 1997 for discussion). Gradability of scales requires demarcation of at least two properties, **endedness** and **polarity**.

Endedness refers to whether a scale is **open** or **bounded**; i.e. whether or not a particular scale terminates at some finite value. A good example of a **bounded scale**

would be pitch as it pertains to human perception, not the actual frequency produced by a sound source. Human perception of pitch only extends so far; above the average limit of 10k, a sound source has no pitch that humans can perceive. Pitch is therefore a bounded scale that ends with a finite value.

An **open scale**, in contrast, continues indefinitely without terminating at a finite value. The height of an object as we measure it ranges between zero and as far as our ability to measure (or imagine) the object. Again, there are no negative values of height assigned in normal environmental interaction.

Another distinction that has been made is the **polarity** of a scale; i.e. the difference between **monopolar scales** and **bipolar scales**²¹. **Monopolar scales** are graded scales that have a starting point at some value (e.g. zero) and continue on in one direction either indefinitely or until some specific value. A good example of a monopolar scale is weight; in its common usage, the weight of objects begins at zero (something that is weightless) and continues on to higher and higher values, assumedly until infinity. There is nothing in our environment that has a negative weight; there is nothing in our environment that tells us that weight need be quantified (i.e. given values of only whole numbers of lbs.) according to some rule. Our ability to grade objects by weight is only limited by ability to gradate our measurements. Weight is therefore a prototypical example of a monopolar open scale.

Bipolar scales are those that have values continuing in both directions from the human point of view. The scale associated with position along the vertical axis from the referent's point of view, labeled **superiority** in this database, has a middle, neutral value associated with objects at the same level as the experiencer. Objects below the viewpoint of the experiencer are said to be low ([-superior]) while those that are above it are said to be high ([+superior]).

3.4.3 Other considerations

There are several other considerations regarding scales worth mentioning. The first of these is **cognitive vs. scientific coding** of scales. As discussed in Nöth (1997), a good case in point is viewing heat from either the cognitive or the scientific point of view.

Hot, warm, cool and cold are all intuitively connected: they are all measures of heat

transfer. In physics all measures of heat are measures of the motion of molecules in some degree above the temperature at which there is no molecular movement (0° Kelvin). As the object warms, the molecular activity increases. Heat, in its scientific sense, is a monopolar, open-ended scale extending from zero to some relative positive limit where an object's chemical bonds have been broken and it ceases to exist. In addition, research has indicated that all temperature differences are sensed with the same free nerve endings; no specific temperature receptors have yet been found (Coren et al. 1978; Sherrick & Cholewiak 1986). The difference in sensation lies in how our body and brain interpret the signals (Craig et al. 1996).

However, this is not in line with how humans experience their environment. Disregarding our cultural affinity for the Fahrenheit and Celsius scales that situate 0° firmly in the center of human experience, we physically experience heat when an external object is more energetic (i.e. hot) than the part of our body we are sensing the object with (known as physiological zero)—if something is at a slightly higher temperature than our skin when we touch it, we will sense warmth; when it is at a much higher temperature, it will feel hot. Conversely, we experience cool, cold, etc. when something is at a lower temperature than our skin. Although being burned by dry ice and fire may both produce a burning sensation and tissue damage, there seems to be a fundamental cognitive difference between them. This experiential difference seems strong enough to justify positing two monopolar scales as opposed to one bipolar scale—i.e. one scale for heat and one scale for coldness—just as Aristotle did long ago.

The second consideration being set aside for now is **scale valency**. There are two such types of bipolar scales associated with antonyms, and hence with the features they represent. In scalar measures such as heat, there is an intermediate region between hot and cold, signified in English with the word tepid or the phrase lukewarm. Scales such as this are trivalent; they have three distinct regions. For other scales, there is no such intermediate region. An example is the scale represented by the English words open and closed. There is no intermediate region between these two states; this scale and others like it are bivalent in nature. The number of valence states, and how to encode them into the database, will be left for future work.

In addition to the valency, another characteristic of scales will need to be dealt with

²¹ unidirectional and bi-directional in Nöth (1997)

at some point: that of **grading**. The phenomenon of grading is familiar to many because of Labov's (1973) landmark study of participants' judgments of cup-like containers. Participants were shown various pictures of containers, and were asked to name them. Depending on its number of handles, its height/width ratio, and even its contents, at some point a 'cup' could just as equally well be called a 'bowl'; a 'cup' might also be called a 'vase', etc. In order to separate such word pairs as the English *warm* and *hot*, *big* and *small*, some sort of marking for degree will need to be encoded into the database. The exact nature of this gradation marking, as with the related valency, are left for future work.

Another consideration that is being put aside for now, but that is no less important than any other, is the role of **speaker orientation** with respect to the object being observed (Zubin & Choi 1984; see also Miller & Johnson-Laird 1976: 394ff for discussion). To say that "object A is in front of object B" is inherently ambiguous in English. Object A could be located in the region considered to be the front of object B. This is termed the **gestalt reading**. However, Object B could be located in the region lying between the speaker and object B. This is termed the **orientation reading**.

For example, in the phrase "the book in front of the chair" the book could be lying on the floor in front of the chair where someone sitting normally puts their feet, regardless of where the speaker is (the gestalt reading). Alternatively, the book could be lying on the floor behind the chair, if the speaker is also standing behind the chair (the orientation reading). While the two readings are normally conflated for positional terms, many languages distinguish readings with different lexical items or phrases. For example, Korean uses one lexical item *kupulin* for objects that are inherently crooked (i.e. a crooked stick), and another lexical item *pittulin* for items that are not straight in reference to their context (i.e. a pole leaning away from the vertical—Zubin & Choi 1984: 336). Since this difference is lexically encoded in languages, it will need to be addressed in the database at some point.

In summation, the following three types of features are currently used in the study: state features, monopolar scalar features, and bipolar scalar features. The exact mechanisms for marking speaker orientation, scale valency and scale gradiency is left for future work. The next section will turn to a description of the database application,

and the data being collected.

4. Database structure & contents

4.1 Database structure

As an application, the cogPrime database currently consists of three browsers—an Overview browser, a Features browser, and a Language browser. The Overview browser displays information for all entered languages. Searches can be made for a particular text entry (adjective, gloss, and if applicable, language-specific script or Chinese character), a particular meta-feature (see §4.2.4 below), or assigned semantic features. Matching results for all languages are then displayed in the main Overview browser. The user can switch between adjective, gloss, script or kanji views; this proves to be quite useful when comparing terminology or feature use between languages. By clicking on a button for a given language, the subset of adjective records corresponding to the search can be accessed in the Language browser for more detailed study.

In addition, languages can be accessed directly via a Language menu listing the languages contained in the database. Searches can be made on individual languages in the same manner as described above. Navigation among records for a given language can be achieved via scrolling through the records, jumping to a specific record, or selecting an entry from a scrollable list.

Similarly, the database of semantic features can be perused with the Features browser, navigated through, and searched upon in the same manner as described above. On both the Features and Language browsers, semantic features are displayed in a hierarchical list of feature categories and features contained within each category. These categories and their features are detailed below in §4.2.

The web version of the database currently consists of similar Overview and Language browsers, with the Feature browser remaining to be added. The access and search functions of the application and web versions are the same. The web version is not dependent on Java or JavaScript for its input or display, and as such should be accessible with any browser supporting tables (e.g. Netscape Navigator 2 on up).

More importantly than its current structure, however, work continues to make both the Language and Feature browsers of both the application and web versions accept new input. In this way, any researcher who wishes to contribute to the database can add

languages and accompanying data, and send the entered records back to the server for inclusion in the main database. Future versions will also include the ability to gloss entries in multiple languages.

4.2 Database contents

Before turning to the specific contents of the database, it is useful to detail the considerations that led to the choosing of that content.

4.2.1 Feature considerations

As noted in §3.1, the features included in the cogPrime database were initially based solely on physical properties of things that can be sensed in our environment. The rationale was that cognitive constructs reflecting concrete objects have been set before language use has begun in earnest. Semantic features reflecting concrete objects' properties, therefore, should be universal in nature—within reasonable bounds of variability, an infant growing up in one country will have the same basic pre-lingual sensory experiences as one growing up in any other country. In contrast, abstract concepts are based on our experiences after language use has commenced, and therefore lexical items representing abstract thoughts, emotions, etc. are inherently dependent on our linguistic background.

However, reflection would make it seem that identifying properties of objects is meaningless without also placing those objects within the cognitive spatial plane; as noted, this is supported by the research discussed in Landau (1994). The set of features therefore also includes those that reflect spatial location. Reflection would also seem to support the inclusion of features reflecting basic judgments of time; these are included as well. While the entire notion of linear time is arguably non-universal, 'basic' judgments—old, young, early, late, etc.—would seem to be universally human.

Likewise, basic color terms were added. While much can and has been said about color recognition and terminology, these features are presently being used as unary, simplistic features without regard to hue, lightness or saturation.²² Color studies suggest that inclusion of features to reflect these concepts may be necessary in the future; i.e.

²² Basic in the sense of Berlin & Kay (1969); e.g. *red* is being used to denote all reference to the "reds"—see Dedrick (1998) pp. 16-17, ch. 2 for discussion.

brightness values have been indicated as being more perceptually salient for colors in the middle of the spectrum than those at the ends (Yoshioka et al. 1996).

Another initial intent was to categorize the set of features based on sensory modality; i.e. all features reflecting properties perceived through visual input. However, upon investigation it quickly became apparent that much of cognitive development is multi-modal; i.e., is dependent on input from more than one sense. This is also true for language acquisition (see the discussion of the McGurk effect in §2.3 above).

In light of this, the feature categories used in this project are not based strictly on sensory modality. Instead they are roughly sensory-based, but are grouped into categories according to judged similarity of the physical properties they reflect. For example, the features associated with the tactile properties of an object can be sensed by two or more modes (a desktop can either look or feel smooth, rough, etc.; the sound generated by the sliding of something over its surface can also serve as a clue).

One last consideration is that of how to best represent the features themselves in the database.²³ One decision that had to be made was what type of features to use—unary, binary (privative or not) or ternary (e.g. Goldsmith 1985; Ringen 1988). In consideration of the discussion in Lyons (1977: 322ff), the initial feature set consists of non-privative binary features, with the sole exception of the unary feature [color].

Secondly, as noted in Lyons (1968: 477ff), the use of common English words as labels is singularly unsatisfactory. This predisposes the reader (or in this case, the user of the database) to think in terms the English meaning of the feature label, coloring their perception of the feature being used to the extent of their English language ability. The best universal alternative, it would seem, is to use small icons for the features. However, this would seem to be both difficult to implement and to use—one could assign a unique icon to each feature entered into the database, but icons small enough to fit into a viewable area will be hard to discern, while easily discernible icons of features assigned to more complex words will require a great deal of screen space. While options are being explored, therefore, the cogPrime database will continue to use English words for features, substituting technical terms for basic words as work progresses, and simply warn the users (particularly native English speakers) not to assign a feature more meaning than it is meant to have.

4.2.2 Adjectives

The initial set of adjectives selected for inclusion in the database was based on the concrete features discussed above. Once the initial set was gathered, the English adjectives reflective of those features (i.e. *long* for the principal dimension of a 3D object) were then listed. Mono-morphemic adjectives were selected since in general derived adjectives in English assign the attributes of the noun or verb they were derived from (e.g. N-y indicates the presence or characteristics of N) or antonymy (e.g. *unmarried* stands in antonymous relation to *married*). This led to a set of app. 50 adjectives. This set was then used as the gloss for the adjectives of other languages' datasets using standard, software- and web-based dictionaries. The set of adjectives for each language, therefore, were initially based on the same features. However, since one-to-one correspondences for adjectives between languages is certainly not the rule, the current number of adjectives for each non-English language varies.

In addition to these initial adjectives, adjectives reflecting the 11 basic color terms of Berlin & Kay's (1969) 'typical stage VII system' were included. Other terms reflective of an object's position in and movement through spatial organization, as well as simple temporal terms, were included to round out the initial set.

4.2.3 Languages

The database currently contains adjective entries for at least some of this core set of adjectives for the following languages: English, Japanese, French, German, Spanish, Uyghur (an Altaic language spoken in mainland Asia and Russia), Bulgarian, and Irish. Other languages whose addition is planned are Italian, Korean, Greek, Hindi, Vietnamese, Russian, and Norwegian. It is hope that the database can be expanded to other language regions as well in the next year to achieve a more balanced representation of language families.

There are inherent problems to be resolved in any sampling of languages and language families like this, of course. For example, Uyghur, like many other Altaic languages, does not utilize an adjectival class in the way that we speak of one in Indo-European languages; Uyghur 'adjectives' function equally well as adverbials with all

²³ Thanks again to Yo M. for pointing me to this reference.

the resulting ambiguity (Hahn 1991: 333). However, it is expected that all languages have some way of attributing a limited feature or set of features to an object; these will be included in the database as well, using modifications in the browsers as required.

4.2.4 Features & feature categories

Although it must be noted that much refinement remains to bring them up to the necessary level (e.g. the in-depth analysis of Japanese numeral classifiers of Matsumoto 1983), the initial set of features and feature categories are detailed below.

The category **status** was created to track the metalinguistic characteristics of each entry, and to allow for searching of the database based on these features. The following features were identified and encoded, all of them state features (see §3.4.1 above):

derived: signifies derivation through some morphological process

unmarked: signifies an adjective that is used as an unmarked term
(e.g. the unmarked English question is "How long is it?")

requiresObject: signifies that an adjective can only modify a concrete object

The category **intrinsicity** contains features that pertain to the chemical composition or properties of an object or substance. The features assigned to this category relate to internal properties not observed by tactile contact; those have been separated into their own category **tactility** discussed below. The features currently assigned to the category **intrinsicity** are all considered monopolar scalar items; the weight, viscosity and visciduity scales are considered to be open while the plasticity scale²⁴ is bounded—a bendable object can have a degree of plasticity varying from 1 (i.e. fully retains its ability to be deformed) to 0 (i.e. stays fully deformed by whatever force). It is not expected that natural languages as opposed to scientific jargon make use of terms for plasticity, but it has been included for completeness:

weight: the weight of an object in the ordinary use of the word

plasticity: a measure of an objects resistance to continuous deformation

elasticity: a measure of an object's ability to return to its original size and shape after being deformed

²⁴ In the sense of its definition in physics: "Capable of undergoing continuous deformation without rupture or relaxation."—from the CD version of the American Heritage Dict. v.4.0.

viscosity: a measure of the resistance a liquid shows to flow

viscidity: a measure of the stickiness of a liquid

The category **size** has been reserved here for the physical dimensions of an object. From the view of humans interacting with their environment, the principal dimensions length, width, depth and height have been supplemented with area and volume to reflect human use of objects with these characteristics. All features used here are treated as being associated with open monopolar scales, their range being limited only by physical constraints of a given reference system:

length: a measure of the principal (longest) dimension of an object

width: a measure of the second longest dimension of an object

thickness: a measure of the shortest dimension of an object

height: a measure of the distance from the ground to the top of an object

(cf. superiority below in the category **spatiality**)

depth: a measure of the inferior extension of a space or liquid

area: a measure of a 2-dimensional flat space

volume: a measure of a 3-dimensional space

Dimensionality is used in the database to mark when an adjective requires an object existing in the specified dimension. For example, the adjective *length* requires only that the referred to object exist in at least one dimension, as is the case with a mathematical line. In contrast the use of an adjective to describe the height of an object requires that the object exist in 3 dimensions; 2-dimensional objects have no height. The dimensional features included in the database as the present time are **1D**, **2D**, and **3D**, although **4D** may be introduced to cover the case of movement through the 3-dimensional space surrounding an observer if it later seems warranted.

Tactility, as is probably obvious, is meant to capture the various sensations experienced through our sense of touch (cf. the category **intrinsicity** above):²⁵

heat: a subjective measure of the heat flow in from the surface being contacted

coldness: a subjective measure of the heat flow out to the surface
being contacted

²⁵ See Craig & Rollman (1999: 308-309) for references to research into the 3-dimensional

pliability: the relative compressibility of the surface being contacted

solidity: the uniformity of an object's internal structure

smoothness: the relative abrasiveness of the surface being contacted

pointedness: how sharp a pointed object is; cf. edgedness below

edgedness: how sharp an edged object is; cf. pointedness below

wetness: to what degree the surface of an object is covered by water

The category **content** was included to cover our experiences with vessels and containers. These items could well be argued to be social constructs as opposed to cognitive ones, since humans have not always used containers. However, it has been demonstrated that infants as young as 2 mos. old seem to have at least a rudimentary knowledge of whether something can be a container or not (Hespos & Baillargeon 2001). Also, even lower mammals seem to be able to tell at a glance whether or not a container holds something of interest; they seem to have some concept of 'empty' or 'has content':

isContainer: signifies whether or not the adjective can modify a container

fullness: the measure of whether the container has been filled or not

Visibility, as the category name suggests, used to categorize how we visually interpret our environment. Initially 3 features have been included: **emissivity**, **reflectivity** and **opacity**. The fourth feature presently in the database, **color**, is a catch-all feature to denote all adjectives denoting the color of a thing. This more common but less precise feature was chosen over the technical notion of **hue**, often defined in terms of the dominant reflected wavelength since the two do not always correspond (see Dedrick 1998:25ff for details). As noted earlier, each simplistic, base color term (as denoted by the list of basic color terms in Berlin & Kay 1969) are thought not to be compositional, but are instead complete characterizations of that color attribute. The first three category features are associated with open monopolar scales, while color is simply being assumed to be a state feature:

emissivity: a measure of how much light an object is producing

reflectivity: a measure of how much light an object reflects

relationship between roughness/smoothness, hardness/softness, and compressional elasticity.

opacity: a measure of how well light passes through an object

color: the color denoted by the adjective

Taste is also fundamental to the experience of assumedly all living creatures. While the first four tastes—salty, sweet, bitter and sour—are well-known through the world, a fifth type of taste bud was discovered fairly recently. That taste bud reacts to gluten, a food constituent found in foods high in protein, and is known as *umami* (Kawamura & Kare 1987). This feature has also been included in the database for completeness, although it is expected that the sensation of this taste is subtle enough that it has not been explicitly encoded in any language. As with **color**, these are all assumed to be state features (i.e. something is either *salty* or *not salty*): **saltiness; sweetness; bitterness; sourness; glutenosity.**

In the category **sound**, loudness and pitch are straightforward-enough inclusions. It also seems worthwhile to include the feature harmony to represent the harmoniousness or dissonance of a sound. While this is certainly a subjective matter, just as one's personal taste in music, there would seem to be a universal mammalian attraction to harmonious sounds and a universal repulsion of discordant ones. Further investigation may be able to sort out whether this feature should remain or not:

loudness: the perceived strength of an acoustic signal

pitch: the perceived pitch of an acoustic signal

harmony: how pleasing to the ear a sound is

The features in this category **spatiality** have to do with the spatial location of an object as opposed to a feature reflecting the object's physical characteristics. **Distance** is fairly transparent; less obvious inclusions are **frontedness** for whether something is in front of us or behind us; **rightedness**, whether something is oriented in a proper direction or is upside-down; and **superiority**, whether something is above the level of the observer or below it:

distance: a measure of how far something is from the observer

frontedness: a measure of how directly in front of the observer something is

superiority: a measure of how far above or below the observer's plane
of reference an object is

rightedness: a measure of how far off an object vertical is from its
perceived axis

speed: a measure of the speed of an object relative to the observer

The category **temporality** was included to cover aspects of time as it applies to the aging of things, durations, and the regular cyclic nature of our environment. As studies have shown, even the lowly fruit fly (*Drosophila melanogaster*) is synched with the circadian rhythms of the Earth (Young 2000). In addition to external cycles, there is the internal cycles of our bodies and our lives that are likely to have been universally encoded in cognition and hence influence language use:

age: a measure of the time since something was born or made

duration: a measure of an amount of elapsed time

lateness: a measure of how far into the normal cycle of a duration-bounded event the object or observer is

Our sense of **smell**, above all other senses, is the most primordial (Whitfield & Stoddart 1984). And researchers are gradually gaining a better understanding of how important our perceptions of pheromones are in our daily social interaction and the cyclic nature of our lives (Seppa 1998; Weller 1998). For this reason it seems justified to include this category. Unlike **taste**, however, which associates a given chemical property with a given set of taste buds, smell is discerned wholly in the brain; there are no 'smell buds' as such (White & Treisman 1997). Because of this complexity, and the fact that the average adult human can discern something on the order of 10,000 different aromas (Axel 1995), the features that should be included in this category are left for the time being.

Again, as noted before, the features detailed above and the categories they are grouped into will certainly change as work progresses, and as other features are added.

5. Summary & further directions

Due to the general compositional nature of language acquisition and analysis of compositional idioms, and the success of previous compositional analyses, it is taken in this paper that a compositional approach to word meaning is a justified and desirable as a part of semantic description. The cogPrime database of semantic features, with initial focus on those features that reflect the concrete properties of concrete objects, seeks to

provide a means of delineating semantic features and comparing the similarities and differences in their use in various languages. By making the database available for download and accessible via the web, it is hoped that it can be expanded beyond its present scope to include other feature and word categories.

The design and coding of the database application and its accompanying web version have taken the better part of the past eight months. Now that the application's form and function is fairly well set, further investigation into the cognitive and semantic literature, refinement of the appropriate features and categories for inclusion, and collection and refinement of the data content can be undertaken. This will include data elicitation in interviews with native speakers of various languages using iconic representations of objects to elicit words used for specific attributes. A web-based questionnaire is planned as well, one that uses graphics instead of word lists, to perform the same elicitation tasks.

There remain, of course, technical difficulties to resolve. Although the database application is being developed on a Macintosh G4 running OS 9, affording built-in multi-lingual font input and output capability, users cannot be expected to using the same platform particularly when accessing the web version of the database. Font coding, then, will have to move to Unicode standards at some point in the future. In addition, although the development environment being used, 4th Dimension™, is cross-platform in nature, challenges of supporting a cross-platform product remain.

In closing, it will be stated one last time that the intention is for this database to evolve into a wider collaborative effort that will not only facilitate cross-linguistic testing of the compositional approach to word meaning, but help shed light on semantic analysis of all kinds.

References

- Aitchison, Jean. 1987. *Words in the mind: An introduction to the mental lexicon*. Oxford: Blackwell Publishers.
- Armstrong, S. L., L. R. Gleitman, and H. Gleitman. 1983. What some concepts might not be. *Cognition* 13: 263-308.
- Armstrong, Susan, ed. 1994. *Using Large Corpora*. Cambridge, Mass.: MIT Press.
- Armstrong, Susan, Kenneth Ward Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann and David Yarowsky, ed. 1999. *Natural language processing using very large corpora*. Norwell, Mass.: Kluwer Academic Publishers.
- Axel, Richard. 1995. The molecular logic of smell. *Scientific American* 273: 154-159.

- Baird, John C., and Kathleen A. Harder. 2000. The psychophysics of imagery. *Perception and Psychophysics* 62: 113-126.
- Beazley, L.D., S.A. Dunlop, D.K. Chelvanyagam and W.M. Ross. 1995. Wiring up the visual system. *Clinical and Experimental Pharmacology and Physiology* 22: 550-558.
- Berlin, Brent, and Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution*. Berkeley: University of California Press.
- Bierwisch, Manfred. 1967. Some semantic universals of German adjectivals. *Foundations of Language* 3: 1-36.
- Bierwisch, Manfred. 1989 The semantics of gradation. In ed. M. Bierwisch and E. Lang, *Dimensional adjectives: grammatical structure and conceptual interpretation*, 71-261. Berlin, New York: Springer-Verlag.
- Blevins, Juliette. 1995. The syllable in phonological theory. In *The handbook of phonological theory*, ed. John A. Goldsmith, 206-244. Cambridge: Blackwell.
- Bloom, Lois. 1993. *The transition from infancy to language: Acquiring the power of expression*. New York: Cambridge University Press.
- Bloom, Paul. 2000. *How children learn the meanings of words*. Cambridge.: MIT Press.
- Bloom, Paul, and Tim P. German. 2000. Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77: B25-B31.
- Bolinger, Dwight. 1965. The Atomization of Meaning. *Language* 41: 555-573.
- Brysbaert, Marc, Wim Fias & Marie-Pascale Noël. 1998. The Whorfian hypothesis and numerical cognition: is 'twenty-four' processed in the same way as 'four-and-twenty'? *Cognition* 66: 51-77.
- Bushara, Khalafalla O., Jordan Grafman and Mark Hallett. 2001. Neural Correlates of Auditory-Visual Stimulus Onset Asynchrony Detection. *Journal of Neuroscience* 21: 300-304.
- Byrne, Richard W. 2000. Evolution of primate cognition. *Cognitive Science* 24: 543-570.
- Chomsky, Noam. 1981. Principles and parameters in syntactic theory. In ed. N. Hornstein and D. Lightfoot, *Explanations in linguistics*. London: Longman.
- Church, Kenneth W., and Robert L. Mercer. 1994. Introduction to the special issue on Computational Linguistics using large corpora. In *Using large corpora*, ed. Susan Armstrong, 1-24. Cambridge, Mass.: MIT Press.
- Clark, Eve V. 1998. Morphology in language acquisition. In *The handbook of morphology*, ed. Andrew Spencer and Arnold M. Zwicky, 374-389. Oxford: Blackwell.
- Clements, George, and Samuel Keyser. 1983. *CV phonology: A generative theory of the syllable*. Cambridge, Mass.: MIT Press.
- Clements, George. 1985. The Geometry of phonological features. *Phonology Yearbook* 2: 225-252.
- Clements, George, and Elizabeth Hume. 1993. The internal organization of speech sounds. In *A Handbook in phonological theory*, ed. John Goldsmith, 245-306. Cambridge: Blackwell.
- Coren, Stanley, Clare Porac & Lawrence M. Ward. 1978. *Sensation and Perception*. New York: Academic Press.
- Craig, A.D., E.M. Reiman, A. Evans and M.C. Bushnell. 1996. Functional imaging of an illusion of pain. *Nature* 384: 258-260.
- Craig, James C., and Gary B. Rollman. 1999. Somesthesia. *Annual Review of Psychology* 50: 305-331.
- Dedrick, Don. 1997. Colour categorization and the space between perception and language. *Behavioral and Brain Sciences* 20:187-188.
- Dedrick, Don. 1998. *Naming the rainbow: Colour language, colour science, and culture*. Dordrecht: Kluwer.
- Dehaene, S., E. Spelke, P. Pinel, R. Stanescu and S. Tsivkin. 1999. Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science* 284: 970-974.

- Deutsch, Dianne. 1987. The tritone paradox: effects of spectral variables. *Perception and Psychophysics* 41: 563-575.
- Deutsch, Dianne. 1992. Paradoxes of musical pitch. *Scientific American* 267: 88-95.
- Deutsch, Dianne. 2000. Bilingual speakers perceive a musical illusion in accordance with their first language. *Journal of the Acoustic Society of America* 2591: 2591.
- Fellbaum, Christiane, ed. 1998. *WordNet: An electronic lexical database*. Cambridge, Mass.: MIT Press.
- Fillmore, Charles. 1982. Frame semantics. In ed. Linguistic Society of Korea, *Linguistics in the morning calm*, 111-138. Seoul: Hanshin.
- Fisher, Cynthia, D. Geoffrey Hall, Susan Rakowitz and Lila Gleitman. 1994. When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. In *The acquisition of the lexicon*, ed. Lila Gleitman and Barbara Landau, 333-375. Cambridge, Mass.: MIT Press.
- Francis, W. Nelson, and Henry Kuchera. 1982. *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.
- Fries, Charles C., and A. Aileen Traver. 1940. *English word lists*. Washington, D.C.: American Council on Education.
- Fodor, Jerry A. 1975. *The language of thought*. New York: Crowell.
- Gleitman, Lila, and Barbara Landau, Eds. 1994. *The acquisition of the lexicon*. Cambridge, Mass.: MIT Press.
- Goldsmith, John. 1990. Vowel harmony in Khalkha Mongolian, Yaka, Finnish, and Hungarian. *Phonology Yearbook* 2: 251-275.
- Gracco, Carol, Clarence T. Sasaki, Richard McGowan, Elizabeth Tierney and John Gore. 1994. Magnetic resonance imaging (MRI) in vocal tract research: Clinical application. *Journal of the Acoustical Society of America* 95: 2821.
- Graziano, Michael S. A., Dylan F. Cooke and Charlotte S. R. Taylor. 2000. Coding the location of the arm by sight. *Science* 290: 1782-1786.
- Hahn, Reinhard F. 1991. *Spoken Uyghur*. Seattle: University of Washington Press.
- Hauser, Marc D., Elissa L. Newport, et al. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition* 78: B53-B64.
- Heider, E. R. 1972. Universals in color naming and memory. *Journal of Experimental Psychology* 93: 10-20.
- Herman, Louis, Stan Kuczaj II and Mark Holder. 1993. Responses to anomalous gestural sequences by a language-trained dolphin: Evidence for processing of semantic relations and syntactic information. *Journal of Experimental Psychology: General* 122: 184-194.
- Herman, Louis M., and Robert K. Ueyeyama. 1999. The dolphin's grammatical competency: Comments on Kako (1999). *Animal Learning and Behavior* 27: 18-23.
- Houde, John F., and Michael I. Jordan. 1998. Sensorimotor Adaptation in Speech Production. *Science* 279: 1213-1216.
- Ikegami, Yoshihiko. 1970. *The semological structure of the English verbs of motion: A stratificational approach*. Tokyo: Sanseido.
- Jackendoff, Ray. 1983. *Semantics and cognition*. Cambridge, Mass.: The MIT Press.
- Jackendoff, Ray. 1996. Semantics and Cognition. In *The handbook of contemporary semantic theory*, ed. Shalom Lappin, 539-559. Oxford: Blackwell.
- Janssen, Peter, Rufin Vogels and Guy A. Orban. 2000. Selectivity for 3D shape that reveals distinct areas within Macaque inferior temporal cortex. *Science* 288: 2054-2056.
- Kager, Rene. 1999. *Optimality theory*. Cambridge: Cambridge University Press.
- Kandel, Eric R. and Larry R. Squire. 2000. Neuroscience: Breaking down scientific barriers to the study of brain and mind. *Science* 290: 1113-1120.
- Katz, Jerrold J., and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language* 39: 170-210.
- Katz, Jerrold J. 1972. *Semantic theory*. New York: Harper and Row.

- Kawahara, Hideki. 1993. Transformed auditory feedback: Effects of fundamental frequency perturbation. *Journal of the Acoustical Society of America* 94: 1883-1884.
- Kawamura, Y., and M.R. Kare, Eds. 1987. *Umami: A basic taste*. Marcel Dekker: New York.
- Kellman, P.J., and E. Spelke. 1983. Perception of partly occluded objects in infancy. *Cognitive Psychology* 15: 483-524.
- Kuhl, Patricia K., and James D. Miller. 1978. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *The Journal of the Acoustic Society of America* 63: 905-917.
- Kuhl, P. K., K. A. Williams, F. Lacerda, K. N. Stevens and B. Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255: 606-608.
- Kuhl, Patricia, and Andrew N. Meltzoff. 1995. Vocal learning in infants: Development of perceptual-motor links for speech. In *Proceeds of the XIIIth International Congress of Phonetic Sciences*, Stockholm: Swedish Royal Institute of Technology and Stockholm University. 146-149.
- Labov, William. 1973. The boundaries of words and their meanings. In *New ways of analyzing variation in English*, ed. Bailey and Shuy: 340-373.
- Landau, Barbara. 1994. Where's what and what's where: The language of objects in space. In ed. Lila Gleitman and Barbara Landau, *The Acquisition of the lexicon*, 259-296. Cambridge: MIT Press.
- Linden, D. E., U. Kallenbach, A. Heinecke, W. Singer and R. Goebel. 1999. The myth of upright vision. A psychophysical and functional imaging study of adaptation to inverting spectacles. *Perception* 28: 469-81.
- Lipka, Leonhard. 1997. The meaning of meaning: Approaches to semantics. In *The locus of meaning*, ed. K. Yamanaka and T. Ohori, 47-61. Tokyo: Kuroshio Publishers.
- Lisman, John E., and Justin R. Fallon. 1999. What Maintains Memories? *Science* 283: 339-340.
- Locke, John L. 1993. *The child's path to spoken language*. Cambridge, Mass.: Harvard University Press.
- Lyons, John. 1968. *Introduction to Theoretical Linguistics*. Cambridge, Mass.: Cambridge University Press.
- Lyons, John. 1977. *Semantics*, 2 vols. Cambridge: Cambridge University Press.
- Matsumoto, Yo. 1993. Japanese numeral classifiers: a study of semantic categories and lexical organization. *Linguistics* 31: 667-713.
- McKeon, Richard, ed. 1941. *Basic works of Aristotle*. New York: Random House.
- McGurk, H., and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264: 746-748.
- Meltzoff, A.N., and M.K. Moore. 1977. Imitation of facial and manual gestures by human neonates. *Science* 198: 75-78.
- Miller, George A., and Philip N. Johnson-Laird. 1976. *Language and perception*. Cambridge, Mass.: Harvard University Press.
- Newmeyer, Frederick J. 1986. *The politics of linguistics*. Chicago: CUP.
- Newmeyer, Frederick J. 1998. *Language form and language function*. Cambridge, Mass.: MIT Press.
- Niida, Eugene A. 1975. *Componential analysis of meaning*. The Hague: Mouton.
- Nöth, Winfried. 1997. The Semantic space of opposites: Cognitive and localist foundations. In *The locus of meaning*, ed. K. Yamanaka and T. Ohori, 64-82. Tokyo: Kuroshio Publishers.
- Nunberg, Geoffrey, Ivan Sag and Thomas Wasow. 1994. Idioms. *Language* 70: 491-538.
- Peterson, Candida, and Michael Siegal. 1995. Deafness, conversation and theory of mind. *Journal of Child Psychology and Psychiatry and Allied Disciplines* 36: 459-474.
- Pinker, Steven. 1987. The bootstrapping problem in language acquisition. In *Mechanisms of language acquisition*, ed. B. MacWhinney. Hillsdale, NJ: Erlbaum.
- Pinker, Steven. 1994. How could a child use syntax to learn verb semantics? In *The acquisition of the lexicon*, ed. Lila Gleitman and Barbara Landau, 377-410. Cambridge, Mass.: MIT Press.

- Pitt, David, and Jerrold J. Katz. 2000. Compositional idioms. *Language* 76: 409-432.
- Prince, Alan, and Paul Smolensky. 1993. ms., *Optimality Theory: Constraint interaction in generative grammar*. Rutgers University.
- Pustejovsky, James, Peter Anick and Sabine Bergler. 1994. Lexical semantic techniques for corpus analysis. In *Using large corpora*, ed. Susan Armstrong, 291-318. Cambridge, Mass.: MIT Press.
- Pustejovsky, James. 1998. *The generative lexicon*. Cambridge, Mass.: The MIT Press.
- Ramus, Franck, Marc D. Hauser, Cory Miller, Dylan Morris and Jacques Mehler. 2000. Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science* 288: 349-351.
- Rapoport, Anatol. 1975. *Semantics*. Reprint of the 1911 *Invitation to semantics*. New York: Crowell.
- Ringen, Catherine. 1988. Underspecification theory and binary features. In eds. Harry van der Hulst and Norval Smith, *Features, Segmental Structure and Harmony Processes*, 145-160. Dordrecht: Foris.
- Rosch, E. 1975a. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104: 192-233.
- Rosch, Eleanor. 1975b. Universals and cultural specifics in human categorization. In ed. S. Bochner, W.J. Lonner and R.W. Brislin *Cross-cultural Perspectives on Learning*. : Halstead Press.
- Rosenblum, L.D., M.A. Schmuckler and J.A. Johnson. 1997. The McGurk effect in infants. *Perception and Psychophysics* 59: 347-357.
- Ross, J. R. 1967. *Constraints on variables in syntax*. Ph. D. thesis, MIT.
- Russell, PA, JA Hosie, CD Gray, C Scott, N Hunter, JS Banks and MC Macaulay. 1998. The development of theory of mind in deaf children. *Journal of Child Psychology and Psychiatry* 39: 903-10.
- Sagey, Elizabeth. 1986. *The representation of features and relations in nonlinear phonology*. Ph.D. dissertation, MIT.
- Sapir, Edward. 1949. *Selected writings of Edward Sapir in language, culture and personality*. David Mandelbaum ed. Berkeley: University of California Press.
- Schwimmer, Brian. 1996. Anthropology on the Internet: A Review and Evaluation of Networked Resources. *Current Anthropology* 37: 561.
- Schyns, P. G., R. L. Goldstone and J. P. Thibaut. 1995. *The development of features in object concepts*. Indiana University Cognitive Science Report 133. Bloomington, IN.
- Schyns, Philippe G., and Luc Rodet. in press. Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition*.
- Seidenberg, Mark S. 1997. Language acquisition and use: Learning and applying probabilistic constraints. *Science* 275: 1599-1603.
- Selkirk, Elisabeth. 1995. Sentence prosody: Intonation, stress, and phrasing. In *The handbook of phonological theory*, ed. John A. Goldsmith, 550-569. Cambridge: Blackwell.
- Seppa, N. 1998. Nailing down pheromones in humans. *Science News* 153:164.
- Sherrick, Carl E., and Roger W. Cholewiak. 1986. Cutaneous sensitivity. In ed. Kenneth R. Boff, Lloyd Kaufman and James P. Thomas *Handbook of Perception and Human Performance: Volume 1—Sensory Processes and Perception*, 12-1 - 12-58. New York: John Wiley and Sons.
- Spelke, Elizabeth. 1994. Initial knowledge: Six suggestions. *Cognition* 50: 443-447.
- Suzuki, Hidekazu. 1998. On the use of linguistic database for linguistic research. In *Report of the Special Research Project for the Typological Investigation of Languages and Cultures of the East and West 1998* 2: 183-197.
- Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in optimality theory*. Cambridge, Mass.: MIT Press.
- van den Brink, G. 1982. On the relativity of pitch. *Perception* 11, 721-731.

- Vihman, Marilyn May. 1996. *Phonological development: The origins of language in the child*. Cambridge: Blackwell.
- Voss, David. 2000. Scientists Weave New-Style Webs to Tame the Information Glut. *Science* 289: 2250-2251.
- Wade, Nicholas. 1982. *The art and science of visual illusions*. London: Routledge and Kegan Paul.
- Weinreich, Uriel. 1972. *Explorations in semantic theory*. The Hague: Mouton.
- Welch, Robert B. 1978. *Perceptual modification: Adapting to altered sensory environments*. New York, Academic Press.
- Welch, Robert B., Bruce Bridgeman, Sulekha Anand and Kaitlin E. Browman. 1993. Alternating prism exposure causes dual adaptation and generalization to a novel displacement. *Perception and Psychophysics* 54: 195-204.
- Weller, Aron. 1998. Communication through body odor. *Nature* 392: 126.
- White, T., and Treisman M. 1997. A comparison of the encoding of content and order in olfactory memory and in memory for visually presented verbal materials. *British Journal of Psychology* 88: 459-469.
- Whitfield, P., and D.M. Stoddard. 1984. *Hearing, taste, and smell: Pathways of perception*. New York: Torstar Books.
- Whorf, Benjamin Lee. 1956. *Language, thought and reality: Selected writings of Benjamin Lee Whorf*. John B. Carroll, Ed. New York: Wiley.
- Wienhold, Götz, and Ulrich Rohmer. 1997. On implications in lexicalizations for dimensional expressions. In *The locus of meaning*, ed. K. Yamanaka and T. Ohori, 64-82. Tokyo: Kuroshio Publishers.
- Wierzbicka, Anna. 1985. *Lexicography and conceptual analysis*. Ann Arbor, Mich.: Coroma.
- Wierzbicka, Anna. 1996. *Semantics: Primes and universals*. Oxford: Oxford Univ. Press.
- Yoshioka, T., B.M. Dow and R.G. Vautin. 1996. Neuronal mechanisms of color categorization in areas V1, V2 and V4 of macaque monkey visual cortex. *Behavioral Brain Research* 76: 51-70.
- Young, Michael W. 2000. Marking time for a kingdom. *Science* 288: 451-453.
- Zubin, David A., and Soonja Choi. 1984. Orientation and gestalt: Conceptual organizing principles in the lexicalization of space. In *Papers from the Parasession on Lexical Semantics*. ed. D. Testen, V. Mishra and J. Drogo, 333-345. Chicago, Chicago Linguistic Society.

J. Kevin Varden
Inst. of Modern Lang. & Culture
University of Tsukuba
1-1-1 Tennodai; Tsukuba
Ibaraki; 305-8571 JAPAN

kevin@varden.com