Representing Idioms: Syntactic and Contextual Effects on Idiom Processing

Ed Holsinger
University of Sourthern California
eschlon@mac.com

1. Introduction

Since *Aspects* (Chomsky, 1965), the prevailing view of the language system has been one in which linguistic behavior emerges as the interaction between two qualitatively distinct components. On the one hand is the lexicon, consisting of a set of learned associations between symbols and concepts. On the other is the grammar, characterized as a set of potentially innate rules that operate over these symbols. The critical distinction between these components is computational. Words are *stored* and *accessed*, sentences are *composed* and *computed*.

Recently, approaches to language have begun to move away from this bifurcated approach. Newer models argue that linguistic knowledge emerges as a consequence of the dynamics of the human cognitive system and the individual's experience with language (Goldberg, 1995, 2006, Tomasello, 2003; Bod 1998; Pierrehumbert, 2001). Under this paradigm the crucial question becomes one of representation. The question of representation can be examined both theoretically (e.g. What sorts of framework maximizes available resources?) and empirically (e.g. Is there evidence that linguistic knowledge is represented one way or another?).

In this paper we will be focusing on an empirical investigation of how the language system represents multi-word units. Multi-word units provide a valuable window into language because they tap into both processing and storage resources. Additionally multi-word exemplars form the base of linguistic experience in usage-based approaches, from which combinatoric knowledge is derived. In this paper we use idiomatic expressions, such as *kick the bucket*, to examine how linguistic knowledge is represented and accessed during on-line comprehension. Idioms provide a valuable test-case in this endeavor, as they exhibit both word-like and structure-like properties.

1.1 Idioms

Much of the early research into idiomatic expressions treated them as words-with-spaces (Bobrow & Bell, 1973; Swinney & Cutler, 1979; Weinreich, 1969). Like words, idioms appear to be arbitrary learned mappings between linguistic form and meaning. Early empirical investigations demonstrated that idioms could be accessed more rapidly than literal expressions and this has been replicated extensively in the literature (Swinney & Cutler, 1979; Gibbs, 1980; McGlone et al., 1994, Ortony et al., 1978). The logic is that idioms are stored and accessed as whole units, and that this sort of direct access is computationally less expensive than the process of accessing and integrating the meanings of individual words into a structure.

More recent work has shed doubt upon the words-with-spaces view. Investigations into idiom meaning has challenged the assumptions that all idioms are compositionally opaque and that the words comprising an idiomatic expression bear no relation to its meaning (Gibbs et al, 1989, 1997). Structural investigations have demonstrated that individuals are aware of the grammatical properties of idiom-internal words (Peterson et al., 2001), and that idioms operate like structures for the purposes of engaging in phenomenon such as structural priming (Konopka & Bock, 2009). Additionally, recent work suggests that the rapid access profile of idioms, the primary evidence in favor of the words-with-spaces approach, also applies to clichés (Tabossi et al, 2009) and more generally to frequent multi-word expressions (Arnon & Snider, 2010).

Taken as a whole, these results support a hybrid representation for idiomatic expressions which preserves their structural properties while still accounting for their more word-like meaning (Cacciari & Tabossi, 1988; Cutting & Bock, 1997; Sprenger et al., 2006). Current models in this paradigm differ in their representations and underlying mechanics, however they all

converge upon the notion that literal processing plays a causal role in the access of idiomatic meaning. Cacciari & Tabossi (1988), drawing evidence from idiom comprehension, proposed the Configuration Hypothesis. Under this model the parser proceeds with literal interpretation until it has incrementally accumulated sufficient evidence that the string in question is idiomatic. At this point the idiomatic meaning is retrieved. Crucially how far down the literal parse the processor goes is dependent upon the degree to which the idiomatic string can be plausibly taken literally, and how well the parser can predict the intended meaning during online comprehension. They refer to this tipping point as the idiom key (see Tabossi & Zardon, 1993).

In a similar vein, Cutting & Bock (1997) proposed a distributed representation for idioms based on speech error data. In a series of experiments, they found that incidence of speech errors increased as a factor of overlap in meaning or structure regardless of whether the structure in question was idiomatic. They proposed that idioms are represented as structural phrasal frames directly associated with conceptual meaning. Sprenger et al. (2006) found that idiom production was facilitated by priming phonological and semantic associates of its literal component words. They proposed that rather than phrasal frames, idioms are represented as *super-lemmas* which operate as a grammatical function over the component lemmas of the idiom. These super-lemmas provide a representation for idioms, and potentially other multi-word collocations, which can account for the idiosyncratic nature of idiom meaning and their varying degrees of structural flexibility while allowing for normal lexical competition between idioms and normal lemmas (e.g. *kick the bucket* vs. *die*) with minimal storage redundancy (see Kuiper et al., 2007).

1.2 Motivation & Predictions

The super-lemma hypothesis provides a detailed model of idiom representation. However, the precise representational content of these super-lemmas is the subject of debate (Tabossi et al., 2009). Additionally, applying this model to idiom comprehension is not trivial. In production, activation spreads from the conceptual layer down to the individual component lemmas. Thus for an idiom such as *kick the bucket*, the super-lemma representation will enter normal competition with other semantically associated lemmas and super-lemmas (e.g. *die, pass away*). Once selected, activation will then spread to the individual component lemmas (e.g. *kick* and *bucket*) with the super-lemma providing pre-computed phrasal configurations for the string.

During comprehension, activation spreads upward from the component lemmas to the conceptual level. The role of the super-lemma in this process is less clear. One possibility is that the super-lemma representation acts as a sort of gate to the conceptual layer. Thus during on-line processing, encountering a syntactic environment incompatible with an idiomatic reading would preclude activation spreading to the conceptual layer. Under this view, strings such as *the bucket was kicked* should not result in consideration of idiomatic meaning as the super-lemma representation contains no appropriate function. Another possibility is that super-lemmas participate in spreading activation during comprehension in the same way as other lemmas. Thus any partial activation of the super-lemma representation will necessarily spread to the conceptual layer. If this is the case we would predict that incompatible syntactic context would not be sufficient to prevent some consideration of the idiomatic interpretation during comprehension.

Unlike syntactic incongruence, however, contextual bias is expected to behave differently. During comprehension the super-lemma hypothesis predicts that some activation of the literal component lemmas is necessary for the super-lemma to become active and for the idiomatic meaning to be accessed. Thus we predict that during comprehension, contextual bias which encourages a *literal* interpretation of the relevant string may influence whether the idiomatic meaning is accessed. However, bias which encourages an *idiomatic* interpretation should still result in some consideration of the literal meaning. We present the results of two experiments designed to investigate how contextual factors (contextual bias and syntactic

incongruence) influence activation of idiomatic meaning during on-line comprehension.

2. Experiment 1: Syntactic Compatibility

Participants (n = 16) were eye-tracked while listening to sentences containing potentially idiomatic strings. Idioms of the form *verb x noun* (e.g. *kick the bucket, find her feet, smell a rat*) were selected based upon the results of an off-line norming study. To manipulate syntactic availability, each of our idioms was inserted into one of two sentential frames. In the Syntactically Available condition the relevant string was inserted into a simple sentence containing a proper name and a time phrase. In the Syntactically Unavailable condition idioms were inserted into a sentence pair such that the relevant string was divided by a sentential boundary (see Figure 1). These sentences along with 60 fillers represent our audio stimuli[1].

In addition to audio stimuli we constructed visual displays consisting of four words. For target items these words were an Idiom Associate, Literal Associate and two Distractors. Idiom Associates were chosen based upon the results of an off-line norming study that asked individuals to list the first three words that came to mind when considering the relevant idiom. Literal associates were associated with the literal meaning of either the verb or noun in the relevant string (see Nelson et al., 1998). The four words were presented all at once, with one word in each corner of the display (see Figure 2). Position of the associates and distractors was balanced.

For each trial, participants were first presented with the visual display and asked to read each of the four words aloud to encourage semantic activation (see van Orden et al, 1988). After reading the words aloud participants pressed a button that started the audio stimuli.

2.1 Experiment 1: Results & Discussion

Figure 3 shows the proportion of looks plots for Syntactically Available and Syntactically Unavailable conditions aligned to the onset of the critical noun (e.g. *bucket*). The patterns look strikingly similar, with both plots exhibiting greater looks to the Literal Associate than the Idiom Associate early in the time-course, followed by later competition between the Literal Associate and Idiom Associate. Statistical analyses confirm this observation, with significant or marginal differences over early time windows and competition in later time windows. This pattern of behavior suggests that individuals first consider the literal interpretation, and that later there is competition between the literal and idiomatic interpretations.

In the Syntactically Available case this result is unsurprising. Both the super-lemma hypothesis and the configuration hypothesis predict that literal processing will precede retrieval of the idiomatic meaning. That idiomatic meaning appears to be active even in the Syntactically Unavailable condition is more problematic. There are two possible explanations. First, it is possible that the late consideration of the idiomatic meaning is a post-processing effect (see Holsinger & Kaiser, 2010). Another possibility is that the view of the super-lemma representation acting as a firm gate to the conceptual layer is too strong. Instead the partial activation spreading from the component lemmas percolates to the conceptual layer despite the incompatible syntax.

3. Experiment 2: Contextual Bias

16 new participants participated in this study. In this experiment the idioms were inserted into the second of a pair of sentences intended to bias the interpretation of the idiom either toward a literal or idiomatic sense (see Figure 1). The former comprised our Literal Bias condition, and the latter our Idiom Bias condition. Aside from this change, the method and procedure were identical to the Experiment 1[2].

3.1 Experiment 2: Results & Discussion

Figure 3 shows the proportion of looks plots for Literally Biased and Idiomatically Biased

[1] In addition to the syntactic manipulation we also manipulated lexical availability by changing the final noun of each idiom to a semantically related noun (e.g. *kick the*

[2] The lexical manipulation from experiment 1 was also included; we again chose to focus on the unaltered idioms in the interest of brevity.

conditions aligned to the onset of the critical noun (e.g. *bucket*). The two plots appear to be mirror images, with more looks to the Idiom Associate throughout the time-course for Idiomatically Biased condition and more looks to the Literal Associate for the Literally Biased condition. Statistical analyses generally support this observation, with significant results over the full analysis region by items only and significant or marginal results over later time windows for Literally Biased trials and sporadically for Idiomatically Biased trials. This pattern of results suggests that contextual bias plays a strong role in the interpretation of these strings. Additionally the relatively weaker results obtained for the Idiomatically Biased trials suggest that literal context exerts a stronger influence on interpretation than idiomatic context.

Contextual bias is known to affect participants' interpretation of idioms (see Titone & Connie 1999). These results replicate these findings and provide a rich time-course of consideration. Additionally, our results suggest that literal bias is more effective driving interpretation than idiomatic bias. This result arises because participants consider the incongruent interpretation more strongly when faced with idiomatic bias than with literal bias. This result is predicted by the super-lemma hypothesis, since idiomatic bias should not be sufficient to fully suppress literal interpretation. Taken with the results of the previous experiment, these results also confirm our hypothesis that contextual bias will behave differently than syntactic incongruence

4. Conclusion

Our results are compatible with contemporary models of idiom representation. These results support the proposal that some degree of literal processing has priority, and is compatible with the view that literal processing is necessary for retrieval of idiomatic meaning. Crucially we found evidence of this literal priority even under contextual bias. We also confirmed our hypothesis that syntactic context influences individuals' processing behavior differently than contextual bias. However, the finding that incompatible syntactic context does not preclude consideration of idiomatic meaning argues against super-lemmas acting as a strong filter during comprehension. Regardless of whether this effect is interpreted as emerging during interpretation or as a post-processing phenomenon, revision of the representational content of the super-lemma and its role during idiom comprehension is required.

One possibility is that during comprehension the parser does not actively check the current structural context against the super-lemma representation. Given that many idioms have some degree of structural flexibility, it may be uneconomical for the parser to check against the potentially large set of possibilities immediately. This approach could explain why playful uses of language such as *his bucket was thoroughly kicked* are possible. During production, however, the super-lemma may be capable of playing a more direct role in shaping the structural configuration of the utterance. If this is the case we predict that syntactic context may play a stronger role when (i) the idiom is less flexible and hence there are less possibilities to check against or (ii) more processing resources are available.

Broadly, our results support a hybrid representation for idiomatic expressions. Idioms are represented as a function from simple lemmas to phrasal segments. These representations serve a dual purpose: they relate an idiom's conceptual meaning to its component lemmas and they provide a representational anchor for specifying information that pertains to the idiom, but not the component lemmas. For idioms this is essential, as features such as degree of syntactic flexibility are relevant only to the idiom. Applied more broadly, however, these intermediate representations can also provide a system for tracking features such as phrase frequency in a way that simultaneously allows for a certain degree of autonomy between the phrase and its component lemmas without divorcing the two entirely within the representational system

References

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67-82.

Bobrow, S., & Bell, S. (1973). On catching on to idiomatic expressions. *Memory and Cognition, 1*(3), 343-46.

Bod, R. (1998). *Beyond grammar: An experience-based theory of language.* Center for the Study of Language and Information.

Cacciari, C. & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language, 27*(6), 668-83.

Chomksy, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Cutting, J. & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition, 25*(1), 57-71.

Gibbs, R. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition, 8*(2), 149-156.

Gibbs, R., Bogdanovich, J., Sykes, J., & Barr, D. (1997). Metaphor in idiom comprehension. *Journal of Memory and Language, 37*(2), 141-154.

Gibbs, R., Nayak, N., & Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language, 28*(5), 576-593.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure.* Chicago UP.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford UP.

Holsinger, E. & Kaiser, E. (to appear). An experimental investigation of semantic and syntactic effects on idiom recognition. Proceedings of the 2010 *Western Conference on Linguistics (WECOL)*. Fresno, CA, November 2010

Konopka, A., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology, 58*(1), 68-101.

Kuiper, K., van Egmond, M., Kempen, G. & Sprenger, S. (2007). Slipping on superlemmas: Multi-word lexical items in speech production. *The Mental Lexicon, 2*(3), 313-357.

McGlone, M., Glucksberg, S., & Cacciari, C. (1994). Semantic productivity and idiom comprehension. *Discourse Processes, 17*, 167-190.

Nelson, D., McEvoy, C., & Schreiber, T. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

Ortony, A., Schallert, D., Reynolds, R., & Antos, S. (1978). Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning and Verbal Behavior, 17*(4), 465–477.

Peterson, R., Burgess, C., Dell, G., & Eberhard, K. (2001). Dissociation between syntactic and semantic processing during idiom comprehension *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1223-37.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (eds.) *Frequency effects and the emergence of lexical structure*. Amsterdam: John Benjamins. 137-57.

Sprenger, S., Levelt, W., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language, 54*(2), 161-84.

Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18*(5), 523-34.

Tabossi, P., Fanari, R., & Wolf, K. (2008). Processing idiomatic expressions: effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(2), 313-27.

Tabossi, P, Wolf, K. & Koterle, S. (2009). Idiom syntax: Idiosyncratic or principled Journal of Memory and Language. *Journal of Memory and Language, 61*, 77-96.

Tabossi, P. & Zardon, F. (1993). The activation of Idiomatic Meaning in Spoken Language Comprehension. In C. Cacciari & P. Tabossi (eds.) *Idioms: Processing, Structure and Interpretation*. Hillsdale, NJ: Lawrence Erlbaum Associates. 145-62.

Titone, D. & Connine, C. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics, 31*, 1655-74.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Harvard UP.

Van Orden, G., Johnston, J. & Halle, B. (1988). Word identification in reading proceeds from spelling to sound to meaning. Journal of Experimental Psychology: *Learning, Memory & Cognition, 14*, 371-86.

Weinreich, U. (1969). Problems in the analysis of idioms. In J. Puhvel (ed.), *Substance and structure of language*. University of California Press. 23-81.

Figures

Example Stimuli	
Experiment 1	**Syntactically Unavailable**
	It was surprising to see someone as skilled as John completely miss the ball when he kicked. The bucket full of orange slices was destroyed when he accidentally missed the ball.
	Syntactically Available
	Mary kicked the bucket last Thursday evening.
Experiment 2	**Idiomatically Biased**
	Swimming with sharks is a dangerous and unpredictable profession. As a result of the shark attack several oceanographers kicked the bucket last Thursday evening.
	Literally Biased
	John spent all day filling things with cement as a nasty prank. Several people broke their toes when they kicked the bucket last Thursday evening and may sue.

Figure 1. Example Stimuli for Experiments 1 and 2

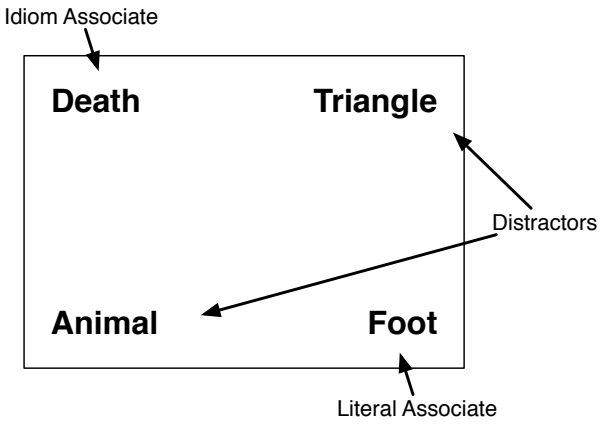

Figure 2. Sample Display for the Idiom *kick the bucket*

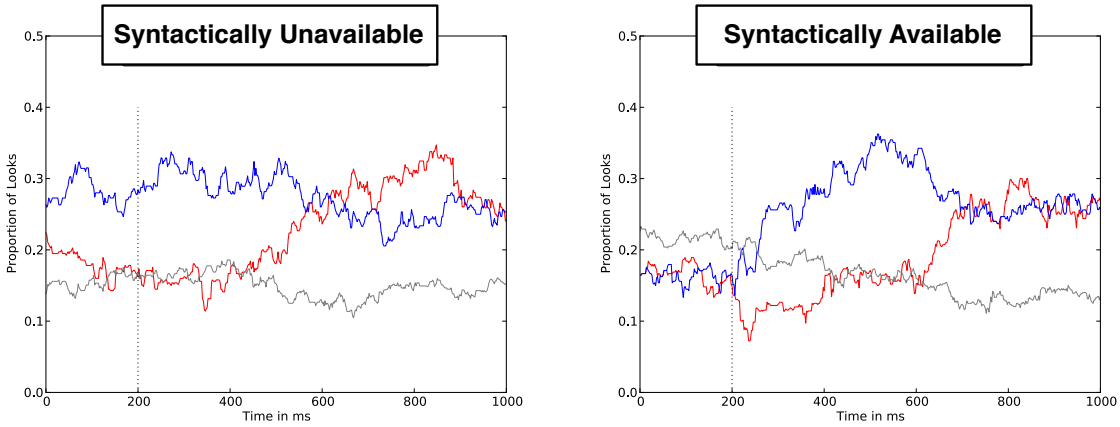

Figure 3. Proportion of looks for Syntactically Unavailable (left) and Syntactically Available (right) conditions aligned to the onset of the noun (e.g. *bucket*). Looks to the Idiom Associate (e.g. *death*) are in red, looks to the Literal Associate (e.g. *foot*) are in blue, and averaged distractors are in grey.

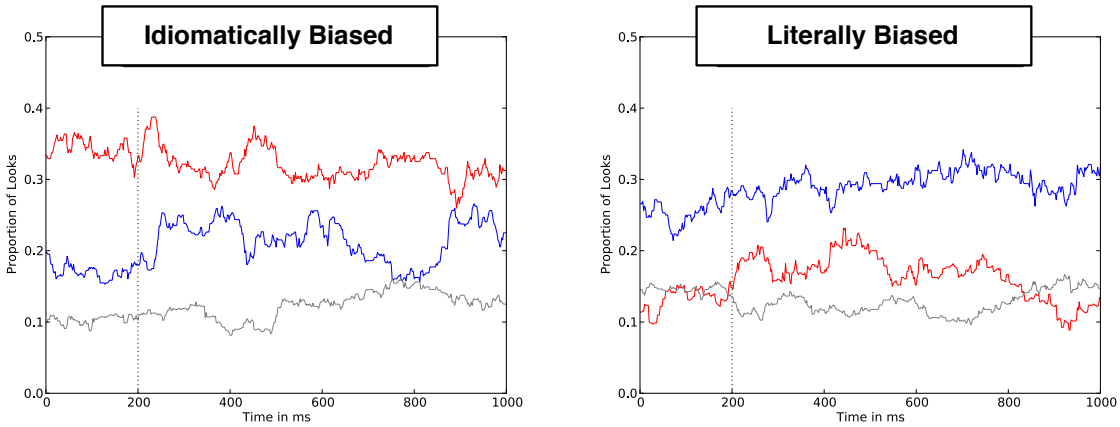

Figure 4. Proportion of looks plots for Idiomatically Biased (left) and Literally Biased (right) conditions aligned to the onset of the noun (e.g. *bucket*). Looks to the Idiom Associate (e.g. *death*) are in red, looks to the Literal Associate (e.g. *foot*) are in blue and the averaged distractors are in grey.