The Construction of Meaning

Walter Kintsch & Praful Mangalath
University of Colorado*

We argue that word meanings are not stored in a mental lexicon but are generated in the context of working memory from long-term memory traces which record our experience with words. Current statistical models of semantics, such as LSA and the Topic Model, describe what is stored in long-term memory. The CI-2 model describes how this information is used to construct sentence meanings. This model is a dual-memory model, in that it distinguishes between a gist level and an explicit level. It also incorporates syntactic information about how words are used, derived from dependency grammar. The construction of meaning is conceptualized as feature sampling from the explicit memory traces, with the constraint that the sampling must be contextually relevant both semantically and syntactically. Semantic relevance is achieved by sampling topically relevant features; local syntactic constraints as expressed by dependency relations ensure syntactic relevance.

In the present paper, we are concerned with how meaning can be inferred from the analysis of large linguistic corpora. Specifically, our goal is to present a model of how sentence meanings are constructed as opposed to word meanings per se. To do so, we need to combine semantic and syntactic information about words stored in long-term

memory with local information about the sentence context. We first review models that extract latent semantic information from linguistic corpora and show how that information can be contextualized in working memory. We then present a model that uses not only semantic information at the gist level, but also explicit information about the actual patterns of word use to arrive at sentence interpretations.

It has long been recognized that word meanings cannot just be accessed full-blown from a mental lexicon. For example, a well-known study by Barclay, Bransford, Franks, McCarrell, and Nitsch (1974) showed that *pianos are heavy* in the context of moving *furniture*, but that they are *musical* in the context of *Arthur Rubenstein*. Findings like these have put into question the notion that meanings of words, however they are represented (via semantic features, networks, etc.), are stored ready-made in the mental lexicon from which they are retrieved as needed. Rather, it appears that meanings are generated when a word is recognized in interaction with its context (Barsalou, 1987). Indeed, there seems to be no fixed number of meanings or senses of a word; new ones may be constructed as needed (Balota, 1990). Furthermore, if meanings were pre-stored, then memory theory would have difficulty explaining how the right meaning and sense of a word could be retrieved quickly and efficiently in context (Klein & Murphy, 2001). Problems such as these have led many researchers to reject the idea of a mental lexicon in favor of the claim that meaning is contextually constructed. Such a view seems necessary to account for the richness of meaning and its emergent character. The question is: How does one model the emergence of meaning in context.

There are various ways to approach this problem (e.g., Elman, 2009). We focus here on statistical methods to infer meaning from the analysis of a large linguistic corpus. Such models are able to represent human meaning on a realistically large scale, and do so without hand coding. For example, latent semantic analysis (LSA, Landauer & Dumais, 1997) extracts from a large corpus of texts a representation of a word's meaning as a decontextualized summary of all the experiences the system has had with that word. The representation of a word reflects the structure of the (linguistic) environment of that word. Thus, machine-learning models like LSA attempt to understand semantic structure

by understanding the structure of the environment in which words have been used. By analyzing a large number of texts produced by people, these models infer how meaning is represented in the minds of the persons who produced these texts.

There are two factors that make models like LSA attractive: One is scale, because an accurate model of something as complex as human meaning requires a great deal of information--the model must be exposed to roughly the same amount of text as people encounter if it is to match their semantic knowledge; the other is representativeness. By analyzing an authentic linguistic corpus that is reasonably representative for a particular language user population, one ensures that the map of meaning that is being constructed is unbiased, thereby emphasizing those aspects of language that are relevant and important in actual language use.

LSA and the other models discussed below abstract from a corpus a blueprint for the generation of meaning--not a word's meaning itself. We argue for a generative[i] model of meaning that distinguishes between decontextualized representations that are stored in long-term memory and the meaning that emerges in working memory when these representations are used in context. Thus, the generative model of meaning that is the focus of this paper has two components: the abstraction of a semantic representation from a linguistic corpus, and the use of that representation to create contextually appropriate meanings in working memory.

Long-term memory does not store the full meaning of a word, but rather stores a decontextualized record of experiences with a particular word. Meaning needs to be constructed in context, as suggested by Barclay et al.'s *piano* example. The record of a lifetime's encounter with words is stored in long-term memory in a structured, well-organized way, for example as a high-dimensional semantic space in the LSA model. This semantic space serves as a retrieval structure in the sense of Ericsson and Kintsch (1995). For a given word, rapid, automatic access is obtained to related information in long-term memory via this retrieval structure. But not all information about a word that has been stored is relevant at any given time. The context in which the word appears

determines what is relevant. Thus, what we know about *pianos* (long-term memory structure) and the context (*furniture* or *music*) creates a trace in long-term working memory that makes available the information about *pianos* that is relevant in the particular context of use. From this information, the contextual meaning of *piano* is constructed. Meaning, in this view, is rich and forever varied: every time a word is used in a new context, a different meaning will be constructed. While the difference in meaning might only be slight at times, it will be significantly varied at others (as when a word is used metaphorically). Recent semantic models such as the Topic Model (Griffiths & Steyvers, 2004; Steyvers & Griffiths, 2007; Griffiths, Steyvers & Tenenbaum, 2007) derive long-term traces that explicitly allow meaning to be contextualized. We use insights from their work to construct a model for sentence interpretation that includes syntactic information. We discuss below how sentence meaning is constructed in working memory. But first we review several alternative models that describe how lexical information can be represented in long-term memory.

The representation of semantic knowledge in long-term memory

For a large class of cases – though not for all – in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in language.
Wittgenstein (1953)

A word is characterized by the company it keeps
Firth (1957)

Language is a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others.
Saussure (1915)

One way to define the meaning of a word is through its use, that is, the company it keeps with other words in the language. The idea is not new, as suggested by the quotations above. However, the development of modern machine learning algorithms was

necessary in order to automatically extract word meanings from a linguistic corpus that represent the way words are used in the language.[ii]

There are various ways to construct such representations. Typically, the input consists of a large linguistic corpus, which is representative of the tasks the system will be used to model. An example would be the TASA corpus consisting of texts a typical American high-school student might have read by the time he or she graduates (see Quesada, 2007, for more detail). The TASA corpus comprises 11M word tokens, consisting of about 90K different words organized in 44K documents. The corpus is analyzed into a word-by-document matrix, the entries of which are the frequencies with which each of the words appears in each document. Obviously, most of the entries in this huge matrix are 0; that is, the matrix is very sparse. From this co-occurrence information, the semantic structure of word meanings is inferred. The inference process typically involves a drastic reduction in the dimensionality of the word-by-document matrix. The reduced matrix is no longer sparse, and this process of generalization allows semantic similarity estimates between all the words in the corpus. We can now compute the semantic distance between two words that have never co-occurred in the corpus. At the same time, dimension reduction also is a process of abstraction: inessential information in the original co-occurrence matrix has been discarded in favor of what is generalizable and semantically relevant.

In Latent Semantic Analysis (Landauer & Dumais, 1997; Martin & Berry, 2007) dimension reduction is achieved by decomposing the co-occurrence matrix via Singular Value Decomposition and selecting the 300 or so dimensions that are most important semantically. A word is represented by a vector of 300 numbers that are meaningless by themselves but which make it possible to compute the semantic similarity between any pair of words. Locating each word in the 300-dimension semantic space with respect to every other word in the semantic space specifies its meaning via its relationship to other words. Furthermore, vectors in the same semantic space can be computed that represent the meaning of phrases, sentences, or whole texts, based on the assumption that the meaning of a text is the sum of the word vectors in the text. Thus, semantic distance

estimates can be readily obtained for any pair of words or texts, whether or not they have ever been observed together. This feature makes LSA extremely useful and has made possible a wide range of successful applications, both for simulating psycholinguistic data and for a variety of practical applications (Landauer, McNamara, Dennis, & Kintsch, 2007; http://lsa.colorado.edu).

The Topic Model (Griffiths & Steyvers, 2004; Steyvers & Griffiths, 2007; Griffiths, Steyvers & Tenenbaum, 2007) represents the gist of a document as a distribution over latent variables called topics. A topic is a probability distribution over words. Documents are generated by choosing words from the distribution of words associated with the topics that represent a document. First a topic is sampled from the set of topics that represents the gist of a document, and then a word is sampled from the probability distribution of words for that topic. This generative procedure is reversed to infer the distribution of topics for the documents in a corpus and the distribution of words for topics.

Topics are frequently individually interpretable: for instance, Griffiths et al. (2007) find a "*printing*" topic characterized by words like *printing, paper, press, type, process, ink*, etc., and an "*experiment*" topic characterized by *hypothesis, experiment, scientific, observation, test*, and so on. Different senses or meanings of a word may be assigned to different topics. Thus, the word *play* has a high probability not only for the *theater* topic, but also for the *sports* topic.

In Table 1 the 20 nearest neighbors of the word *play* are shown both for LSA and the Topic Model. To find the nearest neighbors in LSA, the cosines between *play* and all the words in the corpus are computed and the 20 words with the highest cosine are chosen. Words that appeared fewer than 10 times in the corpus were excluded from the LSA as well as Topic computations in order to reduce noise. The 20 words that have the highest conditional probability in the Topic analysis of the TASA corpus given *play* are also shown in Table 1. The two sets of words are obviously related: six of the words are in common. The LSA neighbors appear to fall into two clusters, corresponding to the

theater sense of *play* and the *sports* sense. The Topic Model yields three clusters: the *theater* and *sport* senses, like LSA, but also a *children-play* sense, containing words like *children, friends,* and *toys*. (These words are, of course, also strongly related to *play* in LSA, they just do not make the top 20).

Table 1

LSA and the Topic Model are by no means the only ways to infer semantic structure from a linguistic corpus. Quite a different approach has been taken in the BEAGLE model of Jones and Mewhort (2007) who model word meaning as a composite distributed representation by coding word co-occurrences across millions of sentences in a corpus into a single holographic vector per word. This vector stores a word's history of co-occurrence and usage in sentences similar to Murdock's (1982) model of episodic memory for word lists. Jones & Mewhort simulate a wide range of psycholinguistic data with their Holograph Model.

Kwantes (2005) proposes a context model of semantic memory which assumes that what is stored is basically a document-by-term matrix as in LSA, but then does not use dimension reduction to compute the vector representing a word but a resonance process modeled after the episodic memory model of Hintzman (1984). A few illustrations are given which demonstrate the ability of this model to simulate human semantic judgments.

In the Hyperspace Analogue to Language model (HAL) of Burgess and Lund (2000) a semantic representation is constructed by moving an n-word window across a text and recording the number of words separating any two words in the window. This procedure essentially weights co-occurrence frequencies by their distance. Unlike most of the methods described above it does not involve dimension reduction.

Note the diverse origins of these models. LSA has its roots in the literature on information retrieval. The Topic Model comes from Bayesian statistics. The other two

approaches mentioned here are semantic memory counterparts of well-established episodic memory models: Jones and Mewhort's holographic model extends Murdock (1982), and Kwantes elaborates Hintzman (1984). Murdock and Hintzman make very different assumptions about the nature of memory – a single holographic trace for Murdock versus multiple traces for Hintzman. It is interesting that both approaches could be extended to semantic memory, suggesting that episodic and semantic memory may differ not so much in their architecture, but more in the nature of what is being stored.

The corpus that is the input to whatever analysis is performed determines the nature of the representation obtained. The TASA corpus, for instance, corresponds roughly to texts a high-school graduate might have read. That is, it serves well to simulate students at that level. But if one wants to analyze texts requiring specialized knowledge, a corpus representative of that knowledge would have to be analyzed. Such a corpus would have to be of a sufficient size to yield reliable results (see Quesada, 2007).

Thus, there are many methods to extract, without supervision, a semantic representation from a linguistic corpus. The question which of the current models is best has no straightforward answers. . They all produce qualitatively similar results, but they are not equivalent. Depending on the task at hand, some models may be more suited than others. For instance, conditional probabilities in the Topic Model can capture some of the asymmetries present in human similarity judgments and association data, which the distance measure used in LSA cannot (Griffiths et al., 2007). The holograph model yields data on the growth of concepts as it is exposed to more and more texts, which is not possible with LSA (Jones & Mewhort, 2007). On the other hand, LSA has been the most successful model for representing text meanings for essay grading (Landauer, Laham, & Foltz, 2003) and summary writing (E. Kintsch et al., 2007). Thus, while particular models might be better suited to certain functions, they all get at a common core of meaning by extracting semantic information from the co-occurrences of words in documents or sentences. However, there is more information in a corpus than that – specifically, information about word order and syntactic structure. Taking into account these

additional sources of information makes a whole lot of difference. We shall describe two such approaches, one using word order and the other using syntax.

The holograph model of Jones and Mewhort (2007) already mentioned is capable of taking into account word order information, in addition to co-occurrence information. Jones and Mewhort use order-sensitive, circular convolution to encode the order in which words occur in sentences. Thus, their model distinguishes between two kinds of information, like Murdock's episodic memory model: item information - which words are used with which other words in the language - and order information, which encodes some basic syntactic information. The model therefore can relate words either semantically (like LSA) or syntactically (clustering by parts of speech, for instance). This dual ability greatly extends the scope of phenomena that can be accounted for by statistical models of meaning. Jones, Kintsch, and Mewhort (2006) have compared the ability of the holograph model, HAL, and LSA to account for a variety of semantic, associative, and mediated priming results. They demonstrate that both word context and word order information are necessary to explain the human data.

Another source of information in a corpus is provided by the syntactic structure of sentences. While word order, especially in a language like English, can be considered an approximation to syntax, there is more to syntax than just that. Statistical models of semantics can exploit this additional information. A model that successfully extracts information about syntactic patterns as well as semantic information from a linguistic corpus is the Syntagmatic-Paradigmatic (SP) model of Dennis (2005). The SP-model is a memory model with a long-term memory that consists of all traces of sentences that have been experienced. It analyzes these traces for their sequential and relational structures. Syntactic information is captured by syntagmatic associations: the model notes which words follow each other (e.g., *drive-fast, deep-water*). Semantic information is captured by paradigmatic associations: words that are used in the same slot in different sentences (e.g., *fast-slow, deep-shallow*). Combining sequential and relational processing in this way enables the SP model to capture the propositional content of a sentence. Hence, the model can answer questions that depend upon that understanding. For instance, after

processing a set of articles on professional tennis, the model performed quite well when asked questions about who won a particular match– but only if the relational and sequential information were combined. After sequential processing, the model answered with some player's name, but only 8% of the time with the correct name, because although it knew that a player's name was required, it was not aware of who played against whom, and who won and who lost. When it was given that information - the relational or paradigmatic associations - the percentage of correct answers rose to 67%.

Being able to register propositional content has important consequences for inferencing. In the following example from Dennis and Kintsch (2008), the SP-model offers a compelling account of what has been called inference by coincidence: Given the sentence "*Charlie bought the lemonade from Lucy,*" people know right away that "*Lucy sold the lemonade to Charlie*" and that "*Charlie owns the lemonade*." According to the SP-model, this is not because an explicit inference has been made, but is rather a direct consequence of understanding that sentence: Understanding a sentence means aligning it with similar sentences that are retrieved from long-term memory. Thus, the Charlie sentence will be aligned with a set of sequential memory traces of the form ***A*** *buy* ***B*** *from* ***C*** where **A** is a set of *buyers*, **B** is a set of *objects bought and sold*, and **C** is a set of *sellers*. Since the same sets also appear in sequential traces of the form ***C*** *sells* ***B*** *to* ***A***, and ***A*** *owns* ***B***, the model knows that *Charlie* belongs to the set of what we call *buyers* and *owners*. Note that sets of words like **A** are extensionally defined as the words that fit into a particular slot: a list of words is generated that are the Agents of *buy*, or the Objects, but semantic roles are not explicitly labeled.

Thus, while the differences between the various bag-of-words models we have discussed are relatively superficial, taking into account word order information or syntactic structure allows statistical models to account for a whole new range of psycholinguistic phenomena at an almost human-like scale. The great strength of all of these models is that they are not toy models and do not depend on hand coding, but nevertheless allow us to model many significant aspects of how people use their language. But do the representations they generate really reflect the meaning of words as that term is commonly understood? These models all decontextualize meaning – they

abstract a semantic representation from a large corpus that summarizes the information in the corpus about how that word has been used in the context of other words. But words mean different things in different contexts, often totally different things, as in the case of homonyms. Even words that have only one meaning have different senses when used in different contexts. Meaning, one can argue, is always contextual. Words do not have meaning, but are clues to meaning (Rumelhart, 1979, as discussed by Elman, 2009).

The construction of word meaning in working memory

How meaning is constructed can be illustrated with reference to LSA. LSA represents the meaning of *play* as a single vector; then why do we understand *play* in one way in the context of *theater* and in another way in the context of *baseball*? We briefly sketch the Predication model of Kintsch (2001; 2007; 2008a; 2008b) that addresses this problem when the context is a single word, before extending it to sentence contexts.

In WordNet (*http://wordnet.princeton.edu*), *bark* has three unrelated meanings, with four senses each for the tree-related *bark* as well as the dog-related *bark*, and a single sense for the ship-related *bark*. In LSA a single vector represents the meanings and senses of *bark*. Thus, LSA by itself does not distinguish between the different meanings and senses of a word. The Predication Model of Kintsch (2001) describes a process that brings about appropriate word senses when a word is used in context from an LSA vector that combines all meanings and senses, generating a context-appropriate word sense. It allows the context to modify word vectors in such a way that their context-appropriate aspects are strengthened and irrelevant ones are suppressed. In the Construction-Integration (CI) model of Kintsch (1998), discourse representations are built up via a spreading activation process in a network defined by the concepts and propositions in a text. Meaning construction in the Predication Model works in a similar way: a network is constructed containing the word to be modified and its semantic neighborhood, and is linked to the context; spreading activation in that network assures that those elements of the neighborhood most strongly related to the context become activated and are able to modify the original word vector. For instance (Kintsch, 2008a), consider the meaning of *bark* in the context of *dog* and in the context of *tree*. The semantic neighborhood of *bark*

includes words related to the *dog*-meaning of *bark*, such as *kennel*, and words related to the *tree*-meaning, such as *lumber*. To generate the meaning of *bark* in the context of *dog*, all neighbors of *bark* are linked to both *bark* and *dog* according to their cosine values. Furthermore, the neighbors themselves inhibit each other in such a way that the total positive and negative link strength balances. As a result of spreading activation in such a network, words in the semantic neighborhood of *bark* that are related to the context become activated, and words that are unrelated become deactivated. Thus, in the context of *dog*, the activation of *kennel* increases and the activation of words unrelated to *dog* decreases; in the context of *tree*, the activation values for *lumber* increases, while *kennel* is deactivated. The contextual meaning of $bark_{dog}$ is then the centroid of *bark* and its dog-activated neighbors, such as *kennel*; that of $bark_{tree}$ is the (weighted) centroid of *bark* and neighboring words like *lumber*. $Bark_{dog}$ becomes more *dog*-like and less *tree*-like; the opposite happens for $bark_{tree}$. Two distinct meanings of *bark* emerge: using the six most highly activated neighbors to modify *bark* from a neighborhood of 500, the cosine between $bark_{tree}$ and $bark_{dog}$ is only .03. Furthermore, $bark_{dog}$ is no longer related to *tree*, cos = -.04, and $bark_{tree}$ is no longer related to *dog*, cos = .02. Thus, context-appropriate word meanings can be generated within a system like LSA, in spite of the fact that what LSA does is to construct context-free word vectors.

Predication also generates meaning that is metaphorical rather than literal (Kintsch, 2008b). For example, the meaning of *shark* in the context of *My lawyer is a shark* can be computed by predication. The neighbors of *shark* activated by *lawyer* include *vicious, dangerous, greedy,* and *fierce*. These nodes are combined with the *shark* vector to generate a new concept $shark_{lawyer}$ whose fishiness has been de-emphasized, but its dangerous character has been retained: the closest neighbors of $shark_{lawyer}$ are *danger, killer, frighten, grounds, killing* and *victims*.[iii] It should be noted, though, that not all metaphors are as simple as in this example. Frequently, the process of metaphor interpretation is much more complex, involving analogical reasoning and not just meaning transfer (for a discussion see Kintsch, 2008b).

While predication has been modeled as a spreading activation network in the CI model, there are some disadvantages to this approach. Introducing a spreading activation network is computationally complex and requires free parameters (how many neighbors are to be searched? what is the activation threshold?). Shifting to a probabilistic model avoids these problems. In the present context, this means replacing LSA with the Topic Model. Meaning is context sensitive in the Topic model: *play* will be assigned to different topics in *theater* and *baseball* documents. Hence, predication with topic features amounts to calculating the conditional probability of each word, given both the argument and the predicate word. Thus, to predicate *Shakespeare* about *play*, we calculate the conditional probability of words given *play*∩*Shakespeare.* Table 2 shows the 20 words with the highest conditional probability given *play*∩*Shakespeare* and *play*∩*baseball.* Predicating *Shakespeare* about *play* neatly picks out the *theater*-related words from the neighbors of *play* and adds other related words. Predicating *baseball* about *play* picks out the *sports*-related words from the neighbors of *play* and adds other sports words. Thus, the mechanism by which predication is modeled in the Topic Model is quite different (and simpler), but it has qualitatively similar results: it effectively contextualizes the meaning of words.

Readers familiar with the Topic model should note that in our assessment and usage of vector representation for words, we simply use the static rows from the word-topic matrix that is a result of applying the inference procedure. Each topic has a multinomial distribution over words and topics are independent. Our word vectors only capture this uncorrelated association across multiple topics and can be seen as multidimensional vectors with each dimension representing the probability of the word conditioned on a particular topic. The Topic Model is designed to be used with more sophisticated techniques capable of deriving context sensitive representations for words. Traditionally the Topic model treats words or propositions as a new document and a disambiguated sharper distribution over topics is inferred using the Gibbs sampling procedure described in Griffiths and Steyvers (2004). For computational tractability the model presented here uses only static representations of words from the initial estimation phase.

Table 2

Predication, then, is a model for a generative lexicon; it is an algorithm for the construction of word meanings when words are used in context. Words are polysemous, but in the models described above their representation in long-term memory is context-free; predication constructs context-appropriate meanings in working memory on the fly, without having to specify word meanings and senses beforehand, as in a mental lexicon. A new meaning for a word emerges every time it is used in a different context. Every time a word is used its meaning will be different – a little different when used in an ordinary way, more so when used metaphorically. Thus, the vector with which LSA represents the meaning of *shark* is altered very little in the context of *swim*, more so in the context of *soup*, and even more in the context of *lawyer*.

Context, in the Construction-Integration (CI) model described above, has only been that provided by another word. In normal language use, however, words are used as part of a sentence and discourse. Sentence structure constrains understanding; syntax determines how a sentence is interpreted. Below we describe a new model that extends the predication model to sentence contexts.

Sentence Meaning

The CI-II model of Mangalath (in preparation) combines ideas from the approaches discussed above with additional notions from the memory and from text comprehension literature. The principal components of the model are:

a. Meaning is constructed in context in working memory from information stored in long-term memory, as in the Construction-Integration model;
b. The general framework is that of the Topic Model;
c. The syntagmatic-paradigmatic distinction is taken over from the Syntagmatic-Paradigmatic-model;
d. A new element is the use of Dependency Grammar to specify syntactic structure;
e. Also new is the distinction between different levels of representation, allowing for a coarse-grained gist and a fine-grained explicit representation.

We first discuss the arguments for introducing dependency grammar and a dual-memory trace and then outline the proposed model.

Syntactic structure. Syntax clearly needs to be considered in meaning construction – but how? There are different conceptions of the role that syntactic analysis plays in human comprehension. For the most part, psycholinguists have assumed that comprehension involves syntactic analysis together with a variety of other knowledge sources to arrive at an accurate and detailed interpretation of a sentence that corresponds more or less to its linguistic representation. There is good reason to believe that human sentence processing is more shallow and superficial than that. Ferreira & Patson (2007) have termed this the "good enough" approach to language comprehension. They argue that comprehenders form representations that are good enough for the task at hand – to participate in a conversation, to answer a question, to update their knowledge – but mental representations are typically incomplete and not infrequently incorrect.[iv] Thus, a model of comprehension should emulate the process of actual human comprehension rather than emulating linguistic analysis. If, indeed, the "comprehension system works by cobbling together local analyses" (Ferreira & Patson, 2007, p.74), the question arises how to model the priority of local analysis.

Current trends in linguistics and psycholinguistics offer useful suggestions. Linguistic theory has taken a turn towards focusing on lexical items together with their associated syntactic and semantic information (Lexicalist-Functionalist Grammar, Bresnan & Kaplan, 1982; Combinatory Categorical Grammar, Steedman, 1996; Tree-Adjoining Grammar, Joshi, 2004; Construction Grammar, Goldberg, 2006). At the same time, psycholinguists have recognized the importance of usage–based patterns with overlapping semantic, syntactic and pragmatic properties in language acquisition as well as language processing (e.g., Garrod, Freudenthal, & Boyle, 1994; MacDonald, Perlmutter, & Seidenberg, 1994; MacDonald & MacWhinney, 1995; Bates & Goodman, 2001; Tomasello, 2001; 2003). Thus, what is needed, is a formalism to decide (a) what the relevant phrasal units are that are combined in a sentence, and (b) a way to construct the meaning of phrasal units from the word vectors of statistical semantics.

Ideally, of course, a model should be able to learn both the latent semantic structure and the syntactic structure from a corpus and to do so without supervision. The CI-II model, however, does not do that, in that it does not specify how syntax is learned. Instead, it focuses on how syntax, once it has been learned, is used in sentence comprehension and production. A suitable grammar that specifies how the words in a sentence are related is Dependency Grammar. As the name implies, Dependency Grammar (Tesniere, 1959; Sgall, Hajicova, & Panevova, 1986; Mel'cuck, 1988) focuses on the dependency relations within a sentence – which is the kind of information one needs in order to specify context in the predication model. Compared with other grammars, Dependency grammar is austere, in that it does not use intermediate symbols such as noun phrase or verb phrase, nor does it use semantic role labels such as agent or object. It merely specifies which word in a sentence depends on which other word, and the part of speech of every word. Automatic dependency parsers (Yamada & Matsumoto, 2002; Nivre et al., 2007) available from the web can be used to perform these analyses. Figure 1 shows an example of a dependency tree overlaid with a propositional analysis of the sentence after Kintsch (1974, 1998). As the example illustrates, dependency units correspond to propositions or parts of propositions. For example, the proposition FLOOD[RIVER,TOWN] in Figure 1 is composed of the dependencies RIVER ← FLOOD and FLOOD → TOWN. Thus, by analyzing the dependency structure of a sentence, information is gained about its propositional structure, which according to many theorists, including Kintsch (1974,1998), is what really matters in comprehension.

Figure 1

Levels of representation. The Topic model (or for that matter, LSA) provides a coarse-grained semantic representation; it is designed to capture the essence – the gist – of a word's meaning, but not all its detail. To represent the TASA corpus, LSA uses 300 dimensions and anywhere between 500 to 1700 topics are needed with the Topic model. In all our experiments, we use a 1195 topic estimate which yielded performance comparable to the results reported in Griffiths, Steyvers & Tenenbaum (2007) . However,

this is not sufficient to specify the finer details of word meanings. Typically a word loads only on a relatively small number of topics. For example, *river* loads on only 18 topics; that number is sharply reduced when words are combined: *river* in the context of *flood* involves 4 topics; – too sparse a representation to be of much use to represent meaning in all its detail. Steyvers & Griffiths (2008) make a similar argument in the context of memory and information retrieval. This is not a defect of the model, but simply a consequence of its goal – to identify the cluster structure (topics) of the semantic space. The gist level representation, whether LSA or Topics, is useful for what it was designed for, but it needs to be complemented by a representation that specifies in detail how words are used. An obvious candidate would be the word-by-word matrix that specifies how often a word is used with another word of the corpus in the same sentence. If we were to cluster such a matrix, we would come up with something similar to the topic structure derived from the word-by-document matrix. But in its unreduced form, the word-by-word matrix provides the kind of detail we need. For the TASA corpus, this means that we now have a second way to represent a word meaning, by means of the 58k words in the corpus. Both relational and sequential information can be captured in this way. The number of times words co-occur in a document yields relational information, but a similar word-by-word matrix based on the frequencies a word co-occurs with other words in a dependency unit can be used to specify how the word is used syntactically.[v] We call this the explicit relational and sequential representation, to distinguish it from the gist-level topic representation. For the word *river*, for instance, the explicit relational trace contains 2,730 items with a non-zero probability and the explicit sequential trace contains 132 items.

It has long been recognized that language involves processing at a general semantic as well as a verbatim level. In the CI-Model of text comprehension, different levels of representation are distinguished, including a verbatim surface structure and a propositional textbase (e.g., Kintsch, 1998). Similarly, theories of memory have postulated dual processes, at the level of gist and at the verbatim level (e.g., Brainerd, Wright, & Reyna, 2002). Statistical models of language need to take into account both latent semantic structure and verbatim form. Latent semantic structure is inferred from

the co-occurrence of words in documents by means of some form of dimension reduction; this provides the gist-level information. In addition, explicit information, both relational and sequential, about how a word is used with other words, is needed to allow for a finer-grained analysis, including the ability to deal with syntactic patterns.

The long-term memory store. The input to LTM is a linguistic corpus such as the TASA corpus, which consists of 44k documents and 58k word types (excluding very rare and very frequent words which are both uninformative for our purposes) for a total of 11m words. The corpus is analyzed in three ways.

First, a word-by-document matrix is constructed, whose entries are the frequencies with which each word appears in each of the documents. Dimension reduction is used to compute the **gist trace** for a word – the 300-dimensional LSA vector, or the 1195-dimensional topics vector.

The second analysis is based on constructing from the corpus a word-by-word matrix, the entries of which are frequencies with which each word has co-occurred with every other word in a document (or sentence). Probabilities are estimated from the frequency counts using the Pitman-Yor process (Teh, 2006), in effect normalizing and smoothing the data. The **explicit relational trace** is thus a vector of length 58k of word-word co-occurrence counts.

The third analysis is performed by first parsing the entire corpus with a dependency grammar parser, the MALT parser of Nivre (Nivre et al., 2007). The **explicit sequential trace** is a vector of length 58k, the entries of which are association probabilities corresponding to how often a word has been used in a dependency unit with another word in the corpus, again using the Pitman-Yor process. Actually, there are two such traces: one for words on the right side of a dependency unit and one for the left side.

The construction of meaning in working memory. The context words in the explicit relational vector are the features which will be used for the construction of

meaning in working memory. A language model is constructed in working memory by sampling these features, subject to local semantic and syntactic constraints. We sketch this procedure for words out of context, words in the context of other words, dependency units, and sentences.

For words out of context, the topic vector (alternatively, the LSA vector) provides a ready representation at the gist level. For many purposes this representation is all that is needed. When a more detailed meaning representation is required, the explicit relational trace is used. However, by itself that vector is so noisy as to be useless: a word co-occurs with many different words that are not semantically related with it. Therefore, a language model needs to be constructed by sampling context words from the explicit relational trace that are semantically relevant. This is achieved by allowing the topic representation of the word W to guide the feature sampling process. A topic is sampled at random from the distribution of topics for W. Then a context word w_i is sampled according to the probability that the selected topic has for all the words in the explicit relational trace. The sampled feature's weight is determined by the probability of selecting that particular topic in the first place (in our experiments all topics are equiprobable); by the probability of the selected word w_i for the topic; and by the probability of the selected word w_i in the relational trace of W. Repeating this sampling procedure and averaging samples generates a feature vector, where the features are the sampled context words. For example, consider the meaning of *kill*, out of context (Figure 2). The 1195-dimensional topics vector for *kill* has 24 non-zero entries and is the gist representation for *kill*. To generate the explicit representation, a topic is selected and a context word w_i is sampled according to its probability on the selected topic and weighted appropriately. The explicit relational vector for *kill* has 58k dimensions, 11,594 with a non-zero count.[vi] The language model generated by sampling topically related features also has 58k dimension, but only 1,711 non-zero entries. In other words, *kill* has occurred in a document with 11,594 other words in our corpus, but only 1,711 of these are topically related to *kill*. These then form the explicit meaning representation for *kill* out of context. Among the most strongly weighted features are the words *insects, chemicals, poison, hunter,* and *pest*.

Figure 2

Now consider predication, that is, the construction of meaning in the context of another word. Exactly the same sampling procedure is used to generate a language model except that the topics controlling the sampling process are now selected from the context word. For instance (Figure 3), to generate the explicit representation of *kill* in the context of *hunter*, topics are sampled from the topic distribution for *hunter* (*hunter* loads on 8 topics). Features are then selected from the explicit relational trace of *kill* constrained by the *hunter*-topics. This generates a vector with 1366 non-zero entries to represent $kill_{hunter}$ (the meaning of *kill* in the context of *hunter*), the top entries of which are, for example, the context words *animals, deer, hunter, wild,* and *wolves*. In other words, while the meaning of *kill* in the TASA corpus was strongly biased in favor of *chemicals* and *poisons*, in the context of *hunter, kill* has to do with *wild animals* and *wolves*.

Figure 3

For another example, consider the meaning of *bright* in the context of *light* and in the context of *smart*. The language model for the former has 1143 entries and the language model for the latter has 1016 entries. However, there is hardly any overlap between them. $Bright_{light}$ is related to *light*, $P(light \mid bright_{light})$=.44, but $bright_{smart}$ is not, $P(light \mid bright_{light})$ = 004.

We have described the construction of meaning of a word in a proposition/dependency unit. We now show how the same procedure is applied to sentences. The model has not yet been extended to deal with complex sentences but we present some preliminary efforts at addressing this difficult task. A sentence contains several interacting somewhat independent propositions. The general strategy we propose is to employ the dependency parse of the sentence to break down the sentence into these independent propositions consisting either of a single dependency unit or two dependency units. The word units now have a role in an order-dependent specification of the proposition's meaning. This propositional unit is first initialized as a directed graph

with vertices represented by the word, its contextualized explicit relational trace, and the edges representing transition probabilities from its explicit sequential trace. This essentially amounts to first contextualizing the words using the explicit relational trace and then checking whether they are combined in a syntactically acceptable dependency unit according to the explicit sequential trace. The meaning of the proposition is now specified as a set of such similar chains the initialized representation can derive.

Consider the sentence *The hunter killed the deer*. The process described above first initializes a chain Hunter->Killed->Deer. We would like this over-specified initial representation to capture the intended meaning as something like Agent{hunter, sportsman, hunters} -> killed{shot, kill ,kills, shoot} -> Patient{deer, bear, lion, wolves}. To enable such generalization, we sample word features from $hunter_{killed}$, $killed_{hunter,deer}$ *and* $deer_{killed}$. The acceptance of the feature chain depends upon the product of the probabilities from the explicit sequential trace. We analyze one such operational instance: Let the words (*deer, kill, sportsman*) be sampled from $hunter_{killed}$ and (shot->bear) be one feature sub-chain from sampling $killed_{hunter,deer}$ and $deer_{killed}$ From the candidate set of complete chains (*deer->shot->bear*), (*kill->shot->bear*) and (*sportsman->shot->bear*), *sportsman->shot* is the only unit with a non-zero probability and hence the only accepted chain. Each feature has a weight associated with it that reflects its construction process: the paradigmatic probabilities resulting from how substitutable a candidate word is semantically, as well as the syntagmatic probabilities resulting from how substitutable it is with respect to the sequential context. A similarity measure for comparing two sentences can be obtained by relative sums of the weights for those features that overlap in two sentence representations. This is not as intuitive a similarity measure as a cosine or a conditional probability, but it does indicate when sentences are unrelated and orders related sentences by rank. Some examples are shown in Table 3.

Table 3

We have not yet systematically evaluated the CI-II model for sentences with a representative set of materials, in part because we are not aware of a generally accepted

benchmark to compare it with. There are, however, two such benchmarks that have been used in the literature to evaluate and compare statistical models of semantics: the free association norms of Nelson and the TOEFL test. The University of South Florida Association Norms (Nelson, McEvoy, & Schreiber, 1998) include human ratings for 42,922 cue-target pairs. There are several ways to compare the human data with model predictions. One way is to find out how often the associate ranked highest by a model was in fact the first associate produced by the human subjects. For LSA and Topic, these values were 9% and 14%, respectively. Since LSA uses a symmetric distance measure and associations are well known to be asymmetric, it is not surprising that the Topics model outperforms LSA (see also Figure 8 in Griffiths et al., 2007). The CI-II model performs about as well as the Topic model (14%), but the really interesting result is the substantial improvement obtained by combining the Topic and CI-II predictions (22%). The combination results in a 135% improvement over LSA. The implications of this result are important, for it provides direct support for the dual memory assumption of the CI-II model. The CI-II predictions are based on the explicit relational trace. However, that does not replace the gist trace (here the Topic model); rather, it complements it. To account for human behavior, both are necessary.

Table 4

Table 4 shows another way to compare models with the data from the Nelson norms. It shows the median rank of the first five associates predicted by the different models. The Topic and the CI-II models yield substantially better predictions than LSA; however, the strongest predictor of data is the CI-II – Topic combination, which produces an 82% improvement in prediction for the first ranked response.

Table 5

The conclusion that both gist level and explicit representations are needed to account for the data is further supported by the analyses of the models' performance on the TOEFL test, shown in Table 5. The TOEFL is a synonym recognition task with four

response alternatives (Landauer & Dumais, 1997). Performance on the TOEFL is predicted best when both gist and explicit information are used, providing further support for the dual-memory model. Note that LSA does quite well with the TOEFL test predictions, better than the Topic model on its own and not much different from the CI-II model. However, combining a gist-model (either LSA or Topic) with the explicit CI-II yields the best results.

Table 6

The predictions for both the free association data and the TOEFL test both involve word meanings out of context. To evaluate the predication component of the CI-II model, we use an example from Jones & Mewhort (2007). To show how their BEAGLE model combines order and context information, Jones & Mewhort note that following *Thomas*_____ , the strongest completion is *Jefferson*, but that additional context can override this association, so that *Edison* becomes the most likely word in *Thomas* _____ *made the first phonograph* (their Table 8). In Table 6 we show that the CI-II model behaves in much the same way, not only favoring *Edison* over *Jefferson* in the proper context, but ruling out *Jefferson* and the other alternatives. Each context sentence was parsed to show which words were directly connected to the target word. In the example above, the relevant context would be *made phonograph*. The sequential trace of *Thomas*_____ contains 143 context words, including the six target words shown in Table 6. A topic is sampled from the gist representation of *made phonograph* (which loads on two topics) and features are sampled from the sequential trace under the control of the *made phonograph* topic sampled. Only *Edison* loads on the topics of *made phonograph*, so it is selected with probability 1.

How does the performance of the CI-2 model compare with other models? There are two considerations for such a comparison: whether the top choice of a model is correct, and how strongly a model prefers the correct choice over the next-best alternative. The CI-2 model not only picks the correct alternative in all six contexts, but it also distinguishes quite sharply between the correct choice and the second best, the average difference being .69. For comparison, BEAGLE also picks all the correct

choices, but the average difference between the correct choice and the second best is only .21. Not surprisingly, LSA and Topic, with their neglect of order information, do not do nearly as well. LSA makes three errors and Topic model four.

Table 7

The same procedure was used to obtain predictions for a set of cloze data by Hamberger (1996). Hamberger obtained human completion data for 198 sentence contexts, 12 of which had to be excluded from our analysis because they included words not in our corpus. Predictions were obtained by turning the task into a 186-alternative multiple-choice task. That is, each model produced a rank ordering of the target items. To obtain the rank orderings for all models only the direct dependency unit was used as context. For instance, for the item *I sewed on the button with a needle and _____, sewed* dictates which topics are allowed, and the probability of accepting a feature from the explicit trace reflects how many times it has been seen with *needle*. The feature weights determine a rank ordering of the target words. Table 7 shows the number of times a model correctly predicts the first associate given by the human subjects (*thread*, in our example) as well as the median rank of the first associate for each model. The CI-II model using both relational and sequential traces performs much better than LSA or the Topic model, but as in our earlier analyses, the best performance is achieved when gist and explicit information is combined.

Conclusions

Statistical models of semantics are based on the analysis of linguistic corpora. The goal of the analysis is to find the optimal (or near-optimal) algorithm that could generate these corpora, given the constraints imposed by the human cognitive system. We have argued that semantic models like LSA describe what is stored in long-term memory. Long-term semantic word memory is a decontextualized trace that summarizes all the experiences a person has had with that word. This trace is used to construct meaning in working memory. Meaning is therefore always contextual, generated from the interaction

between long-term memory traces and the momentary context existing in working memory. Importantly, these interactions involve not only semantic traces, but also syntactic constraints.

We have described a number of different ways to generate the representations in semantic memory, relying on different machine learning techniques, such as singular value decomposition for LSA and a Bayesian procedure for the Topic model. As long as these methods use only the information provided by the co-occurrence of words in documents, they all yield qualitatively similar results, with no obvious general superiority of one model over another. No general, systematic comparison between these models has been made, but in our opinion such a giant bake-off would have little value. Different models have different strengths and limitations. Some are more natural to use for certain purposes than others, and one might as well take as much advantage of this situation as one can. In other words, it seems reasonable to use LSA for essay grading, the Topic model for asymmetric similarity judgments, and so on. Indeed, the fact that all these formally different approaches yield similar results is reassuring: we seem to be getting at real semantic facts, not some method-specific artifacts.

When additional information beyond word co-occurrence is considered, such as word order information, major differences between models emerge. The success of the holograph model BEAGLE in accounting for a wide range of psycholinguistic data provides an illustration of the importance of word order information. However, syntactic structure plays an even more important role in language use. The syntagmatic-paradigmatic model of Dennis (2005) and our own approach, the CI-II model, are two examples of how syntactic information can be incorporated into statistical models of semantics.

Meaning in the CI-II model is constructed in working memory from information stored in long-term memory. Different types of information are available for the contextual construction of meaning, from gist-like information as in the Topic model to explicit memory traces. Relational traces record word co-occurrences in documents;

sequential traces record word pairs that occurred together in dependency units in sentences, obtained from parsing the sentences of a corpus with a dependency parser. When this information is used in working memory to construct a sentence meaning, a language model for the sentence is generated in such a way that it contains only contextually relevant and syntactically appropriate information. Thus, as in the original CI model, the construction phase uses all information about word meanings and syntax that is available in long-term memory, whereas the integration phase selects on those aspects that are contextually relevant.

Future work on the CI-II model will focus on extending it to complex sentences and on developing applications to scoring of short-answer questions in the context of a comprehension tutor. But apart from developing and evaluating the present framework, there are comprehension problems that are beyond the scope of the model discussed here. The most obvious one concerns ways to derive syntactic information through unsupervised learning mechanisms. There are already unsupervised learning models for syntax. The ADIOS model of Solan, Horn, Ruppin, & Edelman (2005; see also Edelman, 2008, Chapter 7) uses statistical information to learn regularities, relying on a graph theoretic approach. The syntagmatic-paradigmatic model of Dennis (2005) infers both semantic and syntactic structure in an unsupervised fashion. One can confidently expect rapid progress in this area. The second gap in the development of statistical models of semantics that needs to be addressed is the limitation of current models to the symbolic level, that is, their neglect of the embodiment of meaning. Computers do not have bodies, but there is reason to believe that it might be possible to model embodiment so as to integrate perceptual and action-based representations with the symbolic representations studied here. Promising beginnings in that respect have already been made by various researchers, such as Goldstone, Feng, & Rogosky (2005) and Andrews, Vigliocco, & Vinson (2009).

There is a third problem, however, that appears more difficult to solve with current methods: discourse comprehension requires not only knowledge of what words mean and how they can be combined, as has been discussed here, but world knowledge

beyond the lexical level – knowledge about causal relations, about the physical and social world, which is not captured by our present techniques. Discourse comprehension involves representations at three levels: the surface or verbatim level, the propositional textbase, and the situation model (Kintsch, 1998). We are dealing here with the first two; future models will have to be concerned with the situation model, as well. Elman (2009) has made a forceful argument that, in order to be successful, theories of language comprehension need to explicitly model event schemas and deal with all the issues that were once considered by schema theory. Some of these problems are addressed by the present model. For instance, the model yields a high similarity value of .4095 for the comparison between *The carpenter cuts wood* and *The carpenter saws wood,* and a low similarity value of .0005 for *The carpenter cuts wood* and *The carpenter cuts the cake.* However, many more complex issues remain that are beyond the scope of the present model.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2009) Integrating attributional and distributional information to learn semantic representations. *Psychological Review*, 116, 463-498.

Balota, D. A. (1990). The role of meaning in word recognition. In G. B. d'Arcais, D. A. Balota & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 9-32). Hillsdale, NJ: Erlbaum

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974) Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior*, 13, 471-481.

Barsalou, L. W. (1987). Are there static category representations in long-term memory? *Behavioral and Brain Sciences, 9*, 651-652.

Bates, E., & Goodman, J. C. (2001). On the inseparability of grammar and the lexicon: Evidence from acquisition. In M. Tomasello & E. Bates (Eds.), *Language development.* (pp. 134-162). Oxford: Blackwell.

Brainerd, C. J., Wright, R., & Reyna, V. F. (2002). Dual-retrieval processes in free and associative recall. *Journal of Memory and Language, 46*, 120-152.

Bresnan, J., & Kaplan, R. (1982). Lexical functional grammar: A formal system for grammatical representation. In J. Bresnan (Ed.), *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 117-156). Mahwah, NJ: Erlbaum.

Dennis, S. (2005). A memory-based account of verbal cognition. *Cognitive Science, 29*, 145-193.

Dennis, S., & Kintsch, W. (2008). Text mapping and inference rule generation problems in text comprehension: Evaluating a memory-based account. In F. Schmalhofer & C. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes* (pp. 105-132). Mahwah, N.J.: Erlbaum.

Edelman, S. *Computing the mind*. Oxford: Oxford University Press, 2008

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. . *Cognitive Science*, 33, 547-582.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211-245.

Ferreira, F. & Patson, N. D. (2007) The 'good-enough' approach to language comprehension. *Language and Linguistics Compass*, 1, 71-83.

Garrod, S., Freudenthal, D., & Boyle, E. (1994). The role of different type of anaphor in the on-line resolution of sentences in a discourse*., 33*, 38-68.

Gigerenzer, G. & GoldsteinD. G. (1996) Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103,650-669.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalizations in language*. Oxford: Oxford University Press.

Goldstone, R. L., Feng, Y., & Rogosky, B. (2005). Connecting concepts to the world and each other. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception in action, memory, and thinking.* Cambridge: Cambridge University Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, 5228-5235.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*, 211-244.

Hamberger, M. J., Friedman, D., & Rosen, J. (1996). *Completion norms* collected from younger and older adults for 198 sentence contexts. *Behavior Research Methods, Instruments & Computers*, 28(1), 102-108.

Hintzman, D. L. (1984). MINERVA2: A simulation model of human memory. *Behavior Research Methods, Instruments, and Computers, 16*, 96-101.

Jones, M. N., & Mewhort, D. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1-37.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55*, 534-552.

Joshi, A. K. (2004). Starting with complex primitives pays off: complicate locally, simplify globally. *Cognitive Science, 28*, 647-668.

Kawamoto, A. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language, 32*, 474-516.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street: Computer-guided summary writing. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Snalysis* (pp. 263-278). Mahwh, NJ: Erlbaum.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W. (2001). Predication. *Cognitive Science, 25*, 173-202.

Kintsch, W. (2007). Meaning in context. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *Latent Semantic Analysis*. Mahwah, NJ: Erlbaum. Pp. 89-105.

Kintsch, W. (2008a) Symbol systems and perceptual representations. In M. De Vega, A. Glenberg, & A. Graesser (Eds.) *Symbolic and Embodied Meaning*. NY: Oxford Univ. Press.

Kintsch, W. (2008b). How the mind computes the meaning of metaphor: A simulation based on LSA. In R. Gibbs (Ed.), *Handbook of Metaphor and Thought*. New York: Cambridge University Press. Pp. 129-142.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language, 45*, 259-282.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review, 12*, 703-710.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. *Assessment in Education, 10*, 295-308.

Landauer, T. K., McNamara, D., Dennis, S., & Kintsch, W. (Eds.). (2007). *Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

MacDonald, J. L., & MacWhinney, B. (1995). Time course of anaphor resolution: Effects of implicit verb causality and gender*., 34*, 543-566.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676-703.

Mangalath, P., (in preparation). Learning structured representations in language: A memory-based approach.

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35-56). Mahwah, NJ: Erlbaum.

Mel'cuk, I. (1988). *Dependency syntax: Theory and practice*. New York: State University of New York Press.

Murdock, B. B., Jr. (1982). A theory for the storage and retrieval of items and associative information. *Psychological Review, 89*, 609-626

Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. (1998) *The University of South Florida word association, rhyme, and word fragment norms.* From http//www.usf.edu/Free Association/

Nivre, J., Hall, J., Chanev, J., Eryigit, A., Kuebler, G., Marinov, S., & Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering, 13*, 1-41.

Quesada, J. (2007). Creating your own LSA spaces. In T. K. Landauer, D. S. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 71-88). Mahwah, NJ: Erlbaum.

Rumelhart, D. E. (1979). Some problems with the notion that words have literal meanings. In A. Ortony (Ed.), *Metaphor and thought*. Cambridge, England: Cambridge University Press.

Sgall, P., Hajicova, E., & Panevova, J. (1986). *The meaning of the sentence in its pragmatic aspects*. Amsterdam: Reidel.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge MA: MIT Press.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences, 102*, 11629-11634.

Steedman, M. (1996). *Surface structure and interpretation*. Cambridge, MA: MIT Press.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *Latent Semantic Analysis*. Mahwah, N.J.: Erlbaum. Pp. 427-448.

Steyvers, M., & Griffiths, T. L. (2008). Rational analysis as a link between human memory and information retrieval. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for a Bayesian cognitive science* (pp. 329-350). Oxford: Oxford University Press.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. Proceedings of the 21st International Congress on Computational Linguistics and the 64th Annual Meeting of the Association for Computational Linguistics.

Tesniere, L. (1959). *Elements de syntax structurale*. Paris: Editions Klincksieck.

Tomasello, M. (2001). The item-based nature of children's early syntactic development. In M. Tomasello & E. Bates (Eds.), *Language development.* (pp. 163-186). Oxford: Blackwell.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MS: Harvard University Press.

Yamada, H., & Matsumoto, Y. (2003). *Statistical Dependency Analysis with Support Vector Machines.* Paper presented at the 8th International Workshop on Parsing Technologies.

Table 1. The 20 nearest neighbors of *play* in LSA and the Topic Model.

20 nearest neighbors by LSA cosine	20 nearest neighbors by conditional probability
play	*play*
playing	game
played	*playing*
kickball	*played*
plays	*fun*
games	*games*
game	pat
volleyball	children
fun	ball
golf	role
costumes	*plays*
actor	important
rehearsals	music
actors	mart
drama	run
comedy	friends
baseball	lot
tennis	stage
theater	toys
checkers	team

Table 2. The 20 nearest neighbors of *play* in the Topic Model, as well as the 20 nearest neighbors of PLAY∩SHAKESPEARE and PLAY∩BASEBALL.

P[W|PLAY]	P[W|PLAY ∩SHAKESPEARE]	P[W|PLAY ∩BASEBALL]
play	play	play
game		game
playing		playing
played		played
fun		
games		games
pat		
children		
ball		ball
role		
plays	plays	
important	important	
music	music	
part	part	
run		run
friends		
lot		
stage	stage	
toys		
team		team
	audience	baseball
	theater	hit
	drama	field
	scene	bat
	actor	sports
	actors	players
	Shakespeare	player
	tragedy	balls
	performed	football
	scenes	tennis
	character	throw
	dramatic	sport

Table 3. Similarity values for some sentence pairs

The hunter shot the deer	Similarity
–The deer was killed by the hunter	.160
–The hunter killed the bear	.093
–The hunter was shot by the deer	0
–The deer killed the hunter	0
The soldiers captured the enemy	
-The army defeated the enemy	.26
-The prisoners captured the soldiers	0
The police apprehended the criminal	
–The police arrested the suspect	.17
–The cops arrested the thief	.018
–The criminal arrested the police	0

Table 4. Median ranks of the first five associates produced by human subjects in the ordering produced by four models.

	LSA	Topic	CI-II	CI-II & Topic
First Associate	49	31	20	9
Second Associate	116	107	62	22
Third Associate	185	196	113	49
Fourth Associate	268	327	180	72
Fifth Associate	281	423	229	90

Table 5. Predictions for 79 items on the TOEFL test for five models. Number attempted is the number of items for which all five words were in the vocabulary used by the model

	LSA	Topic	CI-2	CI-2 & Topic	LSA & Topic
No. attempted	60	45	60	60	60
No. correct	32	24	34	39	42

Table 6. Completion probabilities in seven different contexts

	Jefferson	Edison	Aquinas	Paine	Pickney	Malthus
Thomas <---->	0.21	0.02	0.03	0.13	0.06	0.01
Thomas <----> wrote the Declaration of Independence.	0.68	0	0	0.31	0	0
Thomas <----> made the first phonograph.	0	1.00	0	0	0	0
Thomas <----> taught that all civil authority comes from God.	0	0	0.66	0	0	0
Thomas <----> is the author of Common Sense.	0	0	0	0.12	0	0
A treaty was drawn up by the American diplomat Thomas <--->.	0	0	0	0	0.97	0
Thomas <----> wrote that the human population increases faster than the food supply.	0	0	0	0	0	1

Table 7. Performance of six models on 186 cloze test items from Hamberger (1996)

	LSA	Topic	CI-II relational	CI-II sequential	CI-II rel. & seq.	CI-II & Topic
Number First Associate	43	46	55	68	80	89
Median Rank First Associate	7	6	5	3	2	2

Figure 1. The dependency tree for the sentence "*The furious river flooded the town*" and its propositional structure (indicated by the shaded boxes).

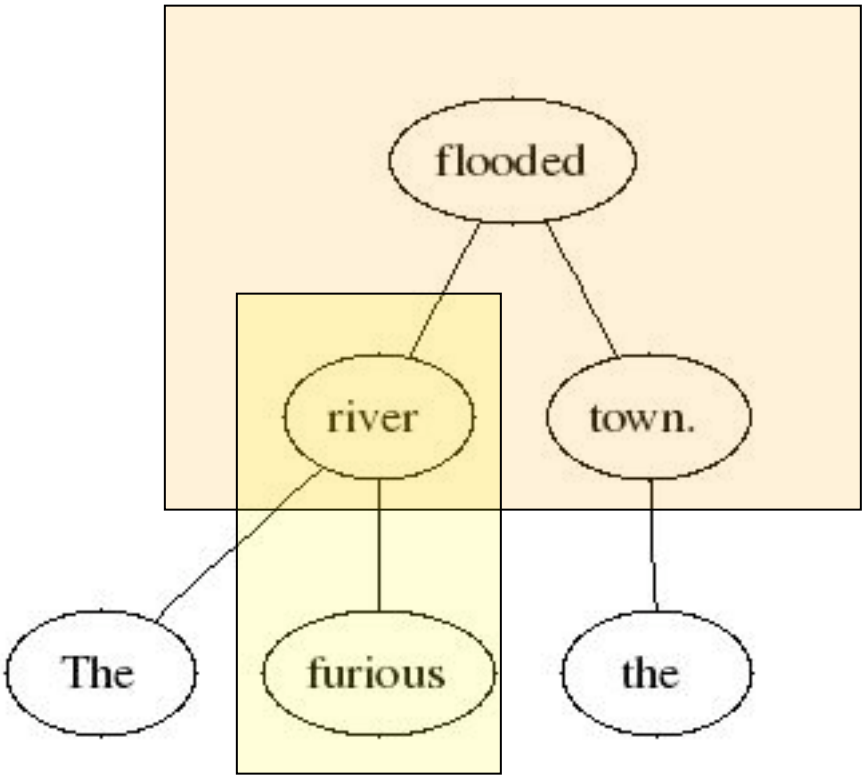

Figure 2. The meaning of *kill* in isolation; numbers indicate number of non-zero entries in a vector

[….24….] 1195 topics for *kill* gist trace
[…….11,594……] 58k contexts for *kill* explicit relational trace

[……1,711…..] 58k contexts for *kill* feature sample
insects
chemicals
poison
hunt
pest
….

Figure 3. The meaning of *kill* in the context of *hunter*;
numbers indicate number of non-zero entries in a vector

[….8 ….] 1195 topics for *hunter* gist trace
[…….11,594……] 58k contexts for *kill* explicit relational trace

[……1,366…..] 58k contexts for *kill(hunter)* feature sample
animals
deer
hunter
wild
wolves
….

[*] The support of the J. S. McDonnell Foundation for this research is gratefully acknowledged. We thank Art Graesser, Danielle McNamara, Mark Steyvers and an anonymous reviewer for their helpful comments.

[i] We use the term 'generative' in a broad sense to refer to a dynamically reconfigurable lexicon where the meaning of a word is generated in working memory based on the context using various heuristics. The term 'generative lexicon' bears no reference or relation to the standard definition in Bayesian machinery.

[ii] When people construct meaning, they consider more than just textual information, but, as we shall see, the semantic representations obtained solely from written texts can be remarkably good approximations to human word meanings. This issue is discussed in more detail in Kintsch (2008a).

[iii] To generate the well-established and distinct meanings of *banks* a simpler algorithm would have sufficed. For instance, the centroids of *banks-money* and *banks-river* are close to the $banks_{money}$ and $banks_{river}$ vectors. However, the enhanced context sensitivity of the predication algorithm is crucial for dealing with the more usual fluid word senses and metaphors. The centroid of *lawyer-shark* makes no sense – its closest neighbors are *sharks, Porgy, whale, bass, swordfish* and *lobsters* – nowhere near the intended meaning of the metaphor.

[iv] The "good-enough" approach is an example of Simon's concept of satisficing (Simon, 1969); other areas in which this approach has been adopted are perception (e.g., Edelman, 2008) and decision making (e.g., Gigerenzer & Goldstein, 1996).

[v] Dependency units are used rather than n-grams because they pick up long-distance relations among the words in a sentence and avoid accidental ones.

[vi] Zero here means almost-zero; because of the smoothing we have used, none of the probabilities are strictly zero.