Generating Example Contexts to Help Children Learn Word Meaning[1]

Liu Liu (liuliu@google.com)
Google Pittsburgh
6425 Penn Ave. Suite 700. Pittsburgh, PA 15206, USA

Jack Mostow (mostow@cs.cmu.edu)
Project LISTEN, School of Computer Science, Carnegie Mellon University
RI-NSH 4103, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Gregory S. Aist (gregory.aist@alumni.cmu.edu)
Applied Linguistics and Communication Studies, Iowa State University
206 Ross Hall, Ames, IA 50011, USA

Abstract

This article addresses the problem of generating good example contexts to help children learn vocabulary. We describe VEGEMATIC, a system that constructs such contexts by concatenating overlapping five-grams from Google's N-gram corpus. We propose and operationalize a set of constraints to identify good contexts. VEGEMATIC uses these constraints to filter, cluster, score, and select example contexts. An evaluation experiment compared the resulting contexts against human authored example contexts (e.g., from children's dictionaries and children's stories). Based on rating by an expert blind to source, their average quality was comparable to story sentences, though not as good as dictionary examples. A second experiment measured the percentage of generated contexts rated by lay judges as acceptable, and how long it took to rate them. They accepted only 28% of the examples, but averaged only 27 seconds to find the first acceptable example for each target word. This result suggests that hand-vetting VEGEMATIC's output may supply example contexts faster than creating them by hand.

1. Introduction

Vocabulary plays a critical role in reading comprehension. "A reader who can pronounce a word but does not know its meaning or crucial facts about it is at a disadvantage in comprehending the text in which it occurs" (Stanovich *et al.* 1991). As the National Reading Panel (NRP 2000) concluded: "Vocabulary is one of the most important areas within comprehension and should not be neglected."

This article focuses on one particular aspect of vocabulary learning – learning word meaning from example contexts. Word meaning includes both *denotation* (objective meaning) and *connotation* (implied meaning and associations) (Chandler 2004). For full mastery of a word, children need to learn both the core meaning and the appropriate use of a word. Since readers must acquire both aspects, effective vocabulary instruction combines explicit explanation with multiple examples of vocabulary in varied contexts (Bolger *et al.* 2008).

Contexts give clues to semantics but also convey many other lexical properties, such as part of speech, morphology, and pragmatics, which help enrich a child's word knowledge base. From the fields of psychology and the learning sciences, it is well established that children can learn word meaning incidentally from contexts (Jenkins 1984; Nagy *et al.* 1985; Schatz 1986; Herman *et al.* 1987; Nagy *et al.* 1987; Schwanenflugel *et al.* 1997; Kuhn and Stahl 1998; Fukkink *et al.* 2001). However, studies also show that incidental vocabulary learning from context is limited and inefficient, and not all contexts are equally appropriate and effective (Beck *et al.* 1983).

[1] This work, performed while the first author was a master's student in the Language Technologies Institute at Carnegie Mellon University, was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080157 to Carnegie Mellon University. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education. We thank Dr. Margaret McKeown for her expertise and assistance, and our lay judges for their participation.

A particular context might not carry enough information to determine the meaning of the word, containing some information but omitting key facets. In fact, most natural contexts are insufficient to infer word meaning (Beck *et al.* 2002), especially for younger readers.

Accordingly, one key issue in vocabulary instruction is identifying how to find or create good example contexts to help learn new words, besides the text where children initially encounter them. Example contexts are usually created by teachers, parents, authors, lexicographers, or, occasionally, educational researchers (McKeown 1985; Bolger *et al.* 2008). A human expert may generate excellent examples, but this process takes time, costs money, and may not be available when needed. Also, human-generated contexts are shaped by the cognitive retrieval and production processes of a person who already knows the word, and might be influenced by when and where the person generates the examples. In contrast, computer-generated contexts can provide systematic and comprehensive coverage.

A standard method to identify good contexts is to retrieve them from an existing corpus. In contrast, we adopt a generation-based approach. Once we've described it, we can explain in detail (in Section 7, Relation to Prior Work) our motivation for adopting it instead of traditional retrieval methods. However, three features of a generation-based approach are worth mentioning here at the beginning of the article. The first is that some of the advantages of a retrieval-based approach, such as providing the word with its common collocates, apply as well to a generation-based approach. The second is that some of the limitations of a generation-based approach, such as the need for hand vetting, also apply to retrieval approaches applied to large corpora: a very large corpus based on general English usage is likely to contain sentences that for one reason or another are not well suited for use in teaching vocabulary to children. (A corpus developed specifically for children would have already had its hand-vetting applied in advance, at corpus construction time, so the work is moved around in time but not avoided entirely.) The third feature is that in various ways a generation-based approach might improve on a selection-based approach: for example, generated examples might provide examples that are better for a specific purpose than those that are available in the corpus; a generation-based approach might also be more efficient since it can test example contexts for suitability as it generates them, rather than retrieving all examples and then filtering them for suitability.

This article describes a system called VEGEMATIC[2] (for "Vocabulary Examples GEnerated autoMATICally"), which generates example contexts to help children learn vocabulary. We focus on "Tier 2" words (Beck *et al.* 2002) – high frequency words used by mature language users across several content areas. These words are unknown to most young children, but important to know. While the method is not intrinsically restricted to Tier 2 words, the article focuses on those words since they are important to teach.

The article also focuses on teaching the core meaning of a word rather than teaching fine-grained distinctions, in line with the reading research literature. For example, *declare* can be used in the sense *state for the record* as in *declare your age*, or *issue a decree* as in *declare an emergency*, or *make a geopolitical claim* as in *declare victory*. The core sense that encompasses all of these is *formally state*, which not only covers these three fine-grained senses but also such uses as *declare my candidacy*, *declared persona non grata*, *declare under penalty of perjury*, and so forth.

The rest of this article is organized as follows. Section 2 describes how VEGEMATIC generates candidate contexts. Section 3 describes constraints on good contexts, and heuristics to operationalize these constraints. Section 4 reports the pilot evaluation of an early version of VEGEMATIC. Section 5 describes extensions that the pilot evaluation motivated. Section 6 evaluates the resulting version on an unseen test set. Section 7 relates VEGEMATIC to previous work and discusses why VEGEMATIC

[2] VEGEMATIC is named after the Ronco Veg-O-Matic kitchen appliance made famous by television ads and parodies. Like its namesake, VEGEMATIC is based on slicing and dicing – not of food but of text, done by Google when it segmented a trillion words of web text into N-grams.

takes a generation approach rather than a traditional selection approach. Section 8 concludes. The Appendix lists examples of VEGEMATIC's output.

2. Context Generation

We now describe how VEGEMATIC concatenates overlapping n-grams to generate contexts.

2.1. Data set

We use Google N-grams (Brants and Franz 2006) to generate candidate contexts. Google N-grams consist of over one billion n-grams and their counts in over one trillion words of text extracted from public web pages, tokenized into lexemes, and segmented into sentences. Google N-grams lists only n-grams with counts of 40 or more. It has been used for many tasks, including spelling correction (Carlson and Fette 2007), machine translation (Hermjakob *et al.* 2008), and other applications (Yu *et al.* 2007; Durme *et al.* 2008).

Google N-grams contain n-grams for *n* from one to five. We use five-grams to generate contexts, as five-grams are the longest n-grams in Google N-grams and hence provide more information about target words. There are 1,176,470,663 five-grams, e.g.:

advantage in a competitive environment	66
advantage in a competitive job	69
advantage in a competitive market	219
advantage in a competitive world	94

Here the number listed after each five-gram is its count.

2.2. Generation methods

VEGEMATIC performs heuristic search in a space of candidate word sequences, starting with a set of five-grams that contain the target word, e.g. *been **extinct**[3] for millions of.* It repeatedly selects a candidate and extends it one word to the left or right by choosing a five-gram (underlined below) that matches the first or last four words, e.g.:

*been **extinct** for millions of*
*been **extinct** for millions of years*
*have been **extinct** for millions of years*
*Dinosaurs have been **extinct** for millions of years*

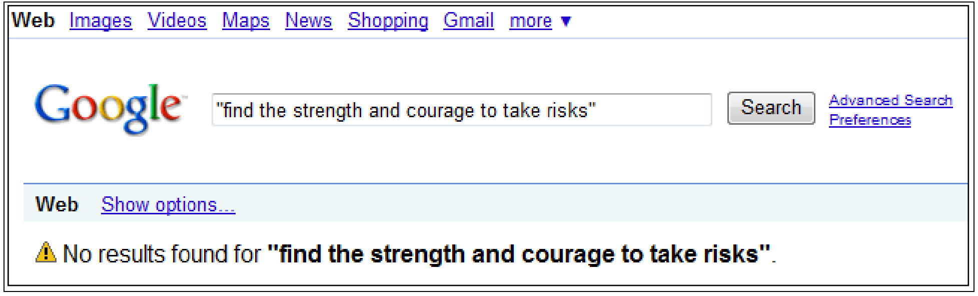
VEGEMATIC incorporates filters to prune candidates that violate constraints, as Section 3 will describe. Using only five-grams containing the target word, it generates contexts at most 9 words long with the target word in the middle and four words on each side of the target word.

This generation method is similar to that used in a context-based machine translation system by Carbonell, Klein, Miller, Steinbaum, Grassiany and Frey (2006). It finds target n-gram translation candidates that overlap maximally with translation candidates for the previous and following n-grams. We exploit the same underlying intuition, namely that generated contexts are locally coherent because the n-grams are long, come from human-written text, and overlap substantially.

In particular, we assume that if one five-gram overlaps with another by four words, then they come from the same set of sentences in the original corpus. When this heuristic assumption holds true, the method reconstructs part or all of one of these sentences. However, when this assumption does not hold true, the method can generate a novel word sequence. We call this phenomenon "crossover" because it combines five-grams from different sentences. The resulting sequence is still locally coherent because each successive five words constitute an authentic five-gram. The longer a generated context, the higher the risk of crossover.

[3] Throughout this article, examples are *italicized* with the target word in ***bold***.

On the positive side, this ability to generate novel contexts can potentially improve on the original contexts, e.g., by streamlining them to eliminate undesirable complexity. For example, the generated context *find the strength and* ***courage*** *to take risks* is novel, as Figure 1 shows. On the negative side, crossover can produce global inconsistencies. For example, *Dollars can make a* ***tremendous*** *amount of work done* is incoherent, although the five-grams are locally consistent.



"find the strength and courage to take risks"

⚠ No results found for **"find the strength and courage to take risks"**.

Figure 1: Screenshot of search results for "*find the strength and* ***courage*** *to take risks*" (7/22/2009)

3. Context Constraints

Not all contexts constructed by concatenating overlapping Google N-grams are good examples. We identified several constraints that characterize good contexts, based partly on expert knowledge and partly on observed deficiencies of some generated contexts.

This section describes each constraint and how we operationalize it as one or more heuristic filters. A filter either eliminates candidate contexts that violate a constraint, or prefers contexts that satisfy the constraint to a greater degree or with higher probability.

3.1. Comprehensible to children

We want contexts that help children learn vocabulary. For a context to help, the child must understand it. If a context contains too many unfamiliar words besides the target word, the child will not understand it well enough to help in learning the target word. For example, the context *It is time to* ***declare*** *victory and go home* is reasonably understandable, assuming the child knows the word *victory*. In contrast, any context containing *...penalties of perjury solemnly* ***declare****...* is useless for teaching ***declare*** to a child who does not know the words *penalties*, *perjury*, or *solemnly*.

One comprehensibility filter eliminates candidates containing more than two hard words. We operationalize "hard" as words that Biemiller (2009) classifies above grade 2 (typically ages 6-7). For words that Biemiller's list omits, we consult a similar list (Paynter *et al.* 2005).

Syntactic complexity is another impediment to understanding a context. Therefore a second filter eliminates candidates that contain relative pronouns, such as *who* or *that*.

3.2. Grammatically correct and complete

Good contexts should be complete, grammatical sentences. Some generated candidates are not grammatical, such as *Southpaw* ***Stout*** *Dem Blog The Scarlet*. (Google N-grams come from text that includes many lists.) Some candidates are incomplete fragments of grammatical sentences, e.g., *Jennifer is very* ***anxious*** *to know about the*.

To filter out ungrammatical contexts, we use the Link Grammar Parser (Sleator and Temperley 1993), a syntactic dependency parser, as a grammar checker. VEGEMATIC eliminates candidates the parser cannot parse as syntactically valid English. While use of an existing parser is not a perfect solution to the challenge of ungrammatical contexts, it does eliminate many of the worst, such as *Jennifer is very anxious to know about the*, which is clearly incomplete and thus not a grammatical sentence.

To filter out incomplete contexts, a second filter requires candidates to start and end with Google N-grams' sentence start and end symbols <S> and </S>.[4] Another filter eliminates candidates that end with modal or auxiliary verbs.

3.3. Word-sense appropriate

Sometimes a target word has multiple senses, only one of which is the target meaning to be taught. A good context is consistent with the target meaning. Contexts that use some other sense of the target word are unhelpful and even confusing. To eliminate them, VEGEMATIC requires contexts to use the target word with a part of speech (POS) compatible with the target meaning.

A filter uses the Stanford Part-of-speech tagger (Toutanova and Manning 2000; Toutanova *et al.* 2003) to determine the target word's POS in a candidate context. If it does not match the POS of the target meaning, the filter eliminates the candidate. For example, if the target meaning of ***force*** is a "group of people," a noun, this filter eliminates the context *do not try to* ***force*** *her to marry a man*, where ***force*** is a verb.

A stronger filter would also exclude contexts with the right POS but a wrong sense of the target word ***force***, such as "strength." Deciding whether the word sense used in a short context matches the target meaning is a tough challenge for the already difficult problem of word sense disambiguation. Instead, Mostow and Duan (2011) used the target meaning to constrain VEGEMATIC's generation process in the first place. However, that work is outside the scope of this article.

3.4. Informative about word meaning

A highly informative context imposes strong semantic constraints on the target word. Such constraints play a substantial role in learning word meanings (Bolger *et al.* 2008).

VEGEMATIC operationalizes semantic constraints as multiple filters. One filter prefers longer contexts as likely to provide more information. A second filter prefers content words over function words (specified by a list) because content words tend to provide more information about target word meaning. This filter eliminates contexts that contain fewer than three content words, including the target word. A third filter prefers words related to the target word, and requires initial five-grams to contain one or more of the 100 most closely related content words. We measure relatedness of two words based on their occurring more frequently than by chance:

Here is the frequency of word in text, and is the probability that words and co-occur in a 5-word window, estimated using Google N-grams. This measure is similar to Pointwise Mutual Information (Church and Hanks 1990).

Overall, these filters prefer contexts with more words overall, more content words, and more related words. For example, consider these two contexts:

Find the strength and ***courage*** *to take risks*

We know it takes ***courage*** *to do so*

Both contexts are 8 words long, but the filters prefer the first context because it contains more content words, including *strength* and *take*, which co-occur with ***courage*** more often than by chance.

3.5. Ordinary prose

Good contexts use normal, classroom-appropriate English. However, we noticed that some of the generated contexts were very web-specific, and likely unfamiliar to young children, e.g., *a* ***Merchant*** *ID and password*.

A lexical filter to avoid web jargon prohibits words much more common on the web than in print, such as *copyright*, *password*, and *download*. We construct a list of web-

[4] To generate more example contexts, the pilot version accepted contexts that started with <*S*> or a capitalized word, or ended with </S> or end-of-sentence punctuation.

specific words by comparing word counts in Google N-grams against the ANC text corpus (http://americannationalcorpus.org/), a broad sample of ordinary English. Of the 1,000 most frequent words in Google N-grams, we classify as web-specific those whose unigram counts in Google N-grams are 20,000 higher than in ANC. (Here 1,000 and 20,000 are heuristic values based on trial and error.) Similar filters prohibit 53 taboo words and phrases, and special symbols such as @.

The web contains many word lists, which lack syntactic structure. Accordingly, a filter eliminates candidates containing more than four consecutive numerals or capitalized words, e.g., *Break Southpaw **Stout** Dem Blog*. Other filters eliminate contexts containing capitalized words other than the first word, the target word, or named entities identified by the Stanford named entity recognizer (Finkel *et al.* 2005).

3.6. Typical of usage and situation

Typicality is an important property of contexts. Helpful contexts should show how words are commonly used, and in what situations. For example, ***celebrate*** is often used in situations like birthdays and anniversaries. A context with the words *My parents are throwing a party to **celebrate** my birthday* is more typical than *He is **celebrating** a pencil.*

We use five-gram counts to operationalize typicality. One filter prefers high-count five-grams in constructing contexts. Another filter prefers candidates with high typicality scores. We score the typicality of a context by averaging the counts of its five-grams.

3.7. Varied and not redundant

Children need to see a word in several varied contexts in order to build up a representation of the word's meaning and acquire enough retrieval cues to access it reliably and efficiently (Bolger *et al.* 2008). Google N-grams are numerous enough for VEGEMATIC to generate diverse contexts for a target word, for example:

- *Members are asked to **declare** that you are 18*
- *He was forced to **declare** a state of emergency*
- *It is time to **declare** victory and go home*

However, some generated contexts are very similar, e.g.:

- *Just **declare** victory and go home*
- *We should **declare** victory and go home*
- *It 's[5] time to **declare** victory and go home*

To eliminate redundancy, we developed the following algorithm:

(1) Partition candidate contexts into clusters that share the same three content words.

(2) Locally, score contexts in the same cluster. To avoid overly similar contexts, pick the highest scored context of each cluster as its representative.

(3) Globally, score and rank the representatives. Output the top *k*, where *k* is the number of contexts needed for vocabulary teaching. If there are fewer than *k*, output them all.

The scoring function used in steps 2 and 3 combines four context features: typicality, length, number of content words, and number of target-related words. To normalize each feature, we divide it by its average value over the set being scored. In step 2, this set is a single cluster; in step 3, it consists of the representatives. We score each context as the mean of its normalized features. A score of 1 indicates average quality. Since feature values are normalized relative to a particular set, a context score depends on the set relative to which it is scored; thus a context chosen as a representative may have different scores in step 2 and step 3.

3.8. Summary and order of filters applied

Table 1 lists in order of application the filters that operationalize the constraints discussed above, with an example and the percentage of remaining candidates pruned by each filter.

[5] Spaces before punctuation are an artifact of tokenization in creating Google N-grams.

For efficiency, VEGEMATIC applies filters as early as possible in its heuristic search (Mostow 1983). The earliest filters apply to initial five-grams, their extensions, and the five-grams used to extend them. Later filters apply to candidate complete contexts. The final filters apply to sets of similar complete contexts in order to choose among them.

Table 1: Context constraints in order applied in VEGEMATIC

Constraint	Filter out	Examples filtered out	% pruned
Filters on five-grams and their extensions			
Informative about word meaning	start five-grams w/o related words	*Grant **Stout** and Eric Coleman*	44.80
Ordinary prose	taboo words	*[unprintable]*	0.10
Ordinary prose	special punctuations or special symbols	*> Have the **courage** to face life cheerfully*	37.00
Ordinary prose	web-specific words	*eBay Buyer Protection **Merchant** info*	1.20
Comprehensible to children	relative pronouns	***declare** the causes which impel*	3.90
Ordinary prose	more than 4 capitalized words or numbers	*Appliances **Merchant** Rating 13 User*	23.50
Comprehensible to children	more than 2 words above Grade level 2	*the authority to **declare** a law unconstitutional*	48.90
Filters on candidate complete contexts			
Complete and grammatically correct	Fails to start with <S> or capitalized word and end with </S> or punctuation	*Months*[6] *and we are **anxious** to hear from you*	99.70
Complete and grammatically correct	end with modal or auxiliary verbs	*I hereby **declare** that the above is*	9.40
Informative about word meaning	fewer than 3 content words	*I was so **anxious** to get back to*	74.30
Complete and grammatically correct	fail grammar checker	*Jennifer is very **anxious** to know about the*	59.40
Word-sense appropriate	POS inconsistent with target meaning	***Stout** also bought books by*	6.40
Ordinary prose	capitalized words except named entities	*I don't feel **anxious** about the R word*	4.10
Filters on sets of generated contexts			
Varied and not redundant	redundant contexts in one cluster	*Report any crime or **suspicious** activity in the area* [vs. *Students should report any **suspicious** activity in the area*]	82.40
Varied and not redundant	contexts with low rank	*Review this **merchant** more from House*	94.40

[6] Google 5-grams are case-sensitive, so words in examples are capitalized or not depending on what 5-grams they came from.

4. Pilot Study

How good are the generated contexts? We did a pilot evaluation of VEGEMATIC as we developed it. Section 4.1 explains how we evaluated contexts; Section 4.2 reports results.

4.1. Methodology

To evaluate the pilot version of VEGEMATIC, we compared the contexts it generated for 10 target words against human-authored contexts. To choose the target words, we started with the 789 words that our vocabulary expert had classified as the "Tier 2" words (Beck *et al.* 2002) in the stories used in Project LISTEN's Reading Tutor (Mostow and Aist 1999). Tier 2 words are used in many situations but unknown to most children, and thus important to teach. Of 15 such words that occurred in two stories, once in each story, we excluded 5 words with multiple parts of speech, and chose the other 10: ***anxious***, ***courage***, ***declare***, ***extinct***, ***merchant***, ***remarkable***, ***slender***, ***stout***[7], ***suspicious***, and ***tremendous***.

To lighten the expert's burden, we used only the 6 highest scored contexts output by VEGEMATIC for each target word (or all of them if fewer than 6). We compared against two sources of human-authored contexts. As a gold standard, we used all example sentences from the WordSmyth children's dictionary (www.wordsmyth.net), which are specifically crafted to illustrate the meaning of each word sense listed. As a sample of naturalistic contexts in which a child would encounter the word during normal reading, we used the Reading Tutor story sentences containing the word. The pilot evaluation set totaled 98 contexts: 57 generated contexts, 21 dictionary sentences (1 to 3 per word, depending on the number of senses listed), and 20 story sentences.

Our vocabulary expert, blind to source, rated all 98 contexts on a five-point Likert scale (1=poor, 3=OK, and 5=good), both on general quality, and on three specific aspects that she proposed in order to capture finer-grained properties of generated contexts. The three aspects were: (1) good use of words, i.e., correct or meaningful use in the intended target sense; (2) the degree to which the context is constraining, or reveals elements of the word meaning; (3) expected comprehensibility to children based on other words or concepts in the context and on its syntactic complexity.

4.2. Results and discussions

As Table 2 illustrates, all three sources of contexts included some that the expert rated as 5. But how did they compare more broadly? Table 3 shows mean ratings and standard errors for each source of contexts. ANOVA revealed a main effect of source for the general score ($F = 10.9$, $p < 0.001$), good use of words ($F = 4.4$, $p < 0.05$), constraining context ($F = 3.7$, $p < 0.05$), and comprehensibility ($F = 7.1$, $p < 0.01$). Pairwise comparisons showed that dictionary contexts surpassed VEGEMATIC examples in general score ($p < 0.001$), in good use of words ($p < 0.05$), in constraining context ($p < 0.05$), and in comprehensibility to children ($p < 0.05$). There was also a trend for the story sentences to be better than the VEGEMATIC contexts in general score ($p=0.051$). No other differences were statistically significant.

Table 2: Examples of contexts rated 5 by expert for target words ***courage*** and ***extinct***

Context	Source
*Find the strength and **courage** to take risks*	VEGEMATIC
*So far, you're the only person who has had the **courage** to step over it.*	Story
*It takes **courage** to stand up for what you believe in.*	Dictionary
*Dinosaurs have been **extinct** for millions of years*	VEGEMATIC
*An **extinct** volcano will not erupt again.*	Dictionary
*He had thought the blue butterfly was **extinct**.*	Story

[7] We inadvertently overlooked its noun sense as a type of beer, as we realized from later examples.

Table 3: Expert ratings of contexts from pilot version of VEGEMATIC and other sources

Evaluation criteria	VEGEMATIC (all)	VEGEMATIC (top half)	Story	Dictionary
General score	2.5 (0.21)	3.9 (0.15)	3.4 (0.28)	4.1 (0.21)
Good use	3.4 (0.22)	4.2 (0.21)	4.0 (0.28)	4.4 (0.20)
Constraining context	3.2 (0.21)	3.9 (0.19)	3.6 (0.21)	4.1 (0.19)
Comprehensibility	2.7 (0.23)	4.2 (0.18)	3.5 (0.33)	4.1 (0.25)

When two methods differ significantly in average performance, one method might be *uniformly* worse than the other – or it might *usually* be worse, with exceptions where it performs reasonably well. If in the latter case the automatically generated contexts contain enough good examples, then a human rater or some automated process could potentially select such examples in a subsequent stage. To determine whether this was the case, we examined the top half of the generated contexts (as rated by our expert.) They compared favorably to story sentences, which suggested that refining VEGEMATIC to filter out the lowest half could make its output as good as story sentences. However, such predictions are overly optimistic because we "tested on the training data," in that we designed some of the filters specifically to eliminate bad sorts of contexts generated for the 10 test words. The pilot evaluation served to guide development to improve VEGEMATIC by fixing observed deficiencies. In contrast, Section 6 will report evaluation on an unseen test set, so as to predict future performance of the improved system. But first we describe the improvements it incorporates.

5. Improvements to VEGEMATIC

The pilot test revealed problems with various aspects of automatically generated contexts.

- **Typicality:** Some low-rated contexts had spuriously high frequency, e.g., *Please check the **merchant** store*. Such contexts were composed of five-grams with high counts; VEGEMATIC chose them because it averaged five-gram counts to quantify typicality of usage. However, as the pilot study revealed, the high counts were not due to actual common usage in English but rather to replication of documents on the Web. Five-grams in particular seem surprisingly vulnerable to such spuriously high counts, as Section 5.1 will discuss.
- **Completeness:** Some low-rated contexts were incomplete or ungrammatical.
- **Balance:** The pilot version scored candidates by averaging their normalized scores on multiple criteria (typicality, context length, number of content words, and number of target-related words), as Section 3.7 described. This formula implicitly assumed that a higher score on one feature compensates for a lower score on another. However, reflecting on the examples produced by the pilot version made clear that context quality depends on satisfying all the criteria, not just some of them.
- **Diversity:** The clustering method still output many redundant contexts.

To generate typical, complete, well-rounded, diverse contexts, we improved how VEGEMATIC generates candidates, scores them, and promotes diversity.

5.1. Generate more typical contexts

To make example contexts more typical, VEGEMATIC no longer starts by picking high-frequency five-grams that contain the target word. Instead, it first chooses the most frequent trigrams containing the target word. Then, for each selected trigram, it chooses the most frequent five-gram that contains the selected trigram. Trigram frequency reflects typicality more reliably than five-gram frequency because trigrams aggregate information across more sentences. Not only are their counts better estimated, they are less distorted by widely replicated documents that render some five-gram frequencies spuriously high.

5.2. Generate complete contexts

To generate complete contexts, VEGEMATIC now uses Google *start* five-grams (which start with <S>) and *end* five-grams (which end with </S>) in two ways – as filters to check whether a generated candidate is complete, and if not, as extensions to complete it. In the latter case, VEGEMATIC extends the candidate with a start and/or end five-gram that overlaps it by 3 words, so as to make it complete but keep it coherent.

5.3. Score multiple criteria

To balance multiple criteria more appropriately, VEGEMATIC first ranks each context by each criterion, and then sorts the ranks worst-first. For example, suppose a context ranks 6th on typicality, 10th on length, 4th on number of content words, and 6th on number of related words. Sorting these ranks worst-first yields the rank vector [10 6 6 4]. VEGEMATIC orders contexts lexicographically by rank vector, e.g., it prefers [10 6 6 4] to [10 7 1 1]. VEGEMATIC uses this order locally to select the best representative of each cluster of similar contexts, and globally to select the best context overall. That is, it prefers better-ranked features but prioritizes the worst-ranked features. In other words, a context is only as good as its worst feature. For example, even the best possible typicality score won't rescue a context with no content words. (Using z-scores instead of ranks might produce a similar order, but our measures do not have normal distributions.)

5.4. Select diverse contexts

To maximize diversity, VEGEMATIC uses novelty detection globally to control the order in which it outputs representative contexts after the first one. In selecting which context to output next, it greedily picks the one least similar to any of the contexts already output. It measures the similarity of an old context to a proposed new context as the number of content words they share, divided by the total number of content words in the new context.

This process is related to the novelty detection task in TREC (http://trec.nist.gov): find sentences both relevant and novel from an ordered sentence list, where a sentence is considered similar to or redundant with previously chosen sentences if its proportion of overlap with them exceeds some threshold. We use such a metric too, but to order contexts rather than to classify them as novel or redundant. Our method resembles Maximal Marginal Relevance (Carbonell and Goldstein 1998) in that it greedily selects the most novel context.

6. Evaluation on Unseen Test Data

How good are the contexts generated after the improvements described in Section 5? Sections 6.1 and 6.2 describe their evaluation by expert and lay judges, respectively. Section 6.1 focuses on measuring quality of generated contexts by comparing them with existing context materials. Section 6.2 focuses on evaluating the productivity of VEGEMATIC by measuring context generation coverage and precision.

6.1. Expert rating of context quality

To evaluate the improved version of VEGEMATIC, we selected 10 previously unseen target words at random from the 789 Tier 2 words in Reading Tutor stories: ***advice, budge, concerned, gained, imagine, intense, obliged, pierced, released,*** and ***vanished***. (Note that some target words are root words while others are inflected.) We evaluated the first 10 contexts that VEGEMATIC output for each target word.

For comparison we chose four sources of human-authored contexts. As in the pilot study, we used examples from the WordSmyth dictionary as a gold standard, and sentences from children's stories as a sample of naturalistic contexts. We added examples that people posted in WikiAnswer (http://wiki.answers.com/Q/FAQ/5070) in response to Example Sentences Questions of the form "What is a sentence using word 'target_word'?" Finally, we also queried each target word in Google and selected the first English sentence containing it in the top 10 retrieved results. To better sample ordinary English, we excluded definitions from online dictionary websites such as the Merriam-Webster Online Dictionary. The resulting test set comprised 192 contexts: 100 generated contexts, 32 dictionary sentences, 36 story sentences, 14 sentences from WikiAnswer, and 10 contexts from Google.

Our vocabulary expert, blind to source, started to rate all 192 contexts the same way as described in Section 4.1, but part-way through, asked us to group all the contexts for each target word together for efficiency, commenting that the contexts were "so repetitive—small variations on similar themes or what looks to be different parts of the same context." Therefore we added the novelty detection heuristic described in Section 5.4, used the modified VEGEMATIC to output a new set of 100 contexts, and pooled them with the human-authored contexts into an updated set of 192 contexts (randomized within word) to rerate from scratch. The quality criteria applied only to individual contexts, not the diversity of the contexts for each target word, so our "peek at unseen data" did not invalidate our results even though it led us to modify VEGEMATIC.

Table 4 shows the mean ratings and standard errors for each source of contexts. ANOVA showed significant effects of source on general score (F = 9.3, p < 0.001), constraining context (F = 4.3, p < 0.01), and comprehensibility (F = 9.8, p < 0.001), but not on good word use. Dictionary examples significantly beat VEGEMATIC examples on general score, constraining context, and comprehensibility. VEGEMATIC examples, story sentences, and Wiki Answers did not differ significantly on general score, but (along with dictionary examples) significantly outscored Google contexts.

Table 4: Expert ratings of contexts from various sources for 10 target words

Evaluation criteria	**Mean (Standard Error)**				
	VEGEMATIC	**Dictionary**	**Story**	**WikiAnswer**	**Google**
General score	3.2 (0.11)	4.2 (0.14)	3.4 (0.20)	3.8 (0.36)	2.1 (0.23)
Good use	4.3 (0.11)	4.5 (0.15)	4.4 (0.14)	4.4 (0.33)	4.4 (0.34)
Constraining context	3.5 (0.11)	4.3 (0.13)	3.8 (0.18)	4.2 (0.32)	3.9 (0.31)
Comprehensibility	3.4 (0.13)	4.6 (0.14)	3.8 (0.21)	4.3 (0.33)	2.5 (0.40)

We wanted to see if the improvements helped. Ideally we would compare the pilot and improved version on the same test set. However, we did not consider it worth our busy expert's time to rate contexts we expected would be inferior. Instead, we compared the improved version's output on unseen test data against the pilot version's already-rated output on the words used to guide its development. Such a comparison is biased in favor of the pilot version, but informative if the improved version turns out to do better anyway.

We needed to control for differences between the two sets of words and for possible changes in expert ratings over time. Therefore we normalized ratings of VEGEMATIC output relative to the expert's contemporaneous ratings of dictionary examples for the same word. More precisely, we computed the *quality ratio* for a word as the mean rating of its VEGEMATIC output divided by the mean rating of its dictionary examples. Intuitively, a quality ratio of *q* asserts that VEGEMATIC output is *q* times as good as dictionary examples, and *q* is generally less than 1. Higher ratios indicate higher quality.

Table 5: Quality ratio of VEGEMATIC to Dictionary examples for pilot vs. improved versions

Evaluation criteria (VEGEMATIC/dictionary)	**Mean (Standard Error)**		***p* value**
	Pilot version	**Improved version**	
General score	0.56 (0.076)	0.76 (0.040)	0.035
Good use	0.70 (0.093)	0.93 (0.049)	0.036
Constraining context	0.68 (0.095)	0.80 (0.045)	0.095
Comprehensibility	0.59 (0.098)	0.72 (0.032)	0.209

Table 5 demonstrates that the improved version outrated the pilot version in quality ratios for all four criteria. ANOVA shows significant differences for general score ($p < 0.05$) and good use of words ($p < 0.05$).

6.2. Lay judgment of coverage and precision

Our evaluation of quality was limited in two important respects. First, although it included 100 VEGEMATIC contexts, they were for only 10 target words. Second, although it employed careful ratings by a leading expert on vocabulary learning and instruction, this level of expertise is not typically available in situations where we envision VEGEMATIC being used to help create educational materials.

To address both limitations, we asked two college educated native English speakers ("rater A" and "rater B") to rate VEGEMATIC output simply as acceptable or not acceptable. Besides the 10 target words used in Section 6.1, we randomly selected 40 more of the 789 Tier 2 words in Reading Tutor stories, excluding the pilot words used in Section 4.1. As the Appendix shows, VEGEMATIC output at least one example for each of the 50 words. For each target word, the raters rated the first 10 contexts output by VEGEMATIC, or all of them if there were fewer than 10, for a total of 449 contexts.

But first we had to define an acceptability criterion clear enough for our raters to apply with high inter-rater reliability. This goal turned out to be difficult enough that it seems worthwhile to discuss our successive attempts as a useful cautionary tale, before reporting the evaluation results.

We initially asked raters to evaluate each context according to three questions intended both to enforce the constraints in Section 3, and to identify which ones VEGEMATIC violated the most, so as to shed light on which aspects future work should prioritize (grammaticality vs. semantic constraint vs. comprehensibility). A simple rating interface incrementally displayed each context and the three questions about it, recording the rater's response and response time to each question:

Please read the sentence silently and then press Enter:

We are willing and anxious to learn more!

Is this example a good English phrase or sentence? [y/n]

Does this example provide useful information about what "anxious" means? [y/n]

Is this example appropriate for a second grader? [y/n]

The program then cleared the screen and went on to the next example.

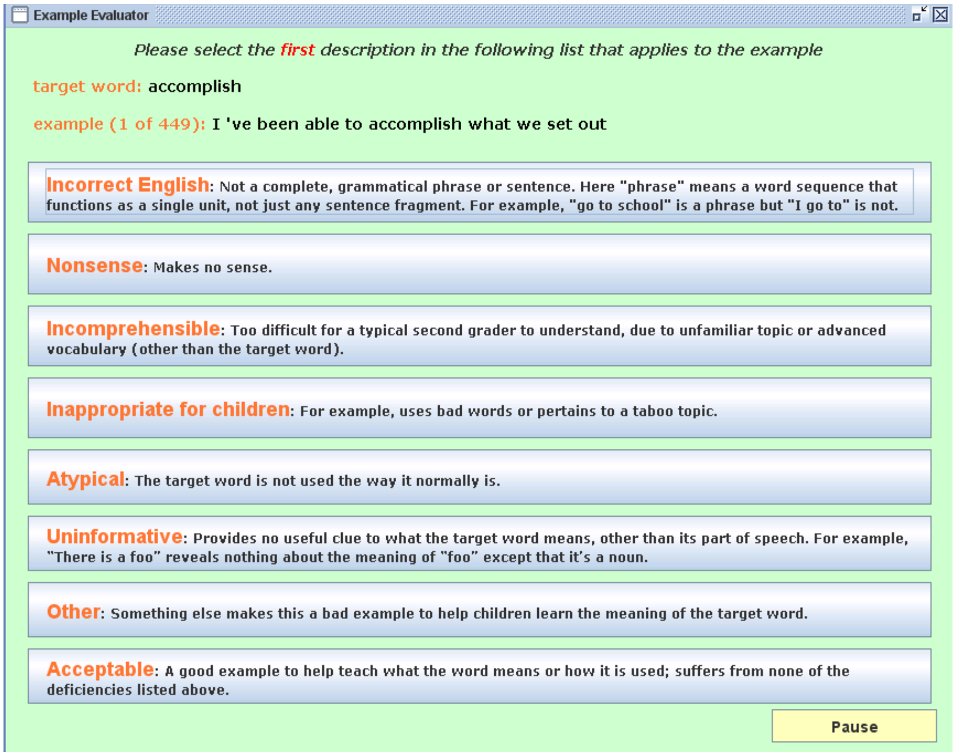
We used the 57 pilot contexts to evaluate inter-rater reliability. Cohen's Kappa for the two raters' responses to the first, second, and third questions was respectively only 0.270, 0.166, and 0.354, and 0.096 for the conjunction of all three responses. Kappa of .4-.6 is considered moderate, .6-.8 substantial, and .8-1 outstanding (Landis and Koth 1977). Our low Kappa values revealed the raters' lack of consensus, presumably due to differing interpretations of the questions.

To elucidate these differences, we selected examples where the raters disagreed and asked them to explain their answers. For each of the 3 questions, we randomly selected 10 examples where the raters disagreed, 5 rated "Yes" by rater A and "No" by rater B, and the other 5 vice versa. Analysis of their explanations showed differing interpretations of vague, ambiguous phrases such as "good English."

To clarify instructions, we rephrased them in terms of 8 finer-grained, more specific categories, e.g. "inappropriate for children," added a brief clarification of each one, e.g., "uses bad words or pertains to a taboo topic," and asked the raters to select the first applicable category for each example. We then calculated their inter-rater reliability both for overall acceptability and for the category selected, since they rejected some of the same contexts for different reasons. Both inter-rater reliabilities were still low, between 0.2 and 0.3, so we asked the raters to discuss the examples they disagreed on until they reached consensus.

As Figure 2 shows, we modified the rating interface accordingly. It displayed the target word and a context, and prompted the rater to click on the first applicable category.

To encourage raters to check for each listed deficiency, the menu of categories listed "Acceptable" last.



Please select the first description in the following list that applies to the example

target word: accomplish

example (1 of 449): I 've been able to accomplish what we set out

Incorrect English: Not a complete, grammatical phrase or sentence. Here "phrase" means a word sequence that functions as a single unit, not just any sentence fragment. For example, "go to school" is a phrase but "I go to" is not.

Nonsense: Makes no sense.

Incomprehensible: Too difficult for a typical second grader to understand, due to unfamiliar topic or advanced vocabulary (other than the target word).

Inappropriate for children: For example, uses bad words or pertains to a taboo topic.

Atypical: The target word is not used the way it normally is.

Uninformative: Provides no useful clue to what the target word means, other than its part of speech. For example, "There is a foo" reveals nothing about the meaning of "foo" except that it's a noun.

Other: Something else makes this a bad example to help children learn the meaning of the target word.

Acceptable: A good example to help teach what the word means or how it is used; suffers from none of the deficiencies listed above.

Figure 2: Context rating interface

The same two raters used the revised rating interface to rate the 449 contexts generated by the improved version of VEGEMATIC, averaging 9.9 seconds per context. In deciding whether an example was acceptable overall, they achieved inter-rater reliability of Kappa = 0.422, considered moderate. However, their inter-rater reliability in selecting the specific category was still low (Kappa = 0.278), reflecting the difficulty of pinning down what's wrong with a context.

We hoped that response times might provide a more sensitive measure of acceptability. However, although correlation between the two raters' response times was significant ($p < .01$), it was too small (.241) to use response time as a reliable measure. Nevertheless, response times were still informative. Table 6 shows the two raters' mean response times, disaggregated by acceptable and not acceptable. Both raters averaged about 3 seconds longer on contexts they rejected, presumably to diagnose their deficiencies. Figure 3 shows mean response times disaggregated by category selected. The two raters had similar patterns of response times, identifying some deficiencies, e.g., "Incorrect English" and "Atypical," faster than others, e.g., "Incomprehensible." Assuming that raters considered successive categories in the order listed on the menu, and selected the first applicable category, one would naturally expect them to take longer to select categories further down on the list. However, category position in the menu did not correlate with response time, so it does not explain these differences.

Table 6: Raters' response times disaggregated by acceptable vs. not acceptable

	Rater A			**Rater B**			**Average (Mean)**
	# of examples	**Mean**	**Std. err**	**# of examples**	**Mean**	**Std. err**	
Acceptable	181	6.70s	0.54s	101	9.19s	1.48s	7.94s
Unacceptable	268	9.47s	0.56s	348	12.18s	0.79s	10.83s

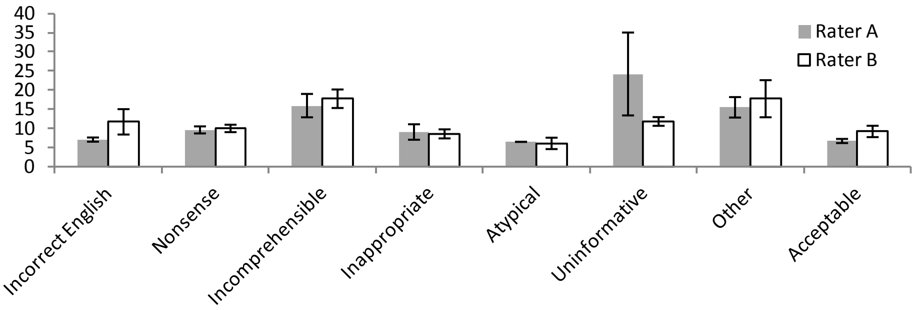


Figure 3: Response times (in seconds, with standard error bars), disaggregated by category selected

Finally armed with a moderately reliable acceptability criterion, we evaluated two metrics: *coverage*, the percentage of words with at least one context rated acceptable; and *precision*, the percentage of contexts rated acceptable. As Table 7 shows, coverage and precision across the two raters averaged 85% and 28%, respectively. If both raters must agree that a context is acceptable, coverage and precision drop to 76% and 19%, respectively.

Table 7: Coverage and precision based on two lay raters

Measure	**Rater A**	**Rater B**	**Rater Mean**	**Both Raters**
Coverage	0.92	0.78	0.85	0.76
Precision	0.36	0.20	0.28	0.19

Has VEGEMATIC reached the point where it is faster to use than to write contexts by hand? Our content developers reported taking 30 seconds to 3 minutes to create the first context for a word by hand, and actually slightly longer for each additional context. In comparison, raters averaged about 10 seconds to classify a context as acceptable or not, and about 27 seconds to find the first context they accepted for a word. (We ignore the time to run VEGEMATIC because it could be done off-line.) Thus hand-vetting VEGEMATIC's output and resorting to manual authoring only if none of the output is acceptable may already be faster than creating a context by hand. However, two caveats are important to mention. First, 27 seconds to find the first acceptable context in a list doesn't say how long it would take to find additional contexts, especially if they must be diverse. Second, even if judged acceptable by lay raters, VEGEMATIC's output may not be as good as contexts written by specialists.

7. Relation to Prior Work

Many sources of knowledge have been utilized in order to help children learn vocabulary, including machine-readable dictionaries and thesauruses such as WordNet (Fellbaum 1998), pictures and sounds for concrete vocabulary, and explanations of words in terms of other words. While other work in reading tutoring (Aist 2001, 2002; Brown *et al.* 2005; Heiner *et al.* 2006) has explored some of these other resources, VEGEMATIC focuses on providing additional contexts to help students learn the meaning of words, and thus VEGEMATIC is most related to prior work on generation and selection of example sentences.

Context generation is a kind of natural language generation, which aims to produce understandable texts in human language from some underlying meaning representation, normally via a three-stage process consisting of content selection, sentence planning, and surface realization (Reiter and Dale 2000). Selection of best realizations has used symbolic, knowledge-based, and statistical methods, e.g., N-gram language models. However, natural language generation typically proceeds from the meaning level to a surface string. In contrast, VEGEMATIC starts from surface level contexts in the form of N-grams, and uses information at the meaning level (such as related words) to guide their extension into contexts that convey some of the meaning of the target word.

Some related work has automated the generation of example sentences in spoken dialogue systems. Dowding et al. (2003) used a grammar to generate example sentences containing specific words (e.g., ***pressure*** and ***commander*** in the sentence *Measure the **pressure** at the **commander**’s seat*) for targeted help in spoken dialogue systems. Our work involves a different population (children), purpose (vocabulary development), and method (combining N-grams).

Other related work has automated the selection of example sentences for vocabulary learning and assessment (Brown *et al.* 2005; Liu *et al.* 2005; Pino *et al.* 2008). Some selection criteria (Pino *et al.* 2008) resemble constraints incorporated in VEGEMATIC as filters. However, the previous methods select complete sentences from an existing language corpus, whereas VEGEMATIC generates contexts. Why prefer generation over selection?

First, context generation can produce more examples than context selection. A simple theoretical argument illustrates this point: Given any corpus, the sentences in that corpus can be recombined to produce (generate) additional possible sentences, many of which are not already in the corpus and hence cannot be selected from it. The novel sentences generated by crossover vary in quality, but some of them are good, such as *Find the strength and **courage** to take risks*. Other methods to generate new sentences are possible, such as grammar-based approaches, template generation, and genetically-inspired recombination.

Second, context generation can also produce more typical contexts than selection does. N-grams aggregate information across many sentences, so their counts reflect typicality of usage. In contrast, the corpus count of a complete sentence is usually 1, which does not indicate whether it uses the target word in a typical way.

Third, context generation could potentially customize output to meet particular goals, somewhat as Oberlander, Karakatsiotis, Isard, and Androutsopoulos (2008) personalize object descriptions in Second Life museums. In the realm of vocabulary learning many customizations are possible, such as using a child’s name, characters from a story, or recently mentioned concepts or words when explaining a difficult word.

8. Conclusion

We now summarize research contributions, limitations, and future work.

8.1. Contributions

This article makes several contributions to automated generation of example contexts to help children learn vocabulary.

First, we introduce the problem of automatic context generation for learning vocabulary. Although the importance of context to learning vocabulary is well-known, context examples used in education have been created by hand or selected automatically from a corpus, and we know of no prior work on generating them automatically.

Second, we show how to generate contexts by combining Google five-grams. We identify several constraints on generating good example contexts, and describe filters to operationalize these constraints. We present heuristic methods for clustering, scoring, and selecting among generated examples. This framework is flexible enough to fit other instructional purposes. For example, we adapted VEGEMATIC to generate contexts for fluency practice simply by disabling filters related to word meaning.

Third, we evaluate quality, coverage, and precision. A vocabulary expert rated VEGEMATIC’s output below dictionary examples but approximately as good for teaching target word meaning as sentences from children’s stories, and better than sentences retrieved by Google. VEGEMATIC achieved over 80% coverage of the target words in terms of outputting at least one context rated acceptable by lay judges, and almost 20% precision in terms of the percentage of contexts they rated acceptable.

Vetting contexts was fast enough to suggest that sifting through VEGEMATIC's output to find acceptable contexts may already be more efficient than creating them by hand.

8.2. Limitations and future work

VEGEMATIC has limitations at syntactic, semantic, pragmatic, and social levels; while addressing these remaining limitations is not the main direction of this article, doing so could serve as fruitful directions for future research.

At the syntactic level, some generated contexts are incomplete or ungrammatical, due mainly to noise in Google N-grams and limitations of the grammar checker. For example, some end-of-sentence five-grams are not really the ends of sentences. A better grammar checker would filter out ungrammatical contexts more thoroughly.

At the semantic level, VEGEMATIC ignores the issue of multiple word senses with the same part of speech. Overcoming this limitation requires constraining VEGEMATIC to generate contexts tailored to specific target meanings, which is addressed in a separate paper (Mostow and Duan 2011) .

At the pragmatic level, crossover generates some contexts with long-distance inconsistencies. For example, *I will have a* ***tremendous*** *impact on my life* is semantically acceptable, but pragmatically awkward.

At the social level, contexts must suit their intended audience. VEGEMATIC outputs some contexts that children lack the background to understand, such as *I declare the motion carried*. Likewise, VEGEMATIC outputs some contexts that violate social norms even though they contain no taboo words, e.g., *She reaches her* ***slender*** *fingers towards my exploding*. A topic filter might mitigate both problems, but human judgment will likely remain necessary to weed out socially inappropriate contexts.

A key follow-on question is how much, how well, and how easily students in the target population can learn vocabulary from contexts generated by VEGEMATIC. Comparison might also be made to additional sources of text, such as example sentences written by in-service teachers whose expertise on children's vocabulary learning presumably lies somewhere between our vocabulary expert's and our lay judges'.

References

Aist, G. 2001. Towards automatic glossarization: Automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. *International Journal of Artificial Intelligence in Education* 12: 212-231.

Aist, G. 2002. Helping Children Learn Vocabulary During Computer-Assisted Oral Reading. *Educational Technology and Society* 5(2): http://ifets.ieee.org/periodical/vol_2_2002/aist.html.

Beck, I. L., McKeown, M. G., and Kucan, L. 2002. *Bringing Words to Life: Robust Vocabulary Instruction*. NY: Guilford.

Beck, I. L., McKeown, M. G., and McCaslin, E. S. 1983. Vocabulary development: All contexts are not created equal. *Elementary School Journal* 83: 177-181.

Biemiller, A. 2009. *Words Worth Teaching: Closing the Vocabulary Gap*. Columbus, OH: SRA/McGraw-Hill.

Bolger, D. J., Balass, M., Landen, E., and Perfetti, C. A. 2008. Contextual variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes* 45(2): 122-159.

Brants, T., and Franz, A. 2006. Web 1T 5-gram Version 1. Philadelphia, PA: Linguistic Data Consortium.

Brown, J. C., Frishkoff, G. A., and Eskenazi, M. 2005, October 6-8. Automatic Question Generation for Vocabulary Assessment. *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, B.C., Canada, pp. 819-826. Stroudsburg, PA, USA: Association for Computational Linguistics.

Carbonell, J., and Goldstein, J. 1998, August 24-28. The use of MMR diversity-based reranking for reordering documents and producing summaries. *Proceedings of* *the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 335-336. New York, NY: ACM.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., and Frey, J. 2006, August 8-12. Context-Based Machine Translation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, Cambridge, Massachusetts, USA.

Carlson, A., and Fette, I. 2007, December 13-15. Memory-based context-sensitive spelling correction at web scale. *Proceedings of the Sixth International Conference on Machine Learning and Applications*, pp. 166-171. Washington, DC: IEEE Computer Society.

Chandler, D. 2004. *Semiotics: The Basics* (2 ed.): Routledge.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22-29.

Dowding, J., Aist, G., Hockey, B. A., and Bratt, E. O. 2003, March 24-26. Generating Canonical Example Sentences using Candidate Words. *Working Papers of the 2003 AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, Palo Alto, California, pp. 23-27. Menlo Park, CA: AAAI Press.

Durme, B. V., Qian, T., and Schubert, L. 2008, August 18-22. Class-Driven Attribute Extraction. *22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, UK, pp. 921–928. Stroudsburg, PA: Association for Computational Linguistics.

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

Finkel, J. R., Grenager, T., and Manning, C. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan, pp. 363-370. Stroudsburg, PA: Association for Computational Linguistics.

Fukkink, R. G., Blok, H., and Glopper, K. d. 2001. Deriving word meaning from written context: A multicomponential skill. *Language Learning* 51(3): 477-496.

Heiner, C., Beck, J. E., and Mostow, J. 2006, June 26-30. Automated Vocabulary Instruction in a Reading Tutor. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, pp. 741-743. Lecture Notes in Computer Science, Vol. 4053. Berlin: Springer Verlag.

Herman, P. A., Anderson, R. C., Pearson, P. D., and Nagy, W. E. 1987. Incidental acquisition of word meaning from expositions with varied text features. *Reading Research Quarterly* 22(3): 263-284.

Hermjakob, U., Knight, K., and Iii, H. D. 2008. Name Translation in Statistical Machine Translation - Learning When to Transliterate. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, pp. 389-397. Stroudsburg, PA: Association for Computational Linguistics.

Jenkins, J. R., Stein, M., & Wysocki, K. 1984. Learning vocaulary through reading. *American Educational Research Journal* 21: 767-787.

Kuhn, M. R., and Stahl, S. A. 1998. Teaching children to learn word meaning from context: A synthesis and some questions. *Journal of Literacy Research* 30(1): 119-138.

Landis, J. R., and Koth, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159-174.

Liu, C.-L., Wang, C.-H., Gao, Z.-M., and Huang, S.-M. 2005, June 29. Applications of lexical information for algorithmically composing multiple-choice cloze items. *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, Ann Arbor, Michigan, pp. 1-8. Stroudsburg, PA: Association for Computational Linguistics.

McKeown, M. G. 1985. The acquisition of word meaning from context by children of high and low ability. *Reading Research Quarterly* 20: 482-496.

Mostow, J. 1983. Machine transformation of advice into a heuristic search procedure. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (eds.), *Machine Learning*, pp. 367-403. Palo Alto, CA: Tioga.

Mostow, J., and Aist, G. S. 1999. Giving help and praise in a reading tutor with imperfect listening -- because automated speech recognition means never being able to say you're certain. *CALICO Journal* 16(3): 407-424.

Mostow, J., and Duan, W. 2011, June 24. Generating Example Contexts to Illustrate a Target Word Sense. *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, Portland, OR, pp. 105-110. Stroudsburg, PA: Association for Computational Linguistics.

Nagy, W. E., Anderson, R. C., and Herman, P. A. 1987. Learning Word Meanings from Context during Normal Reading. *American Educational Research Journal* 24(2): 237-270.

Nagy, W. E., Herman, P. A., and Anderson, R. C. 1985. Learning words from context. *Reading Research Quarterly* 20(2): 233-253.

NRP. 2000. *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (00-4769). Washington, DC: National Institute of Child Health & Human Development. At www.nichd.nih.gov/publications/nrppubskey.cfm.

Oberlander, J., Karakatsiotis, G., Isard, A., and Androutsopoulos, I. 2008, April 9-12. Building an adaptive museum gallery in Second Life. *Proceedings of Museums and the Web: the international conference for culture and heritage on-line*, Montréal, Québec, pp. 749-753.

Paynter, D. E., Bodrova, E., and Doty, J. K. 2005. *For the love of words: vocabulary instruction that works, grades K-6*. San Francisco: Jossey-Bass.

Pino, J., Heilman, M., and Eskenazi, M. 2008. A selection strategy to improve cloze question quality. *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems*, Montreal, Canada, pp. 22-34.

Reiter, E., and Dale, R. 2000. *Building natural language generation systems*: Cambridge University Press.

Schatz, E. K., & Baldwin, R. S. 1986. Context clues are unreliable predictors of word meanings. *Reading Research Quarterly* 21: 439-453.

Schwanenflugel, P. J., Stahl, S. A., and McFalls, E. L. 1997. Partial Word Knowledge and Vocabulary Growth during Reading Comprehension. *Journal of Literacy Research* 29(4): 531-553.

Sleator, D., and Temperley, D. 1993, August 10-13. Parsing English with a link grammar. *Third International Workshop on Parsing Technologies*, Tilburg, NL, and Durbuy, Belgium.

Stanovich, K., West, R., and Cunningham, A. E. 1991. Beyond phonological processes: Print exposure and orthographic processing. In S. Brady and D. Shankweiler (eds.), *Phonological Processes in Literacy*, pp. 219-235. Hillsdale, NJ: Lawrence Erlbaum Associates.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. 2003, May 27 - June 1. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of the Human Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, pp. 252-259. Stroudsburg, PA: Association for Computational Linguistics.

Toutanova, K., and Manning, C. D. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, Hong Kong, pp. 63-70. Stroudsburg, PA: Association for Computational Linguistics.

Yu, L.-C., Wu, C.-H., Philpot, A., and Hovy, E. 2007, November 11. OntoNotes: Sense Pool Verification Using Google N-gram and Statistical Tests. *Proceedings of the OntoLex Workshop at the 6th International Semantic Web Conference (ISWC 2007)*, Busan, Korea.

Appendix: Examples of Contexts Output by VEGEMATIC

To give a flavor of VEGEMATIC's output, both good and bad, we present a representative sample selected from the examples rated by lay judges as described in Section 6.2. To avoid unconscious bias, we used the following selection procedure:

1. For good examples, choose randomly from the examples rated acceptable by both raters, marked +. If none, choose randomly from examples rated acceptable by only one rater, marked +?. If none, show a blank line, marked +/.
2. For bad examples, choose randomly from the examples rated unacceptable by both raters, marked –. Every word had at least one such unacceptable example.

Target word		**Context**
accomplish	+?	it is possible to accomplish great things in life !
	–	what you need to accomplish the work described herein
advantage	+?	We have a huge advantage over the average person ?
	–	each with a distinct advantage in today 's fast
advice	+?	We are pleased to give advice on getting the best
	–	Thanks for all the advice you need for home care
amuse	+	Small things amuse small minds !
	–	Your best way to amuse is the game ?
ancient	+	Learn the language of ancient Rome
	–	It was the ancient capital of the State Address
appease	+	He is trying to appease the masses of people
	–	in order to appease the public 's health
arranged	+	These meetings are arranged by the author !
	–	Photos are arranged into four groups based
ascend	+	From here we ascend from earth to heaven
	–	They ascend into the hill of Mars
boasting	+/	
	–	and two bedroom apartments boasting a living space ?
budge	+	He refused to budge in the past week
	–	He refused to budge in the past three
ceremonies	+	The grand opening ceremonies for the 2006 program
	–	the opening ceremonies of the American dream
compete	+	People in your neighborhood compete for your business with pleasure
	–	You 'll be able to compete on a level playing
concerned	+	I think people are more concerned with quality of service
	–	It did not seem concerned about the war Permalink
consented	+	I would never have consented to the application form

	–	Remember that the person consented to taking a test ?
creature	+	It 's the most beautiful creature he had ever seen
	–	Design a deep sea creature act
custom	+	Looking for custom made men 's dress shirts
	–	All available in custom sizes and colors are available
delicate	+	We cut through the delicate fabric in no time !
	–	order to preserve the delicate balance of life and death
enormous	+	It has an enormous collection of free online games
	–	Even with the enormous amount of people with disabilities
entire	+?	This will copy the entire contents of the measuring range
	–	We carry the entire line of products in BOOKS
eventually	+	Even if you eventually decide to go ?
	–	Say what we will eventually find a way to
fatigue	+	The pain and fatigue caused by blood clots
	–	edges to help prevent foot fatigue
ferocious	+	It is a ferocious attack on the village
	–	It is a ferocious attack on the United
force	+	It 's a very powerful force to be reckoned with
	–	through a direct sales force and customer service levels
fret	+	Land on the first fret with your index finger
	–	Meets on the first fret with your index finger
gained	+	France had gained control of the West Bank
	–	the years we have gained a good deal ?
gasped	+?	She moaned and gasped for breath !
	–	The young gay man gasped as he looked ?
gasp	+	Laugh until you gasp for breath !
	–	I hear a gasp from the other two
gratefully	+	I would gratefully appreciate any information !
	–	This information will be gratefully received so that we know
harshly	+	Do not judge me too harshly
	–	He was harshly critical of the Bush
imagine	+	I always try to imagine what might have happened
	–	Do not even start to imagine what it would really help
impulsive	+	They tend to be impulsive and live life !
	–	Such thoughtless and impulsive buying will most likely come
incredible	+	Witness the incredible true story of three women
	–	Just look at our incredible deals on a wide variety
injury	+	Maybe you have suffered an injury in the fourth quarter
	–	Show all types of brain injury can be caused by
intense	+	I felt an intense desire to find the best
	–	I have an intense fear of becoming fat
jeer	+/	
	–	now jeer the time hard edge poem

misfortune	+	I had the misfortune of dealing with terrorism
	–	have to suffer the misfortune of being placed !
nimble	+/	
	–	The light and nimble handling of the new millennium
obliged	+	I shall be obliged to return to home
	–	It would be highly obliged if given the chance
observed	+	City Council observed a moment of silence
	–	concentrations were observed in the brain ?
opportunity	+	This is a great opportunity to take advantage of
	–	students with the opportunity to become better prepared ?
overcome	+	Learn to overcome your fear of public speaking
	–	p1 design can not overcome the power of sale
perhaps	+?	He said it was perhaps the worst of all
	–	Pray for them and perhaps even use the information ?
pierced	+?	Her heart was pierced through the heart muscle
	–	Need to get my eyebrow pierced !
recognized	+?	The company also has been recognized for a long time
	–	the world 's most recognized names in the country ?
released	+	Watch the funniest movies released for 1999 !
	–	The hotel rooms will be released back into the wild
searched	+	How often have you searched for a store name ?
	–	Add this item also searched for 1 x 10
startled	+	She was suddenly startled by a loud noise ?
	–	I have been startled by the sight and sound
sturdy	+	It provides a sturdy handle for easy carrying
	–	Prints are shipped in sturdy tubes to ensure safe operatin
torrents	+/	
	–	More xxx torrents are available on line
vanished	+	And then it vanished into the thin air ?
	–	it has vanished from the sky moissanite