Towards Computational Guessing of Unknown Word Meanings:
The Ontological Semantic Approach

Julia M. Taylor (jtaylor1@purdue.edu)
CERIAS, Purdue University & RiverGlass, Inc
West Lafayette, IN 47907 & Champaign, IL 61820

Victor Raskin (vraskin@purdue.edu)
Linguistics & CERIAS, Purdue University
West Lafayette, IN 47907

Christian F. Hempelmann (chempelm@purdue.edu)
Linguistics, Purdue University & RiverGlass, Inc
West Lafayette, IN 47907 & Champaign, IL 61820

Abstract

The paper describes a computational approach for guessing the meanings of previously unaccounted words in an implemented system for natural language processing. Interested in comparing the results to what is known about human guessing, it reviews a largely educational approach, partially based on cognitive psychology, to teaching humans, mostly children, to acquire new vocabulary from contextual clues, as well as the lexicographic efforts to account for neologisms. It then goes over the previous NLP efforts in processing new words and establishes the difference—mostly, much richer semantic resources—of the proposed approach. Finally, the results of a computer experiment that guesses the meaning of a non-existent word, placed as the direct object of 100 randomly selected verbs, from the known meanings of these verbs, with methods of the ontological semantics technology, are presented and discussed. While the results are promising percentage-wise, ways to improve them within the approach are briefly outlined.

Keywords: guessing word meaning, natural language understanding, ontological semantic technology

Unknown Words in Text

Along with ambiguity, unattested input is one of the major problems for natural language processing systems. An NLP system is robust only if it can deal with unknown words. Yet, to deal with such words only makes sense when the rest of the sentence is understood. We take an approach here similar to that of a human learner that encounters an unfamiliar word and is able to approximate its meaning based on the rest of the sentence or its subsequent usages in other sentences.

There are some suggested strategies in the human acquisition and understanding of unknown words. Some cases stand out as easy and almost self-explanatory. One of these cases is when a word is immediately explained. Such an explanation may be introduced by a *that is* phrase (*To lose weight, one may have to follow a diet, that is, to limit the amount of food and to avoid eating certain foods.*), or by apposition (*Computers programs follow algorithms, ordered lists of instructions to perform*.), or by examples (*The earliest records of felines, for example, cats, tigers, lions, or leopards, are from millions of years ago.*), or by providing the presumably known opposites for comparison through words like *but, rather then, not* (*It is frigid outside, rather than warm and comfortable like yesterday*.).

Both in the case of human acquisition of new vocabulary and the machine attempt at guessing its meaning, these somewhat trivial instances, where the meaning of a new word is immediately explained, either by giving its definition or by examples, present no particular interest for us here. Besides, such cases are rather rare in regular expository texts because most writers do not bother to allow for vocabulary deficiency with regards to words with which they are well familiar themselves. Thus, it is the non-trivial cases, those without an attached explanation or description, that it is necessary to address when one is interested in designing a computer system for natural language understanding.

On the other side of the spectrum lie words that can only be guessed through their functional description, not necessarily following the first use of an unknown word. These functional descriptions should be gathered throughout the document, or a number of documents, narrowing the original functional description, if necessary, or supplying other facets of it. For example, *They used a Tim-Tim to navigate their way to the cabin on the lake. It took them almost half a day. They hadn't checked if the maps had been recently updated on the device, and spent hours looking for roads that no longer existed.* From the clues in the first sentence, *Tim-Tim* can be understood as a navigation instrument (including an atlas or a map) through an inverse function of the instrument of navigation. Since no other devices are mentioned, this navigation instrument can be considered the device from the third sentence whose maps can be periodically updated. It is essential, therefore, in situations of dispersed clues that co-reference (or antecedence) be established correctly—in this case, between *device* and *Tim-Tim*.

Towards the middle of the spectrum are the cases where the description may immediately follow the first use of the word but without being helpfully triggered by phrases like *for example* or *that is* (*He was rather taciturn. He didn't like*

small talk, rarely spoke in society, hardly said anything other than hello when meeting people.). The word *taciturn* in the first sentence is used as a description of *he.* The second sentence continues the description of the same person. We could assume that *taciturn* somehow overlaps with this description, and then, at the very least, we know that the unknown word refers to a person's temperament and/or communication style, and it is on the quiet side.

The difference between this text and the previous example is that the former consistently describes a person, to the point that *that is* could be added in (*He was rather taciturn, that is, he didn't like small talk, rarely spoke in society, hardly said anything other than hello when meeting people.*). The latter one does not contain a focused description but rather disperses the clues throughout the short narrative, and the meaning of the word has to be derived based on what function it could play in the situation.

We are ultimately interested in the more challenging case of dispersed clues, where the functional details must be collected to identify the meaning of the unknown word. This paper describes an experiment where functional details of a single sentence determine the meaning of the word.

Previous Research on Unknown Words

Human Vocabulary Acquisition

The problem of understanding unknown words has been addressed both with regard to humans, in first and second language acquisition, and to computers. With humans, it is known as vocabulary acquisition and enrichment. As children and second language learners increase their vocabulary, enabling as well as expediting this process has been an important educational goal. Both methodologies for helping children and students guess the meanings of the unknown words from contextual clues and the metrics for evaluating such methodologies have been discussed in detail (see, for instance, Bransford and Johnson, 1972; Gipe, 1979; McKeown, 1985; Nagy et al., 1985; Rankin and Overholser, 1969—but cf. Schatz and Baldwin, 1986; and for a more contemporary perspective, Wagner et al., 2006).

There is also a number of useful anonymous websites with exercises and helpful tips for guessing the meanings of unknown words without looking them up in dictionaries or encyclopedias[1].

When dealing with human vocabulary acquisition, four types of situations are typically recognized (cf. Nagy et al., 1987):

- words that are in the oral vocabulary but not in the reading vocabulary
- new meanings for words that are already in the reading vocabulary with one or more other meanings
- words that are in neither the oral vocabulary nor their reading vocabulary and for which there is no concept available but for which a concept can be easily built
- words that are neither in the oral vocabulary nor the reading vocabulary, for which there is no concept available, and for which a concept cannot be easily built

Concepts are used in this kind of research in an undefined, somewhat pre-scientific sense of general ideas, possibly underlying synonyms (see, for instance, Wisniewski, 1996, 1997a,b, 2000), the sense that came to be used in much later ontologies as means to control terminological usage (see Raskin et al., 2008).

Metrics are created to measure the strength of contextual support, determined on the basis of raters' judgments as to how much information the text provides about the meaning of an unfamiliar word (see Nagy et al., 1987). Success is also seen as depending on the readability of a text, a controversial measure, usually defined in terms of sentence length and difficulty of words, and on the "density of hard words."

These parameters of human cognition pertain clearly to the computation of learning new words. Thus, to compare with the four learning situations considered in human learning above, two different types of unattested input can be encountered in computational systems: unique unattested words with no other (related or unrelated) senses already known and new senses of known words. For each of those, a concept may be available in the existing ontology or needs to be acquired, thus resulting in similar four situations.

Lexicography

Partially, the NLP effort concerning unknown words overlaps with another human effort, namely, dealing with neologisms, a traditional concern of lexicography, the art—rather than the science—of dictionary making. Algeo (1977, 1980, 1991, 1993) provides frequently cited information on the source and typology of words entering the language (cf. also Barnhart, 2007; Lehrer, 2003; O'Donovan and O'Neil, 2008; Sheidlower, 1995; Simpson, 2007). Recent developments in electronic media have led to new spellings of known words (e. g., *4 u*), their new forms or senses (*to (un)friend*) and new words (*twitter*). This has engendered a recent concern for the normalization of e-mail/SMS text (Aw et al., 2006; Choudhury et al., 2007).

Natural Language Processing

While some NLP efforts focus on determining just the part-of-speech (POS) of an unknown word (Mikheev, 1997; Murawaki and Kurohashi, 2008; Ciramita, 2002), others attempt to guess its lexical/semantic class, most prominently if it is a proper name (Smarr and Manning, 2002; Bikel et al., 1999; Collins and Singer, 1999; Cucerzan and Yarowsky, 1999; Buchholz and Van Den Bosch, 2000, Nadeau and Sekine, 2009). This results in a large-scale effort in named entity recognition, especially in languages like Chinese and Korean, where there is no helpful capitalization, and many names, including foreign ones, utilize the characters for regular words, using their phonetic values to approximate the foreign pronunciation. Most NLP work on unknown words is done in the statistical and/or

[1]English-Zone.com, www.sinclair.edu/centers/tlc/... /learning_words_from_context_clues.pdf

machine learning paradigm (Weischedel et al., 1993: Campbell and Johnson, 1999: Ciaramita and Johnson, 2003; Curran, 2005), without 'understanding' the contextual clues.

Guessing Meaning of Unknown Words With the Ontological Semantic Technology

In this section, we will demonstrate in somewhat simplified non-proprietary technical detail, how the meaning of an unknown word is determined on the basis of the full understanding of everything else in the sentence. This will be done with the methods and resources of the Ontological Semantic Technology (OST—see Raskin et al., 2010, Taylor et al., 2010, and Hempelmann et al., 2010).

At the core of OST are repositories of world and linguistic knowledge, acquired semi-automatically (Hempelmann et al., 2010, Taylor et al., 2010) within the approach and used to disambiguate the different meanings of words and sentences and to represent them. These repositories, also known as the static knowledge resources, consist of the ontology, containing language-independent concepts and relationships between them; one lexicon per supported language (for example, English), containing word senses anchored in the language-independent ontology which is used to represent their meaning; the Proper Name Dictionary (PND), which contains names of people, countries, organizations, etc., and their description anchoring them in ontological concepts and interlinking them with other PND entries; and a common sense rules resource. A conforming lexicon and ontology, as well as PNDs and common sense rules, are used by the Semantic Text Analyzer (STAn), a software, developed by RiverGlass Inc., that produces Text Meaning Representations (TMRs) from the text that it reads. The format of TMRs conforms to the format and interpretation of the ontology. The processed TMRs are entered into InfoStore, a dynamic knowledge resource of OST, from which information is used for further processing and reasoning.

Thus, just as in cognitive psychology, underlying some of the pedagogical research on vocabulary acquisition, where "the reader is seen as building a mental representation of the textual meaning based on information contained in the text and on the activation of complementary knowledge resources (van Dijk & Kintsch, 1983, Johnson-Laird, 1983)" (Rieder, 2002: 54), STAn is constructing TMRs by processing text with the help of the OnSe knowledge resources. From this perspective, the unknown word tasks can be seen as finding a formalized solution to a cloze test (Taylor, 1953), where every nth word of a test is deleted and the participant is asked to reconstruct these omitted words. Because this "inferencing" of the meaning of an unknown word is done by humans on the basis of context as well as language and world knowledge, OST models context as the syntactic environment of the unknown word mapped onto the concepts found in this environment and the constraints these concepts place on the word. The concepts and their properties represent the world knowledge required for the task.

As such our approach a more mature version of work on unknown words in NLP like that of Granger (1977), 'mature' here meaning not only that our resources are much richer, but also that the unknown word task is only one of the many that OST undertakes in the course of its processing of text towards the representation of its meaning. In this paper, we are illustrating our approach by the inferencing of noun meanings in relation to the meaning of the verbs of which they are direct objects, like Granger does. More recent approaches in the same vein include Cardie (1993), Hastings and Lytinen (1994), and not least the work by Wilks and colleagues on "lexical tuning" (Wilks, 1978; Wilks and Catizone, 2002), much in the spirit of very rich semantic resources underlying OST.

While other NLP approaches emphasize neologisms as the focus of their unknown-word effort (for a very recent overview see Cook, 2010), we realistically expect unattested input to contain existing words which have not yet found their way into our lexicon. Even with a 100,000-sense lexicon, only 10% or so of the lexical resources of a natural language would be covered, and unlike native speakers who make do with well under 50,000 words in their vocabularies (cf. Nation, 2006), a contemporary NLP application will typically go into highly specific technical terms or seek the explanation of a very rare word.

Computational Experiment

To guess any word in any syntactic position is an overly complex task. We simplified it for this paper by considering only those words that play the role of the direct object of a transitive verb, as a starting point. The selection task was twofold: we needed to select a number of verbs to test, and then we needed to select sentences that we could test these verbs in.

The description of the senses in the English OST lexicon contains an annotation field for the purposes of providing a human-readable definition and an example of the word sense in a sentence. The annotation fields were filled in the process of a sense acquisition, long before the experiment in this paper was thought of, and the acquirers did not have any constraining instructions in producing the examples. The format of the definition and example is not much different than that of any dictionary. It serves no purpose for the computer, which reads machine readable syn-struc, sem-struc and extracts the needed information from there.

The annotation examples are considered to be exemplars of sentences that the software should be able to process. Since the examples are free creations by acquirers, independent from our task, we considered them to be as appropriate for the task as any corpus selection, and it saved us the effort of looking for one verb match within a corpus. Thus, we selected these examples as the test sentences and replaced the direct object in each example with a word *zzz*– an unknown word to the system. The computer's task was to find ontological concepts that could be an interpretation of the word, based on the provided sentence.

Our lexicon contains 4469 senses of transitive verbs, not including verbs that could be either transitive or intransitive. From among the 4469 candidates, we randomly selected verbs until we reached 100 that could be processed using the example sentences that resulted in correct sense interpretation of the verb. We considered 189 verb senses, 59 of which contained no examples, and 30 of which produced an interpretation of a verb unacceptable to a human expert. In other words, the computer misunderstood the verb meaning.

The remaining 100 verbs whose example sentences passed the acceptability rating were considered for the unknown word test. Each direct object, defined by the syn-struc, was replaced in the example sentence with *zzz*. For example, the sentence for the verb *rethink, She decided she would rethink the new curtains before buying them for the whole house* became *She decided she would rethink zzz before buying them for the whole house.*

We added a file to our English lexicon with the word *zzz* and over 2000 senses of it, one for every event and object in our ontology. Thus, when processing the altered examples, STAn was able to consider every object and event as possible meanings of the unknown word *zzz*. To detect the meaning of *zzz*, the system should interpret the rest of the sentence, according to its ontological knowledge, while filling possible interpretations of *zzz*. The text-meaning representation (TMR) of the original sentence is:

```
(DECIDE
    (AGENT(HUMAN(GENDER(FEMALE)))
    (THEME(CONSIDER-INFO(ITERATION(MULTIPLE))
        (AGENT(HUMAN(GENDER(FEMALE)
        (THEME(INFORMATION
            (HAS-TOPIC(CURTAIN(NOVELTY(HIGH))))))
        (BEFORE(BUY
            (THEME(CURTAIN(HAS-LOCALE(HOUSE))))))
    )))
```

When zzz is inserted, the TMR becomes:

```
(DECIDE
    (AGENT(HUMAN(GENDER(FEMALE)))
    (THEME(CONSIDER-INFO(ITERATION(MULTIPLE))
        (AGENT(HUMAN(GENDER(FEMALE)
        (THEME(???))
        (BEFORE(BUY
            (THEME(???(HAS-LOCALE(HOUSE))))))
      )))
```

Looking at the above TMR, the semantic text analyzer needs to find the concepts that can satisfy the following:

- it is something that a human can rethink or it is information about something that a human can rethink
- it is a theme of BUY
- it is located in a HOUSE

Combining these clues, we have limited knowledge for determining a narrow sense of *zzz*—anything that fits into a house and can be bought can work here: furniture, decorative items, wall paint, china, etc. The resulting broad categories highlight the difficulty that a system faces: a concept denoting all décor works as well as that for a miniature.

Note that the first clue of the sentence about rethinking is practically useless: any object (physical, mental, or social) that can serve as THEME for thinking works. She could be rethinking a party, a paper topic, or curtains (the first and the third theme require the handling of ellipsis). Thus, without the second and the third clues, the interpretation could have been left as wide as any object or event.

We consider (native-speaker-) acceptable any interpretation that is reasonable within the context of the sentence, without any outside knowledge or emphasis. Thus, we consider it possible for a football player to be infuriated when a dog barks, for plants to be imported, and for water vehicles to be caulked.

As demonstrated above, the task did not necessarily restrict possible interpretations to a small number of concepts. To get a better handle on what the analyzer offered as its guesses, we considered the top five interpretations (TMRs) for each sentence, if the output contained at least five, and all of the interpretations when there were less than five. We then took the fraction of correct and incorrect interpretation of *zzz* in this sentence compared to the overall number of meanings considered (the largest overall number could be, of course, 5). Thus, if the system suggested 2 acceptable results and 3 unacceptable ones, we reported 0.4 and 0.6 respectably. If the system suggested only 1 result and it was acceptable, it was still counted as 1.

We found that the system suggested unacceptable meanings of *zzz* for 34.4% of the 100 senses. Out of the 65.6% acceptable meanings, 13% were considered to be no worse than a human could do.

In some cases (n=5), the analyzer used the intended meaning of the verb in the sample sentence, but switched the meaning of the verb when *zzz* was inserted. When the sentence made more sense with a different meaning of the verb with the chosen interpretation of *zzz* (n=2), it was counted as acceptable.

STAn generally prefers finer grain concepts to the coarser grain. Such a preference achieves the selection of, for instance, a human female over a general animal female in the resolution of an unreferenced usage of the pronoun *she*. Such a preference, however, usually backfires with the unknown word task, where it would be smarter to select the most generic concept for the constraints and narrow it down further only when more details are available. It is the selection of fine grain concepts that dominated the category of the acceptable but not preferred meanings. Thus, the example sentence *The constantly barking dog infuriated the neighbors,* once the word *neighbors* was substituted with *zzz*, led to 'the constantly barking dog infuriated the wide receiver' as the first interpretation. 'Wide receiver' here refers to a player position in American football, and one can imagine, in principle, a situation where that fury could be quite possible. The corresponding TMRs for the original example sentence and the *zzz* are shown below.

```
(anger(experiencer(personal-role))
    (cause(make-noise(volume(high))(pitch(low))
        (agent(dog))(iteration(multiple))
```

)))
(anger(experiencer(wide-receiver))
(cause(make-noise(volume(high))(pitch(low))
(agent(dog))(iteration(multiple))
)))

A perfect solution was achieved on sentences like *He shucked the corn*, with the original sentence and the *zzz*-replacement interpreted by STAn as:

(remove(theme(plant-part))
(agent(human(gender(male))))
(start-location(grain)))
(remove(theme(plant-part))
(agent(human(gender(male))))
(start-location(seed nut grain)))

At the opposite end of spectrum lie the sentences that were not interpreted by STAn at a level acceptable for a human judgment. One such sentence was *The engine emitted steam* and the substituted version *The engine emitted zzz.* The unacceptable interpretation of *zzz* was that of *shampoo, beer, wine,* and *yogurt*. Such misinterpretations are typically caused by the unnecessarily relaxed ontological constraints on some events. In this case, the event EXUDE (anchoring concept of this sense of *emit*) has a default theme of GASEOUS-MATERIAL, or LIQUID-MATERIAL, resulting in the acceptability of the above substances. On the other hand, the almost 2:1 ratio of the acceptable interpretations suggests that most of the ontology is well constrained.

We also wanted to know whether direct objects could be found using n-grams or other techniques that would take a subject and a verb of the sentence as input and return a possible direct object. We randomly selected ten verbs from our sample and ran a subject + verb query against a database of English concordances using the Brown, BNC written and BNC spoken[2] corpora. Only one query out of ten produced a non-zero result. Reducing the query to a single word, indicating the verb, produced seven non-zero results.

A similar search on Google produced many results, thus lowering a possibility that the selected verbs are not used in common speech. While the small number of attempted queries against the corpora should not be taken as a conclusive result, the number can be used as an indication of failure of finding appropriate words using non-conceptual representation (even for computational purposes). Thus, at least in guessing unknown words, some form of conceptual representation and conceptual hierarchy should be used for an attempt of approaching human-level competence.

Summary

We have demonstrated on an admittedly restricted purview that a meaning-based computational system of language understanding is capable of guessing the meaning of unknown words from the context, with the clues determined similarly to the way humans approach it. In the illustrated case, the context consisted primarily of the ontologically defined meaning of the known words directly related to the target word syntactically. Further improvements in the ontology and the lexicon as well as better grain size management within the software should improve the guessing results within a single sentence. Coreference and ellipsis resolution will facilitate bringing several sentences with their clues together and thus further improve the processing and interpretation of unknown words within the approach.

[2] http://www.lextutor.ca/concordancers/concord_e.html

References

Algeo, J. (1977). Blends, a structural and systemic view. *American Speech*, 52(1/2), 47–64.

Algeo, J. (1980). Where do all the new words come from. *American Speech* 55(4), 264–277.

Algeo, J. (Ed.) (1991). *Fifty years among the new words*. Cambridge: Cambridge University Press.

Algeo, J. (1993). Desuetude among new words. *International Journal of Lexicography* 6(4), 281–293.

Barnhart, D. K. (2007). A calculus for new words. *Dictionaries,* 28,132–138.

Beck, I., McKeown, M., & McCaslin, E. (1983). All contexts are not created equal. *Elementary School Journal,* 83, 177-181.

Bikel, D., Schwartz, R., & Weischedel, R. (1999). An algorithm that learns what's in a name. *Machine Learning* 34, 211–231.

Bransford, J., & Johnson, M. (1972). Contextual prerequisities for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior,* 11, 717-726.

Buchholz, S., & Van Den Bosch, A. (2000). Integrating seed names and n-grams for a named entity list and classifier. *Proc. of the 2nd International Conference on Language Resources and Evaluation* Athens, Greece.

Campbell, D. A., & Johnson, S. B. (1999). A technique for semantic classification of unknown words using UMLS resources. *Proc. of the AMIA Symposium* (pp. 716–720).

Cardie, C. (1993). A case-based approach to knowledge acquisition for domain-specific sentence analysis. *Proc. of the Eleventh National Conference on Artificial Intelligence* (pp. 798–803).

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., & Basu, A. (2007). Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition,* 10(3/4),157–174.

Ciaramita, M. (2002). Boosting automatic lexical acquisition with morphological information. *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, Philadelphia, PA (pp. 17–25).

Ciaramita, M., & Johnson, M. (2003) Supersense tagging of unknown nouns in WordNet. *Proc. of EMNLP* (pp. 594-602).

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. In P. Fung and J. Zhou (Eds.), *Proc. of EMNLP/VLC'99*. College Park, MD: ACL.

Cook, P. (2010). *Exploiting linguistic knowledge to infer properties of neologisms*. Ph.D. thesis, University of Toronto.

Cucerzan, S., & Yarowsky, D. (1999). Language independent named entity recognition combining morphological and contextual evidence. In P. Fung and J. Zhou (Eds.). *Proc. of EMNLP/ VLC'99*, College Park, MD: ACL.

Curran, J, R. (2005). Supersense tagging of unknown nouns using semantic similarity. *Proc. of ACL*, Ann Arbor, MI: (pp. 26–33).

Dijk, T. A. van & Kintsch, W. (1983). *Strategies of discourse comprehension*. Orlando: Academic Press.

Gipe, J. (1979). Investigating techniques for teaching word meanings. *Reading Research Quarterly,* 14, 624-644.

Granger, R. H. (1977). FOUL-UP: A program that figures out the meanings of words from context. *Proc. of the Fifth International Joint Conference on Artificial Intelligence,* Cambridge, MA (pp. 172–178).

Hastings, P. M. & Lytinen, S. L. (1994). The ups and downs of lexical acquisition. *Proc. of the Twelfth National Conference on Artificial Intelligence* (pp. 754–759).

Hempelmann, C. F., Taylor, J. M., and Raskin, V. (2010). Application-guided ontological engineering, H. A. Arabnia, D. de la Fuente, E. B. Kozerenko, and J. A. Olivas (Eds.), *Proc. of International Conference on Artificial Intelligence*.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Lehrer, A. (2003). Understanding trendy neologisms. *Italian Journal of Linguistics,* 15(2), 369–382.

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

Mikheev, A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics,* 23(3):405–423.

Murawaki, Y., & Kurohashi, S. (2008). Online acquisition of Japanese unknown morphemes using morphological constraints. *Proc. of EMNLP* (pp. 429–437).

Nadeau, D., & Sekine S. (2009). A survey of named entity recognition and classification. In S. Sekine & E. Ranchhod (Eds.), *Named entities: Recognition, classification and use*. Amsterdam: John Benjamins.

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

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal,* 24(2), 237-270.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review,* 63, 59-82.

O'Donovan, R., & O'Neil, M. (2008). A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. *Proc. of the 13th Euralex International Congress* (pp. 571–579).

Rankin, E., & Overholser, B. (1969). Reaction of intermediate grade children to contextual clues. *Journal of Reading Behavior,* 2, 50-73.

Raskin, V., Buck, B., Keen, A., Hempelmann, C. F., & Triezenberg, K. E. (2008). Accessing and manipulating meaning of textual and data information for information assurance and security and intelligence. In F. Sheldon (Ed.), *Proc. of the Fourth Cyber Security and Information Intelligence Research Workshop (CSIIRW'08)*, ACM Digital Library.

Raskin, V., Hempelmann, C. F., & Taylor, J. M. (2010). Guessing vs. knowing: The two approaches to semantics in natural language processing. In A. E. Kibrik (Ed**.**), *Proc. of Annual International Conference Dialogue*.

Rieder, A. (2002). A cognitive view of incidental vocabulary acquisition: from text meaning to word meaning. *VIEWS,* 11(1-2), 53-71.

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

Sheidlower, J. T. (1995). Principles for the inclusion of new words in college dictionaries. *Dictionaries,* 16, 33–44.

Simpson, J. (2007). Neologism: The long view. *Dictionaries,* 28:146–148.

Smarr, J., & Manning, C. D. (2002). Classifying unknown proper noun phrases without context. *Technical Report dbpubs/2002-46, NLP Group*, Stanford, CA: Stanford University.

Taylor, J. M., Raskin, V., & Hempelmann, C. F. (2010). On an automatic acquisition toolbox for ontologies and lexicons in Ontological Semantics, *International Conference on Artificial Intelligence*.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly,* 30, 415–433.

Wagner, R. K, Muse, A. E., and Tannenbaum, K. R. (Eds.) (2006). *Vocabulary Acquisition: Implications for Reading Comprehension.* New York, NY: Guilford Press.

Weischedel, R. M., Schwartz, R. M., Ramshaw, L. A., & Palmucci, J. (1993). Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics,* 19(2), 359-382.

Wilks, Y. (1978). Making preferences more active. *Artificial Intelligence,* 11(3), 197–223.

Wilks, Y., & Catizone, R. (2002). Lexical tuning. *Proc. of CICLING 2002* (pp. 106–125).

Wisniewski, E. J. (1996). Construal and similarity in conceptual combination. *Journal of Memory and Language,* 35, 434–453.

Wisniewski, E. J. (1997a). When concepts combine. *Psychonomic Bulletin and Review,* 4, 167–183.

Wisniewski, E. J. (1997b). Conceptual combination: Possibilities and esthetics. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 51–81). Washington, DC: APA.

Wisniewski, E. J. (2000). Similarity, alignment, and conceptual combination: Reply to Estes and Glucksberg. *Memory & Cognition,* 28, 35–38.