Search by Fuzzy Inference in a Children's Dictionary

St-Jacques, C.; Barrière, Caroline

Publisher's version / Version de l'éditeur:

Computer Assisted Language Learning, 18, 2005-07

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*Search by Fuzzy Inference in a Children's Dictionary **

St-Jacques, C., and Barrière, C.
July 2005

* published in Computer Assisted Language Learning. Volume 18, Number 3. July 2005. pp. 193-215. NRC 48513.

Canada

Computer Assisted Language Learning
Vol. 18, No. 3, July 2005, pp. 193–215

Routledge
Taylor & Francis Group

Search by Fuzzy Inference in a Children's Dictionary

Claude St-Jacques[a] and Caroline Barrière[b]
[a]*University of Ottawa, Canada;* [b]*National Research Council of Canada*

This research aims at promoting the usage of an online children's dictionary within a context of reading comprehension and vocabulary acquisition. Inspired by document retrieval approaches developed in the area of information retrieval (IR) research, we adapt a particular IR strategy, based on fuzzy logic, to a search in the electronic dictionary. From an unknown word, searched for by a learner, our proposed fuzzy inference process makes it possible to retrieve relevant lexical information from any entry of the dictionary. Furthermore, it organises this information in the form of semantic maps of an adaptable size surrounding the query word. Manual construction of such semantic maps are seen as being effective for helping learners in vocabulary acquisition and reading comprehension tasks. Our research leads to a capability for building them automatically. Using concrete examples, we provide details of the calculation for the construction of semantic maps as well as for the retrieval of information. We introduce a software component which could be integrated in a computer assisted language learning (CALL) environment to promote vocabulary acquisition in L1.

Children's dictionary use for vocabulary acquisition

Research in cognitive sciences (Clark, 1993, 2003) and applied linguistics (Singleton, 1989) enlightens the children's lexicon acquisition as an active dynamic process. In fact, Clark (2003) argues by quoting *l'Émile* of Rousseau (1995), that young children already have a simplified grammar at kinder age (5-years-old) that enables them to understand new words. However, Clark (1993) also argues that "any theory of acquisition must be a processing theory, a theory of changing performance" (Clark, 1993, p. 259). Furthermore, authors like Singleton (1989) view language acquisition

*Corresponding author. Interactive Language Technologies Group, National Research Council Canada, Gatineau, Québec, Canada K1A 0R6. Email: Caroline.Barriere@cnrc-nrc.gc.ca

ISSN 0958-8221 (print)/ISSN 1744-3210 (online)/05/030193–23

DOI: 10.1080/09588220500173377

as a life long process only interrupted by the senescence or injuries affecting intellectual capacity.

The early years of this process are particularly important and show a large vocabulary growth. Depending on how vocabulary is defined (base forms or not, receptive or productive, oral or print), somewhat of a concensus among researchers states that approximately 2,000 to 3,500 distinct words are learned yearly (Lehr, Osborn, & Hiebert, 2004). Yet, this tremendous learning achievement is partially explained by the fact that students lack a deep understanding for many of the words they know. Beck, McKeown, and Omanson (1987) have distinguished three levels of word knowledge—the unknown, acquainted, and established levels. Nevertheless, such acquisition rates obviously cannot come from direct teaching in the classroom, and students learn many of the words that make up their vocabularies by *incidental learning* (Anderson, 1996).

This relation between reading comprehension and vocabulary acquisition is acknowledged by researchers in cognitive science and first language acquisition. "One of the most persistent findings in reading research is that the extent of students' vocabulary knowledge relates strongly to their reading comprehension and overall academic success" (Lehr et al., 2004). For instance, Nagy, Herman, and Anderson (1985) argue that the learning of a new word is frequently related to the incidence of reading or repetitive listening. Results of several empirical studies (Chun & Plass, 1996; Hulstijn, Hollander, & Greidanus, 1996; Zimmerman, 1997) confirm that hypothesis.

Researchers closer to the problem of teaching vocabulary acquisition through reading propose a *scaffolding* model (Wood, Burner, & Ross 1976), a valuable instructional technique still in use today (Taylor, Pressley, & Pearson, 2002), in which the proportion of responsibility for reading experiences is gradually released from the hands of the teacher to the hands of the students. The gradual release is important to prepare students for their lifelong learning processes, as their reading performances will continue to evolve in part through vocabulary acquisition performed within their numerous reading situations in autonomous settings.

Graves, Juel, and Graves (2004, p. 238) mention three strategies included in scaffolding that help students to become independent word learners: using context cues, using word parts, and using the dictionary. The first two strategies help the learner infer the meaning of an unknown word. The third strategy, the use of the dictionary, could be used to confirm such an inference. Unfortunately, dictionary usage seems problematic for young students (Miller & Gildea, 1987). And even today, "a student's experience in working with dictionaries is often limited to instruction in alphabetical order, using guide words, and understanding pronunciation keys" (Graves et al., 2004). Among the possible guidelines given, learners can be told to look through all definitions and find the one that makes the most sense in the context in which the word is used.

This certainly opens a challenging dimension in the role that computational lexicography can play to promote the young learner's dictionary as a significant component of a computer assisted language learning (CALL) environment. Our

research moves in this direction. More specifically, our concern in the present paper is with the development of a software tool for dictionary exploration, useful in L1, especially in an autonomous setting.

This tool is developed based on a solid theoretical background coming from information retrieval research, and is mostly aimed at pushing the boundaries of the present-day static view of the dictionary. We explore the lexicographical conditions, which can allow a certain dynamicity in the use of an electronic dictionary to benefit a young learner in his vocabulary acquisition, and favor a change in performance in his reading abilities.

The static view comes from the structural partition of the paper-form of a dictionary into a macrostructure of alphabetically ordered lexical entries and a microstructure of definitions and contextual examples. For any specific unknown word that is searched for, this static view provides a single entry point within the dictionary, that is through the unique lexical entry of that word found in the macrostructure. This entry point leads to the retrieval of a unique corresponding microstructure which will often leave the users feeling they have received insufficient assistance (too little information) or feeling overwhelmed (too much information, e.g. with polysemous words) without having completely answered their contextual request.

Our research aims at helping users in their search for information by viewing the overall microstructrure of the dictionary as a corpus of text providing multiple contexts of word definitions and usages which can be selected and presented. An empirical study by Gipe (1980) showed that young children learn a new word more easily when it has been introduced to their knowledge base, not only by one, but by several contexts, with one of them giving the definite meaning. Furthermore, Pearson and Studt (1975) illustrated with one experiment that the frequency of a word and the richness of the context facilitate the identification of a word by a novice reader. Among the things emphasised by the National Reading Panel (2000), regarding the role of vocabulary in reading instruction, is that repetition, richness of context, and motivation may also add to the efficacy of incidental learning of vocabulary.

We suggest that through dictionaric inferences (St-Jacques & Barrière, 2004) within the microstructure of the dictionary, the latter can be transformed into a dynamic environment, adaptable to each reading situation in which unknown words are found. Before the detailed presentation of our approach, we will first look at two examples that explore the limitations of the presently held static view of the dictionary, in which each macrostructure entry is seen in isolation. We then suggest that the dynamicity of an electronic dictionary is a necessary condition for increasing its value in helping young learners.

In Need of More Flexible Dictionary Search Approaches

The first example we explore is a text, *Lucy and the chickens,* from TEA[1] (Texas Education Agency), dedicated to the development of reading skills of children in Grade 3 (8-years-old). Reading through the text, a young reader discovers that Lucy loves the animals on the family farm a lot, and that "she enjoyed collecting eggs from

the henhouse" . Let us suppose that the young reader does not know the meaning of "henhouse", and furthermore, that this word is not clarified elsewhere in the text. A good reader having a sufficiently developed vocabulary, background knowledge and reading skills will infer the meaning of the word from the context by thinking that eggs are collected in the farm building where the hens are living.

A search in a dictionary could validate the situational relevance of that inference, but unfortunately, given their limited coverage, children's dictionaries sometimes prove to be useless for a child. In this particular example, the American Heritage First Dictionary[2] (AHFD), a dictionary specifically aimed at children from 6- to 9-years-old, does not contain the word "henhouse" among its 1800 words. Although dictionary editors decide on their dictionary coverage based on frequency lists to include high frequency words, it is impossible for them to prejudge the whole context of usage and needs of a child.[3]

Given the absence of "henhouse" in the dictionary, the vocabulary acquisition scaffolding process (Graves et al., 2004) would provide the child with at least three other entry points. The first one, "egg", is specifically mentioned in the sentence (context-cue). The second one, "chicken", as part of his background knowledge, has certainly frequent association with egg[4] in this context. The third one, "hen", is a guess at the compositional nature of the word "henhouse" with "house" as a known lexical-semantic unit and "hen" remaining as unknown.

Table 1 shows the three entries for "egg", "chicken" and "hen" from the AHFD. The first two do not contain any cues about the name of the buildings found on a farm; only the third entry point "hen" is a good track as the young reader encounters the word "chicken" providing a link into his background knowledge.

With this example, we do not wish to show that the nomenclature of dictionaries is always defective, but on the contrary, that relevant information is more easily available if we break away from the traditional mode of searching only through their macrostructure. This is in accordance with Humblé (2001), who suggests that to exploit the computational capacity of the electronic dictionary is to adapt it to the users. In response to empirical studies that show that a user is often unable to find the information deeply buried in a dictionary and the proposal that the user should be

Table 1. AHFD entries for egg, chicken, hen

Entry	Definition
Egg	An egg is a smooth round shell with a baby animal inside of it
	Birds grow inside eggs until they are ready to hatch
	Many people eat eggs from chickens for breakfast
Chicken	A chicken is a kind of bird
	Chickens are raised for eggs and meat
	A hen is a bird
	Female chickens are hens
Hen	Hens lay eggs

trained to develop their abilities to use that tool correctly, Humblé explains that "the teaching of dictionary skills always sounds a little like trying to adapt the user to the washing machine instead of the opposite" (Humblé, 2001, p. 48).

Despite the optimism of Humblé (2001), who believes that the dictionary in an electronic format will have minimal limits on the contents of collocations, other researchers, such as Rossi (2000), state that the dictionary definitions remain synthesised and argue that the kind of definite meaning included in a dictionary cannot ensure the conditions for the re-employment of the word in other possible contexts by a young learner.

In the *Le petit Robert des enfants,* the creator, Josette Rey-Debove (1990), shows innovative efforts to counter Rossi's statement. She creates a dictionary in paper-form by using what she calls "phrastic definitions" that try to overcome the difficulty children face when trying to understand more formal or obscure definitions. This lexicographic technique is founded on two particular characteristics. First, an autonymic connotation of the lexical entry is given, i.e. the dictionary presents the entry as a sign and quotes it at the same time. Second, glossed examples are used to enlighten the definition of a word with its linguistic demonstration in different examples, each followed by an explanatory gloss.

Our concern is not primarily that the information is synthesised or insufficient in the dictionary, but that it is dispersed and therefore not easily accessible in order for the reader to build a good contextual picture of an unknown word. In fact, pictures, in a literal sense, are often used in children's dictionaries like AFHD to facilitate the understanding of new words, as well as to help a child establish a direct link with the sensorial world. Although pictures are useful for words which designate concrete objects, they do not provide an understanding of the interrelation of an object to other objects in the world.

To address both the information dispersion problem and the word interrelation understanding problem, we turn to the possibility given by computers for the automatic construction of semantic maps created directly from the information within the microstructure of electronic dictionaries.

Semantic Maps for Flexible Dictionary Search

Semantic maps are well known to language teachers and are recognised for their value in vocabulary acquisition. Semantic maps are an "effective means to expand a student's knowledge of words with which they are already familiar but which have multiple meanings or are part of an extensive network of related words" (Johnson & Pearson, 1984; Pittelman, Heimlich, Berglund, & French, 1991). Moreover, "the use of semantic mapping has been empirically demonstrated to facilitate student success in vocabulary development" (Foil & Alber, 2002).

Not many reading strategies can be used, as graphic strategies are, at the preview stage before reading, during the reading process itself, and at the stage after reading (Dowhower, 1999). As a pre-reading activity they can activate background knowledge of the students surrounding manually selected words that will be in the text. As a

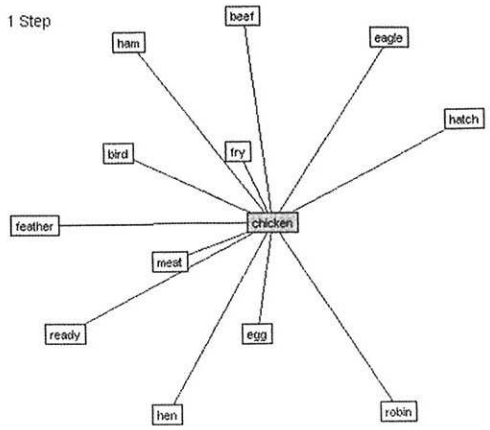
post-reading activity they can show the relation between the different elements in the text. These two often interactive teacher–learner activities are not possible in an autonomous setting. Here we suggest to view reading as a problem-solving activity and to use automatically generated semantic mapping during reading to help the student resolve an unknown word problem which prevents him from understanding a sentence.

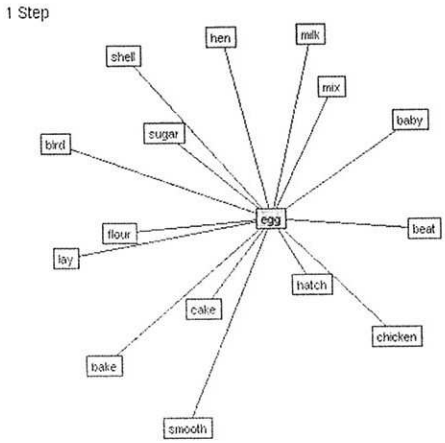
Within a computer environment, where it is easy to visualise and navigate through semantic maps, the learner would never consider a word through a fixed definition, but instead, throughout the situational relevance of his inference with some semantic fields.

By automatically generating a semantic map around an unknown word pointed out by a young learner during his reading, this would help him not only to understand that word, but also to establish its position in a network of paradigmatic relations, such as synonymy, hyperonymy and meronymy, which abstracts the understanding of the word away from the context of usage. The semantic maps can help structure the acquisition of the lexicon by facilitating the user's investigations into categories of objects, allowing the user to reinvest their knowledge about the categories of objects already known in order to identify the new word. Let us return to our previous example with "henhouse" as an unknown word. Figures 1(a), 1(b), and 1(c) show three semantic maps automatically generated by our approach (explained later as a pseudo-thesaurus construction), which respectively find the fuzzy associations around the words "chicken", "egg" and "hen", previously mentioned as three possible entry points in the AFHD.

We have gained here on two fronts. First, compared to a macrostructure access (as shown in Table 1), providing a single partially successful entry point, all three inference routes "egg", "chicken" and "hen" can now show the three-way relation between the three words. Second, semantic maps contain many known words to a child, and therefore can easily evoke some lexical contexts already experienced in the real world to help him couple the meaning of a new word with his background knowledge.

Semantic maps are quite interesting in themselves as they give the user a quick view of a semantic field; and they can further be used as an intermediate step for querying into the dictionary. We can gather the words found in the semantic map into a bag of words used together to launch a query into all of the microstructures of the dictionary in order to retrieve the most relevant information pertaining to the understanding of this bag of words. The microstructure retrieved will contain examples and definite explanations of words to help the learner further organise his mental lexicon. To explain this idea further, we take a second text, *The farm animals*,[5] in which the author uses a picture for each sentence to facilitate the understanding of his story for a young child. For example, the first sentence says, "The animals that live in a farm all make special sounds" and just beside the text, the author presents a picture of a barn. Let us suppose that from this picture, a child infers that "a farm is a building where the animals are living". By looking in the AFHD, this inference can be invalidated but not rectified, because in the "farm" entry the microstructure explains that "A farm is





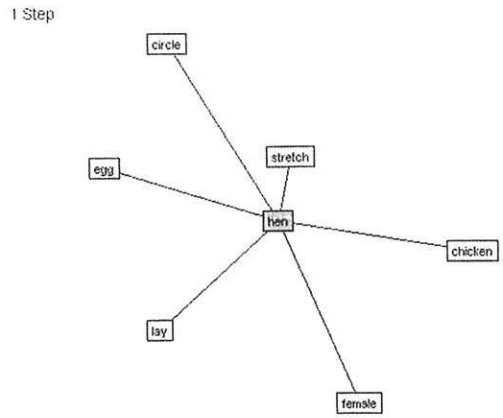


Figure 1a – Semantic map for "chicken" built from the AHFD; 1b – Semantic map for "egg" built from the AHFD; 1c – Semantic map for "hen" built from the AHFD

an area of land. People grow food and raise animals on farms". Therefore, the false hyperonymic link hypothesis (farm-building) is removed, but no other link is suggested, such as the correct one (barn-building).

Let us see how semantic maps can help, and construct one around "farm", as shown in Figure 2. From the association strengths shown, we set a threshold at 0.1 to select the words closest to farm. The use of a threshold, called α-cut, will be explained later in the artice. The derived bag of words contains {farm(1.0), barn(0.166), raise(0.12), horse(0.117), cattle(0.111)} and is used to search within the microstructure for relevant information. The exact search algorithm will also be explained later on, and the present steps will be referred to as query expansion and segment (definition) retrieval. This search allows for the retrieval of relevant definitions in the microstructure, as shown in Table 2, in a decreasing order of relevance (we only put the top 6 for the sake of conciseness).

It is interesting to notice that for polysemous words such as "country" (Table 2, last row), the retrieval will highlight the appropriate sense. Here, "country", Sense 2, is in opposition to Sense 1, which defines it as a geographical area uniting people with the same laws. This is an advantage of viewing all senses in the microstructure as separate entries; it provides a filtering mechanism for polysemous words.

The macrostructure (entries) corresponding to the retrieved definitions shown in Table 2 can also be graphically organised to illustrate their distance to the central word, as shown in Figure 3.

According to Clark (1993), two pragmatic principles of conventionality (e.g. greenhouse versus henhouse) and contrast (e.g. chicken versus hen) make the mapping of a new word feasible for children, and in particular, the contrast principle helps children to distinguish words in their mental dictionaries. However this

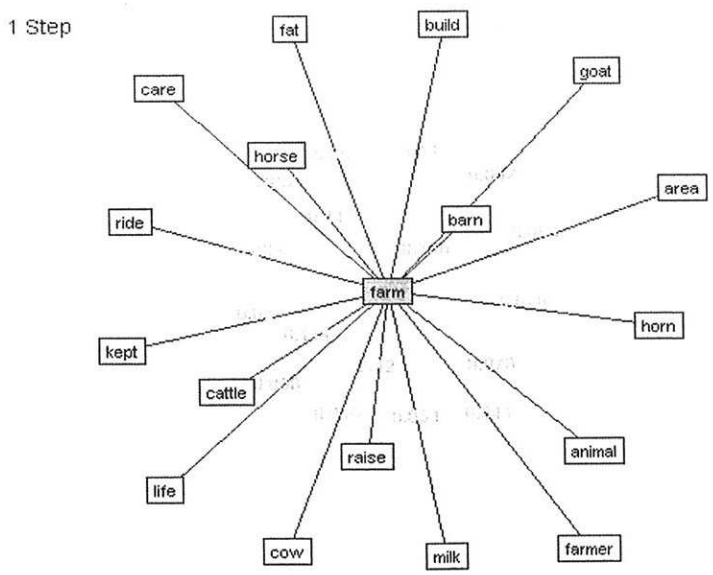


Figure 2. – Semantic map around "farm" as built with the AHFD

Table 2. Definitions relevant to the understanding of the word farm

Entry	Rel.	Definition
Barn.1.n.	0.5	A barn is a kind of building on a farm The barn is where the farm animals stay at night Farm machines and food for the animals are kept in the barns, too
Farm.1.n.	0.33	A farm is an area of land People grow food and raise animals on farms
Horse.1.n.	0.167	A horse is a large animal with long legs Horses live on farms People like to ride horses
Stable.1.n.	0.167	A stable is a building on a farm where horses and other animals are kept
Farmer.1.n.	0.167	A farmer is someone who works on a farm Farmers start to work early in the morning
Country.2.n.	0.167	The country is an area away from a city There are forests, fields, and farms in the country

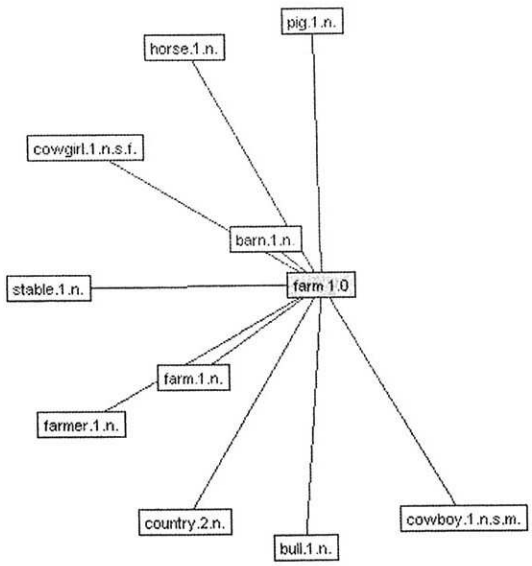


Figure 3 – Graphical view of relevant definitions

reasoning is based on the differentiation between two words sharing a certain part of the same semantic field. Our retrieval approach is quite interesting in this sense, since it first establishes the semantic field (semantic map), and then can present more detailed information within that field (relevant definitions), allowing a learner to form the necessary contrasts to structure his mental lexicon.

We will come back to this two-step process later on, as we explain the formal computation of (i) pseudo-thesaurus construction; and (ii) query expansion and retrieval, both within a lexicographic context. However, first we will discuss our inspiration from the area of information retrieval to describe the theoretical approach of our logical reasoning.

An Inferential Approach to the Dictionary Access

Information retrieval (IR) on the Internet presents a very similar situation to the macrostructure-microstructure problem described earlier for dictionary access. In IR, entry points are the sets of key words available to the user, and the web pages of interest are among all possible web pages on the Internet. Here again, the user can often be dissatisfied by Boolean search engines (allowing AND/OR combinations of keywords) either presenting him with an excessively reduced set of Web pages, or submerging him with a set of irrelevant hits too large to explore. A better access to the targeted information depends on the background knowledge of the user and his capacity to expand or refine his query by proposing other keywords that are more appropriate for his request. IR researchers try to withdraw the burden of query expansion/refinement from the hands of the user by exploring approaches which allow for an automatic inference of a set of terms acting as an intermediary level between the vocabulary used in documents and the user's original query keywords.

According to Turtle and Croft (1991), the IR inferential approach has existed since Cooper's work on relevance logic. Cooper's strategy (1971) consists of deducing a relevant statement from a question, on the condition that it belongs to the set of statements of possible logical answers to the interrogative premise. Subsequently, the concept of relevance was widened by Wilson (1973) in inductive contexts. In such cases where the information desired by the user of a system depends on a particular context and his state of knowledge, Wilson (1973, p. 459) suggests that the relevance of an argument can be induced from the support of the conclusion to the premise, without being such an implication in the sense of the classical logic.

In a previous study, we investigated the inferential activity surrounding the use of a dictionary (St-Jacques & Barrière, 2004) and our theoretical account likely converges with the concerns of Wilson (1973) regarding the logical inference in IR and the situational relevance. That paper presented possible logical reasonings allowing access to significant information with respect to a particular context of dictionary use. The activity of a dictionary search, called "dictionaric inference", was formally represented by (*Ax R By*), as a logical relation *R* linking a premise *Ax* (macrostructure) with a conclusion *By* (microstructure).

This inferential process in IR is seen in recent work by Nie (2003). According to his strategy, a query expansion gives rise to an inferential process based on the entailment of a premise towards the knowledge available within a thesaurus automatically built from the textual corpus. In this research, we suggest a similar inferential approach by using the dictionary as a textual corpus for the construction of a thesaurus to upgrade the user query words to the vocabulary of the lexicographical resource.

Taking advantage of the inferential approach, we suggest a strategy of dictionary searching by way of a progressive expansion of a query. Such an expansion relieves the user from constantly looking through the macrostructure. Rather, the expansion is guided by a fuzzy inference which allows for an automatic presentation of new microstructures in an order of decreasing similarity to the original query.

Our query expansion strategy rests on an automatic construction of a fuzzy pseudo-thesaurus (Miyamoto, 1990) generated from the microstructure of a dictionary. A pseudo-thesaurus is a broad matrix giving the strength of association between all pairs of lexical units of a textual corpus. According to Kim and Choi (1999), such association values can be carried out by the measurement of similarity between the pairs of terms based on their collocations. Two strategies prevail as a basis for this calculation of similarity: the fuzzy thesaurus approach from Miyamoto (1990) and the similarity thesaurus approach from Qiu and Frei (1993) and Vechtomova *et al.* (2003). The first one borrows its methods from the theoretical framework of fuzzy logic, more particularly Zadeh's fuzzy set theory (1965), while the second one leans on statistics and collocation probabilities.

We chose the method of fuzzy thesaurus for two main reasons. First, it supports three relational types of association of terms often used in IR: broader-term, narrower-term and related-term. Therefore, it provides the flexibility to explore these different types of associations which have been considered important by researchers in IR, such as Kim & Choi (1999), Qiu & Frei (1993) and Nie (2003). These researchers have suggested some query expansion strategies to assign weights to terms based on both their types and degrees of association with the query terms, as found in a textual corpus. We have not yet ventured in this area, and instead have focused solely on the related-term association following Greenberg (2001) who showed that this type of association tends to be more important than others by significantly increasing the recall of documents.

Second, the fuzzy thesaurus approach not only ensures an expansion starting from one word, but it also allows queries with multiple terms. Since there are so many polysemous words, this is a very interesting and valuable feature allowing for the addition of domain indicators to orient the search in a more promising direction. We will develop this idea later on when we explain in detail the steps and calculations involved in pseudo-thesaurus construction and dictionary searches.

Fuzzy Pseudo-Thesaurus and Progressive IR in the AHFD

Let us take a concrete example of the automatic generation of a pseudo-thesaurus and trace, step-by-step, the stages of a progressive search in the electronic dictionary entitled, *American Heritage First Dictionary* (AHFD) that was introduced earlier.

Before the computation of similarity between terms, we must carry out a set of pre-processing steps, first to find the canonical forms of words using some stemming rules, second to apply a stop word filter for frequently occurring or insignificant words and, third to eliminate the hapax, which are terms occurring only once in the corpus. Then, we proceed to the dictionary segmentation using one entry as a

Table 3. Segments of text for the entry "block" in the AHFD

Block1	A *block* is a piece of *wood* or *plastic* or *stone*
	It has *straight sides* and is *usually shaped* like a *square* or a *rectangle*
	Children play with *blocks*
Block2	A *block* is an *area* of a *city*
	It has *four streets* for *sides*
	Tom and *Jim walked around* the *block*

segment of text, except for polysemous terms, where each sense is unbundled into a distinctive segment. Table 3 shows the significant words in bold for the lexical entry "block", segmented in its two senses. The notion of *significance* is based on filtering, proportional to the corpus size, and in this particular case is set to keep words occurring between 2 and 100 times.

In choosing this pattern of segmentation as well as the method of computation of a pseudo-thesaurus, described below, our approach of similarity measure between terms becomes typical of a short span collocation. According to Vechtomova, Robertson, and Jones (2003), collocation can be of a short or a long span. The first type of collocation is found in short segments of text like a sentence or paragraph and expresses grammatical restriction or a lexical limitation on the use of terms. The second type of collocation arises instead from the semantic existence of a relation between words according to a long span like a whole document. A similarity measure based on long span segments acquires its precision as the number of documents and the dimension of the corpus increase. For the present research, indeed, a similarity measure based on short span segments is better adapted to the structure of dictionaries and the limited size of the corpus it generates.

Subsequent to the segmentation, each sense of a lexical entry is represented by a vector of significant terms in a vector space. Each term in a vector is assigned a weight as calculated from its frequency of occurrence within that segment. The structure of a lexical entry in AHFD is such that the first sentence initially gives the definition, and then one or two sentences follow to display an explanation or an exemplification of a word sense. Taking this entry structure into account, we wish to emphasise the value of the defining sentence, and choose to modulate the weight of its terms by doubling their values for occurrence in that sentence. From the definitions in Table 3, we calculate, as shown in Table 4, weights for the terms[6] in vectors for block1 and block2 respectively.

This table gives a representative sample of the terms' weights for lexical entries of the AHFD written in a matrix form, which we will refer to as an *entry-term* matrix.

The generation of the pseudo-thesaurus is based on the evaluation of the fuzzy association by linking together two terms in the AFHD. Assume that if $T = \{t_1, t_2, \ldots t_n\}$ is the set of significant terms and $E = \{e_1, e_2, \ldots e_m\}$ is the set of entries of the AHFD and let $\mu_{e_k}(t_i)$ be the function giving the membership degree of a term t_i to an entry e_k, then the degree of similarity between two terms t_i, t_j (T is given by the relation of similarity $R(t_i, t_j)$. According to Miyamoto (1990), we use the Jaccard

Table 4. The weight of significant words for the two senses of the entry "block"

	Block (t_1)	Wood (t_2)	Plastic (t_3)	Stone (t_4)	Side (t_5)	Shape (t_6)	Square (t_7)	Rectangle (t_8)	Children (t_9)	Area (t_{10})	City (t_{11})	Four (t_{12})	Street (t_{13})	Walk (t_{14})	Around (t_{15})
Block1 (e_1)	3	2	2	2	1	1	1	1	1	0	0	0	0	0	0
Block2 (e_2)	3	0	0	0	1	0	0	0	0	2	2	1	1	1	1

coefficient to compute the degree of association of related terms (RT), namely given by the following equation:

$$R(t_i, t_j) = \frac{\sum_E \min\left[\mu_{e_k}(t_i), \mu_{e_k}(t_j)\right]}{\sum_E \max\left[\mu_{e_k}(t_i), \mu_{e_k}(t_j)\right]}$$

Suppose now that we generate a pseudo-thesaurus from Table 4. Approximating a term's degree of membership[7] to an entry by its number of occurrences, we have for instance $\mu_{e_1}(t_2) = 2/N$, i.e. the membership degree of the term "wood" to the entry "block1".

Let us calculate, for example, the fuzzy similarity between the terms "wood" and "plastic" for the entries e_1 and e_2 in Table 4. So, if we have e_1: $t_2(2)$ $t_3(2)$ and e_2: $t_2(0)$ $t_3(0)$, then the fuzzy similarity is $R(t_2, t_3) = \frac{\min(2,2)+\min(0,0)}{\max(2,2)+\max(0,0)} = 1$. The result of this calculation for all of the pairs made of t_1,t_2,t_3,t_4 is shown in Table 5 as an extract of a pseudo-thesaurus P.

The actual pseudo-thesaurus, constructed on the entire AHFD, is a matrix similar to the one presented in Table 5, but of much larger dimensions to establish the similarity between all pairs of words present in the dictionary entries. From this large pseudo-thesaurus (matrix), we can extract a row (or a column since the matrix is symmetric) which represents the similarity between a particular searched word and all other words. This information is then graphically represented via a semantic map, as we presented earlier in Figures 1 and 2, in which the searched word is placed at the center of the map, and link lengths encode the similarities to other words.

However the pseudo-thesaurus can be used in a more refined operation called query expansion which provides a way to combine multiple rows in the matrix to incorporate the influence of context in the organisation of the semantic map. This is quite useful for polysemous words for which we would like to orient the map toward a particular sense.

We present another example chosen to emphasise the importance of context in query expansion and retrieval of significant information, and the flexibility of the fuzzy approach in allowing queries to be sets (groups of words) rather than individual words.

Table 5. Extract of a fuzzy thesaurus

		Block	Wood	Plastic	Street
	Block	1	33	33	14
$P=$	Wood	33	1	1	0
	Plastic	33	1	1	0
	Street	14	0	0	1

If we consider that a fuzzy set Q represents a user's query with terms t_i denoted by the membership function $\mu_q(t_i)$ and that the original query is augmented by the fuzzy thesaurus $P(t_i,t_j)$, then the expanded query A is obtained by the standard composition $A(t_j) = \max_{\mu(t_i,t_j)\in T} \min[\mu_Q(t_i), \mu_P(t_i,t_j)]$ between these relations. The standard composition of a fuzzy relation, like $A = Q^{\circ}P$, is usually described as an operation of *max-min* composition between fuzzy relations.[8]

Suppose a query Q of a user is made with the terms "block" and "toy". If we denote the membership degree of these terms[9] by the vector $Q = \begin{bmatrix} t_1 & t_2 \\ 1 & 1 \end{bmatrix}$ and the relevant part of the fuzzy pseudo-thesaurus is given by the matrix $P = \begin{matrix} t_1 \\ t_2 \end{matrix} \begin{bmatrix} t_1 & t_2 & t_3 & t_4 \\ 1 & 1 & .09 & .09 \\ 1 & 1 & 0 & 0 \end{bmatrix}$ where t_1,t_2,t_3,t_4 are respectively the terms "block", "toy", "igloo" and "stone", then we automatically obtain the following vector of values for the augmented query: $A = \begin{bmatrix} t_1 & t_2 & t_3 & t_4 \\ 1 & 1 & .09 & .09 \end{bmatrix}$. The vector A is obtained by applying the values of the query on each column of the matrix P as follows: $\mu_A(t_1) = \max$ [min(1,1), min(1,1] = 1, $\mu_A(t_2) = \max$[min(1,1), min(1,1)] = 1, $\mu_A(t_3) = \max$[min(1.09), min(1,0)] = .09, $\mu_A(t_4) = \max$[min(1.09), min(1,0)] = .09.

Such an augmented query can then be projected onto the entry-term matrix to select its relevant part. Table 4, once normalised[10] could be seen as an extract of such an entry-term matrix on which to project the augmented query. If we call B the entry-term matrix and we apply the augmented query A on it, as a fuzzy operation of composition of relations once again, then we obtain the result S for a search in the AHFD by applying the *max-min* composition $S = A^{\circ}B$.[11]

Now that we have introduced the formalism for a flexible query, let us reconsider our aim to progressively enlarge the access for a user to the dictionary's vocabulary by showing how we can gradually increase the user's original query. The process of computation governing the automatic generation of a pseudo-thesaurus lends itself to the idea of progressive similarity between words. Moreover, the query expansion requires a limitation in the number of terms to be added. The fuzzy set theory enables us to handle that limitation by using the concept of α-cut. According to Bouchon-Meunier (1994), an α-cut produces an approximate subset of the original fuzzy set by limiting its objects to the ones having a membership degree higher than the level of the parameter α. Thus, the original query of a user $Q = \begin{bmatrix} t_1 & t_2 \\ 1 & 1 \end{bmatrix}$, made of the terms "block" and "toy", can be gradually expanded following one of the subsets $A_\alpha$expressed in Table 6 according to the threshold value of $\alpha = \{.08, .06, .04\}$. Note that to be concise, we are not repeating the terms of the preceeding subsets A_α[12] in Table 6, so the lists should be seen as cumulative.

In the same way, we obtain the gradually expanded query A_α shown in Table 7 for a request made with the terms "block" and "neighbourhood".

Relaxing the threshold value of an α-cut includes more and more terms in the expanded query. Although this strategy introduces more terms for a dictionary search, an excessive expansion has the effect to include some related terms with small association values to the original query.

In a learning context, we can opt for a conservative approach in which fewer highly related terms are included in the expanded query set. Nevertheless, to satisfy curious readers wanting to see more words, our software module allows semantic maps of

Table 6. Expanded queries A_α to an original request using "block" and "toy"

α	A_α^1 (block, toy)										
08	Block (1.0)	Toy (1.0)	Side (.08)								
06	Sold (.077)	Jim (.077)	Mr (.077)	Box (.073)	Pile (.071)	Fit (.067)	Neighbour (.067)	Square (.067)	Half (.065)	String (.064)	Baby (.061)
04	Female (.059)	Stone (.059)	Tom (.056)	Straight (.05)	Up (.048)	Street (.047)	Put (.044)	Small (.041)	Store (.040)		

Table 7. Expanded queries A_α to an original request using "block" and "neighborhood"

α	A_α^2 (block, neighbourhood)									
08	Block (1.0)	Neighbourhood (1.0)	Costume (0.1)	October (0.1)	Collect (0.09)	Holiday (0.09)	Side (0.08)			
06	Jim (.077)	Mr. (.077)	Neighbour (.077)	Candy (.077)	Kim (.071)	Power (.071)	Quietly (.071)	Rectangle (.067)	Square (.067)	Electricity (.067)
04	Female (.059)	Stone (.059)	Storm (.059)	Tom (.056)	Straight (.05)	Almost (.05)	Street (.047)	Everyone (.043)	Little (.043)	

different sizes to be viewed. Figures 4(a), and (b) show the semantic maps for $\alpha = 0.4$ and $\alpha = 0.8$ as taken from Table 7 for the query combination "block – neighborhood".

Earlier, we emphasised that semantic maps are interesting in themselves, and here we further add that they become especially interesting if they can be contextually adapted. Furthermore, their words are used to find relevant segments from the electronic dictionary. Once again, it is easy to produce a progressive presentation of the results because each segment extracted from the AHFD receives a grade of membership to the expanded query and these values can be used for ordering the output. Again, we can apply an α-cut on these values of membership degrees to limit the number of results shown to the learner. Table 8 and Table 9 provide the results of the computation of a search in the AHFD made by applying the *max-min* composition $S = A^1_{\alpha=.08} \circ B$ and $S = A^2_{\alpha=.08} \circ B$ respectively corresponding to the expanded queries given in the first rows of Table 6 and Table 7.

In both tables, the first column indicates which lexical entries correspond to the relevant segments retained by the query among the whole microstructure of the AHFD. The second column shows the degree of relevance for each segment. The third column presents the segments of the dictionary text giving the relevant information concerning the user's query. To emphasise the difference between the two senses of "block", we only include the exclusive microstructures for each in the tables.

For the sake of completeness, we include the relevant information retrieved for both block1 and block2 in Table 10.

Table 8. Unique relevant segments to an expanded query originally expressed by "block" and "toy"

Macrostructure	Relevance	Microstructure
Store	0.333	A store is a place where things are sold
		Shoes are sold in shoe stores
		You can buy toys in a toy store
		Some stores are very big and have all kinds of things to sell
Sled	0.167	A sled is a toy
		People ride on sleds over the snow
		Sleds are made of wood, metal, or plastic
Gather	0.167	To gather means to come together or put together
		People often gather to listen to music
		Peter gathered up all his toys and put them in one box
Collect	0.167	To collect means to put things together in a group
		We collected our toys into a pile
Divide	0.167	To divide means to change one big thing into two or more smaller things
		Mary divided the apple into two halves
		Joel divides his toys in three piles

Table 9. Unique relevant segments to an expanded query originally expressed by "block" and "neighborhood"

Macrostructure	Relevance	Microstructure
Neighbourhood	0.333	A neighbourhood is an area where neighbours live
		Kim knows almost everyone in his neighbourhood
Quiet	0.167	To be quiet means to make very little sound
		Our neighbourhood is very quiet at night
Power	0.167	Power means electricity
		Our neighbourhood had no power after the big storm
Halloween	0.167	Halloween is a holiday
		It comes on the last day of October
		People wear costumes on Halloween
		Then they may go out to collect candy in their neighbourhood
Clown	0.1	A clown is a person who makes people laugh
		Clowns wear funny costumes
		They play tricks
		Many clowns work in a circus

The generation of the pseudo-thesaurus on the AHFD allows for the presentation of semantic maps to a young reader as shown in Figures 4 (a) and (b). As mentioned, this will help his understanding of new words through their association with other words. Furthermore, the pseudo-thesaurus is at the core of the query expansion process to access specific relevant entries such as those shown in Tables 8 and 9. This will help the learner find relevant information to his query and organise his mental lexicon.

Discussion

It is interesting to note that "Although the National Reading Panel (2000) cites computer technology as a promising technique for increasing vocabulary, little research yet exists to provide direction for computer-related instruction" (Lehr et al., 2004). Wood (2001) suggests that the potential lies in certain capabilities not found in printed materials, such as game-like formats, animations and online dictionaries and reference materials. Even earlier, Baker et al. (1995) hinted at the potential of computer applications: "helping students understand the relation between words through semantic maps seems especially suited to computer software" . The tool, we suggest, is certainly a step in this direction, promoting dictionary usage via the use of automatically constructed semantic maps in an autonomous setting.

To allow lifelong learning of vocabulary through reading and reciprocically help in reading comprehension via vocabulary acquisition, it is important to help learners in autonomous settings easily find meaning of unknown words, or mostly confirm meanings they infer from contextual and word-part clues. Since the usefullness of semantic maps has been established in the literature, they provide a natural path into

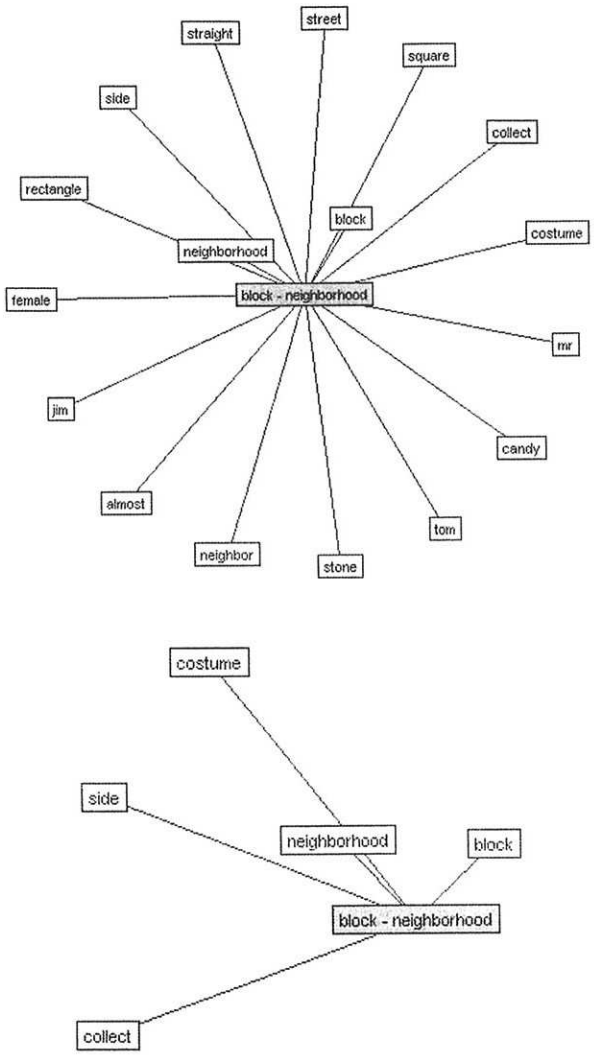


Figure 4a) Semantic map from "block – neighbourhood" with $\alpha = 0.4$ 4b) Semantic map from "block – neighbourhood" with $\alpha = 0.8$

an efficient usage of dictionary. We propose a tool which can automatically, in real time, generate a semantic map for the words contained in the dictionary and which can take into account the contextual constraint from the occurrence context.

We were concerned with how dispersed the information was in the dictionary. We now make the information available through our pseudo-thesaurus construction and visualisation. We were also concerned with helping a learner find appropriate information about an unknown word during a reading task. With the contextual searches allowed within our tool, via a multi-words query, we address this concern. Finally, we were concerned with helping the students in their mental lexicon

Table 10. Relevant segments for both queries including "block"

Macrostructure	Relevance	Microstructure
Block2	0.333	A block is an area of a city It has four streets for sides Tom and Jim walked around the block
Block1	0.333	A block is a piece of wood or plastic or stone It has straight sides and is usually shaped like a square or a rectangle Children play with blocks
Igloo	0.167	An igloo is a kind of house It is made of blocks of snow People who live in cold places where there are no trees sometimes build igloos
She	0.167	She means a female person Mrs Perez lives on our block She is our neighbour
Toy	0.167	A toy is an object that people play with Wood blocks, dolls, and kites are toys

organisation. Our tool allows for the presentation of related-term associations via semantic maps and also favours access to specific information through relevant definitions retrieved.

We certainly wish to push our study further to better integrate the dictionary search into an online reading environment, allowing for automatic identification of context words to bias the search, and help the learner navigate through the dictionary. This would also provide an appropriate setting for further evaluation of the impact of using semantic maps for vocabulary acquisition. The next step in our semantic mapping process is to make an attempt at clustering the words which are part of the semantic map as is often suggested to language teachers as an exercise to do with students. For example, in Foil and Alber (2002) a semantic map around the word "serpent" would include groups related to "What is it?" (animal, reptile, snake), "What is it like?" (long, scaly, legless, slithery, scary) and "What are some examples?" (cobra, python, king, copperhead, cottonmouth). This is a challenging task for future research as grouping criteria are diverse and complex (Lehr et al., 2004). Lastly, we will pursue the possibility of adapting our process to other dictionaries.

Note

1. See www.tea.state.tx.us/student.assessment/resources/online/2003/grade3/read.htm.
2. Copyright obtained from Houghton Mifflin for an electronic reproduction of the *American Heritage First Dictionary*, originally in paper form, to be used for research purposes.
3. Hayes and Ahrens (1988) they identify 30.9 rare words per 1,000 in a children's book. In comparison, newspapers have 68.3, and prime-time adult shows 22.7.

4. In the *Edinburgh Association Thesaurus* (www.eat.rl.ac.uk), empirical data is gathered about word associations. When egg is the trigger word, top associations are with yolk, hen, cup, bacon and chicken.
5. You can find this story at www.magickeys.com/books/farm/index.html.
6. In the example, we voluntary omit the significant words "straight", "usually", "Tom" and "Jim" from Table 1 only to limit the size of the table.
   Membership degree to a fuzzy set is always defined by the interval of values $\mu_e(t) \rightarrow [0,1]$. So, we should normalise the frequencies of occurrences by a large positive number N, but according to Miyamoto, this parameter disappears in the calculation of the fuzzy similarity. For the mathematical details of the automatic generation of a pseudo-thesaurus, we refer the reader to Miyamoto (1990).
7. We refer the reader to Klir and Yuan (1995) for the mathematical explanation of the composition of fuzzy relations.
8. We give Boolean values to the terms in the original query; otherwise we would have to find a way to assign a weight to them with respect to the user's expectation or certainty. This is outside the scope of the type of search assumed here by a young reader in a dictionary
9. In the program, we normalise the weight of terms by the highest frequency of a word found across every vector of the AHFD. The max–min composition can be used as the operator of what St-Jacques and Barrière (2004) called the "dictionaric inference" (*Ax R By*) when we infer from a dictionary $A(t_j)^{\circ}B(t_i,t_j)$.
   To be precise, we should say that the α-cut is applied to the values of $P(t_i,t_j)$ before obtaining A by *max–min* composition: $A = Q^{\circ}P$ except that the *min* value is always from the pseudo-thesaurus because we give a Boolean value (1) to the query's words. By extension, we are saying here that α-cut is also the limitation for the weight of the terms in an expanded query A_{α}.

References

Anderson, R. C. (1996) Research foundations to support wide reading. In V. Greaney (Ed.), *Promoting reading in developing countries*. Newark, DE: International Reading Association, 55–77.

Baker, S. K., Simmons, D. C., Kameenui, E. J. (1995) Vocabulary acquisition: Curricular and instructional implications for diverse learners, Technical report, No. 14, National Center to Improve the Tools of Educators (NCITE). Retrieved from http://idea.uoregon.edu: 16080/~ncite/documents/techrep/tech14/html

Beck, I. L., McKeown, M. G., & Omanson, R. C. (1987). The effects and uses of diverse vocabulary instructional techniques. In M. G. McKeown & M. E. Curtis (Eds), *The nature of vocabulary acquisition*. Hillsdale, NJ: Erlbaum, 147–163.

Bouchon-Meunier, B. (1994). *La logique floue*. Paris: PUF.

Clark, E. (1993). *The lexicon in acquisition*. Cambridge: Cambridge University Press.

Clark, E. (2003). *First language acquisition*. Cambridge: Cambridge University Press.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval, 7*, 19–37.

Chun, D., & Plass, J. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal, 80*, 183–198.

Dowhower, S. L. (1999) Supporting a strategic stance in the classroom: A comprehension framework for helping teachers help students to be strategic. *The Reading Teacher, 52*, 672–688.

Foil, C. R., & Alber S. R. (2002) Fun and effective ways to build your students' vocabulary. *Intervention in School and Clinic, 37*(3), 131–139.

Gipe, J. P. (1980). Use of a relevant context helps kids learn new word meanings. *The Reading Teacher, 33*(4), 398–402.

Graves, M. F., Juel, C., & Graves, B. B. (2004). *Teaching reading in the twenty-first century*, Allyn and Bacon Publishers, Boston.

Greenberg, J. (2001). Automatic query expansion via lexical–semantic relationships. *Journal of the American Society for Information Science and Technology, 52*(5), 402–415.

Hayes, D. P., & Ahrens, M. (1988). Vocabulary simplification for children: A special case of "motherese", *Journal of Child Language, 15*, 395–410.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words, *The Modern Language Journal, 80*, 327–339.

Humblé, P. (2001). *Dictionaries and Language Learners*. Frankfurt am Main: HAAG.

Johnson, D. D., & Pearson, P. D. (1984). *Teaching reading vocabulary*. New York: Holt, Rinehart & Winston.

Kim, M. C., & Choi, K. S. (1999). A comparison of collocation-based similarity measures in query expansion. *Information Processing and Management, 35*, 19–30.

Klir, G. J., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic: Theory and applications*. Upper Saddle, NJ: Prentice Hall.

Lehr, F., Osborn, J., & Hiebert, E. H. (2004). A focus on vocabulary, research-based practices in early reading series, Regional Educational Laboratory at Pacific Resources for Education and Learning Publishers, 26 pages. Retrieved from http://www.prel.org/products/re-/ES0419.htm/

Miller, G. A., & Gildea, P. M. (1987). How children learn words. *Scientific American, 257*(3), 94–99.

Miyamoto, S. (1990). *Fuzzy sets in information retrieval and cluster analysis*. Dordrecht: Kluwer Academic Publishers.

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

National Reading Panel. (2000). *Teaching children to read: An evidence based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington DC: National Institute of Child Health and Human Development.

Nie, J. Y. (2003). Query expansion and query translation as logical inference. *Journal of the American Society for Information Science and Technology, 54*(4), 335–346.

Pearson, P. D., & Studt, A. (1975). Effects of word frequency and contextual richness on children's word identification abilities. *Journal of Educational Psychology, 67*(1), 89–95.

Pittelman, S. D., Heimlich, J. E., Berglund, R. L., & French, M. P. (1991). *Semantic feature analysis: Classroom applications*. Newark, DE: International Reading Association.

Qiu, Y., & Frei, H. P. (1993). *Concept based query expansion*. Paper presented at the ACM SIGIR International Conference on Research and Development in Informational Retrieval, Pittsburgh.

Rey-Debove, J. (1990). *Le petit Robert des enfants*. Paris: Dictionnaires Le Robert.

Rossi, M. (2000). Autonymie et monstration du signe dans les dictionnaires pour enfants. In J. Authier-Revuz, S. Branca-Rosoff, M. Doury, G. Petiot, & S. Reboul-Touré (Eds.), *Proceedings of "Le fait autonymique dans les langues et les discours", Syled (Systèmes, langues, énonciation et discursivité)*, Université de la Sorbonne Nouvelle (Paris 3), 5–7 October 2000.

Rousseau, J. J. (1995). *L'Émile ou de l'education*. Paris: Gallimard.

Singleton, D. M. (1989). *Language acquisition: The age factor*. Clevedon: Multilingual Matters.

St-Jacques, C., & Barrière, C. (2004). L'inférence dictionnairique: de la créativité poétique à celle du raisonnement flou. *Cahiers de lexicologie, 85*, 129–155.

Taylor, B. M., Pressley, M., & Pearson, P. D. (2002). Research-supported characteristics of teachers and schools that promote reading achievement. In B. M. Taylor, & P. D. Pearson (Eds.), *Teaching reading: Effective schools, accomplished teachers*. Mahwah, NJ: Erlbaum.

Turtle, H., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transaction on Information Systems, 9*(3), 187–222.

Vechtomova, O., Robertson, S., & Jones, S. (2003). Query expansion with long-span collocates. *Information Retrieval, 6,* 251–273.

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval, 9,* 457–471.

Wood, D. J., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem-solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100.

Wood, J. (2001). Can software support children's vocabulary development? *Language Learning and Technology, 5,* 166–201.

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control, 8*(3), 338–353.

Zimmerman, C. B. (1997). Do reading and interactive vocabulary instruction make a difference?: An empirical study. *TESOL Quarterly, 31,* 121–140.