Behavior Research Methods, Instruments, & Computers
2004, 36 (2), 203-212

Automatic classification of dysfunctional thoughts: A feasibility test

KATJA WIEMER-HASTINGS and ADRIAN S. JANIT
Northern Illinois University, DeKalb, Illinois

PETER M. WIEMER-HASTINGS
DePaul University, Chicago, Illinois

and

STEVE CROMER and JENNIFER KINSER
Northern Illinois University, DeKalb, Illinois

The identification of dysfunctional thoughts is a central effort in cognitive therapy. This paper describes the first version of a computer module that classifies dysfunctional thoughts automatically. It is part of COGNO, a system we are developing to give automatic feedback on dysfunctional thoughts. The system uses rules that were developed from language markers identified in a sample of 149 dysfunctional thoughts. The system was tested with an independent set of 112 example thoughts. The system detects the majority of dysfunctional thoughts, but works reliably only for some thought categories. Automatic thought classification may be a first step toward developing natural dialogue systems in cognitive therapy.

According to cognitive therapy, even small events such as forgetting an appointment can make individuals feel depressed or anxious if unwarranted negative interpretations of the events are generated, such as "That's just like me, I forget everything," or "I blew it; they'll never want to meet with me again." Overly negative interpretations typically make the event look worse than it really is. As such, they reflect inaccurate assumptions and conclusions, which can be classified as cognitive fallacies (Irwin & Bassham, 2003). Such thoughts have the power to make individuals feel hopeless, guilty, angry, or discouraged. The assumption that dysfunctional thoughts underlie mood (and other) disorders is at the heart of cognitive therapy, which aims to alter negatively biased thoughts, also referred to as "dysfunctional thoughts" or "cognitive distortions." Dysfunctional thoughts generally express negative perceptions of oneself, others and the world, and the future (Blackburn & Eunson, 1989; Sacco & Beck, 1995). In clinical populations, dysfunctional thoughts become default interpretations and are believed to be accurate (Beck, 1967, 1976; Sacco & Beck, 1995), which is why they are often called "automatic thoughts." Dysfunctional thoughts also occur in nonclinical populations, but less frequently (Kumari & Blackburn, 1992). Everybody generates occasional inaccurate interpretations, but depressed individuals are characterized by an overall, systematic negative bias (Beck, 1967). Research has shown a significant relation between the frequency of dysfunctional thoughts and the severity of clinical symptoms (Fennell & Campbell, 1984).

Twelve types of dysfunctional thoughts were proposed by the pioneer of cognitive therapy, Aaron T. Beck (1976; Beck, Rush, Shaw, & Emery, 1979). Examples of dysfunctional thoughts are mind reading (e.g., "She hates me") and negative predictions (e.g., "I will fail this test"). The complete list of the types of dysfunctional thoughts incorporated in this study is displayed in Table 1, with examples for each (some of the labels differ from those generated by Beck). Different dysfunctional thought categories reflect different beliefs. For example, "should" thinking is often associated with a perfectionist attitude and the belief that one has to get everything right in order to be a worthy person. Magnification of a problem shows a strong belief that only negative things are to be expected, which leads to selective attention to events that confirm the belief. Such negative beliefs or schemas are very strong and change resistant in depressed individuals (Hollon & Beck, 1979).

One core aspect of cognitive therapy is a process called cognitive restructuring, in which patients are trained to recognize dysfunctional thoughts and to form adaptive alternative interpretations (Beck, 1967, 1976). This is achieved through a variety of techniques, which, if used regularly as homework or in therapy sessions, reduce symptoms such as depression or anxiety (e.g., Burns & Auerbach, 1992; Burns & Spangler, 2000). The change of dysfunctional thoughts induced by cognitive structur-

We thank Keith Millis, Ira Bernstein, Rich Carlson, and an anonymous reviewer for very helpful comments on an earlier version of this article. Thanks are due also to Pat Sanchez, who worked on the collection of the dysfunctional thoughts in our training set. Correspondence should be addressed to K. Wiemer-Hastings, Department of Psychology, Northern Illinois University, DeKalb, IL 60115 (e-mail: katja@niu.edu).

Table 1
Categories and Examples of Dysfunctional Thoughts After A. T. Beck

Thought Category	Example
All-or-nothing thinking	I have completely failed at this task.
Negative predictions	I will never graduate.
Disqualifying the positive	Anyone could have done this.
Emotional reasoning	I feel guilty—I must have forgotten something.
Labeling	What an idiot I am!
Magnification	I forgot her birthday—I am such a lousy friend!
Mind reading	He thinks I am no good.
Overgeneralization	This kind of thing always happens to me.
"Should" thinking	I should have remembered the appointment.
Personalization	He is angry because he doesn't want me here.

ing techniques makes an independent, significant contribution to patients' improvement (Persons & Burns, 1985). Overall, the effectiveness of the modification of thoughts has been demonstrated by many outcome studies showing that cognitive therapy is highly effective, sometimes more effective than other therapy forms (Butler & Beck, 2001; Dobson, 1989; Gloaguen, Cottraux, Cucherat, & Blackburn, 1998; Strunk & DeRubeis, 2001). Cognitive therapy is particularly effective in reducing various forms of depression and anxiety (Butler & Beck, 2001).

In cognitive restructuring, several techniques can be used to elicit a patient's thoughts; these thoughts are then analyzed with respect to dysfunctional components. One classic technique for this is the "daily record of dysfunctional thoughts" (DRDT), a technique in which daily events, associated emotions, original thoughts, and adaptive responses are recorded in columns (Beck, 1976). The restructuring process requires that the patient recognize that a thought is negatively biased and maladaptive, which can be achieved through dialogues with the therapist or thought analysis in homework. Then, the thoughts can be challenged through reasoning or experiments. In the DRDT, patients practice composing adaptive responses to their own thoughts. In a recent study, a nonclinical population of college students used the DRDT for 1 week (Janit, Wiemer-Hastings, & Cromer, 2004). None of these students had scores on depression or anxiety measures that indicated a mood disorder. Students showed significant improvements in depression and anxiety scores after using the DRDT for a week.

Computers in Therapy

In recent years, a variety of computer systems have been developed, mostly with the objective of making therapy more cost-effective and more widely and easily available. In a review, Kenardy and Adams (1993) mentioned four main categories of computer use in therapy: assessment, data collection, training of professionals, and training of patients. Newer models are designed predominantly for assessment (e.g., Newman, Consoli, & Taylor, 1997) and for education about self-help strategies (e.g., Wright et al., 2002).

A few general trends have emerged from research on computer use in the clinical domain. First, patients show fairly good acceptance of computers in the therapeutic context. In fact, some studies on computer use in assessment report that patients feel more comfortable providing personal information on a computer than telling it to a therapist (Erdman, Klein, & Greist, 1985; Hart & Goldstein, 1985). At the same time, researchers emphasize the importance of designing user-friendly computer programs to make them accessible to people who are not computer literate (see Wright et al., 2002, for a good example of a user-friendly system). These efforts should further increase acceptability of computers in therapy. Second, many researchers have strongly advised that computer programs be used in conjunction with therapy administered by a human to ensure ethical and effective procedures (e.g., Agras, Taylor, Feldman, Losch, & Burnett, 1990). However, some evidence suggests that improvements can been achieved by use of computer programs alone (e.g., Selmi, Klein, Greist, Sorrell, & Erdman, 1990). As computer systems become more sophisticated, they may become increasingly useful as independent contributors to therapy.

Cognitive Therapy Applications

Because of its highly structured techniques (Selmi et al., 1990; Stuart & LaRue, 1996), the area of cognitive therapy seems promising for computer applications. Because cognitive and cognitive-behavioral therapies are so highly structured, it is possible to create quite sophisticated and useful computer tools without a need for natural language understanding. For example, patients can easily interact with the programs by providing ratings, giving yes-and-no answers, or just following system prompts to challenge a particular belief (Newman, 1999). The systems can choose the next step on the basis of a structured decision tree in which the interaction path is determined by patient responses.

It does not come as a surprise, then, that in recent years, several programs have been developed specifically for use in cognitive or cognitive-behavioral therapy (Gruber, Moran, Roth, & Taylor, 2001; Newman, 1999; Proudfoot et al., 2003; Selmi, Klein, Greist, Johnson, & Harris, 1982; Selmi et al., 1990; White, Jones, & McGarry, 2000; Wright et al., 2002). Cognitive restructuring techniques have been successfully implemented in some of these systems. For example, cognitive restructuring in The Stress Manager (Newman, 1999), an ap-

plication for Generalized Anxiety Disorder, asks a series of questions that prompt the user to rate emotions, critically analyze their thoughts for irrational beliefs, and review the evidence for the beliefs. If the user identifies particular dysfunctional thoughts, the system provides some suggestions for how to restructure them.

These systems can be easily integrated with therapy because they support the patient's homework between therapy sessions. The ability of the patient to apply techniques learned in therapy between sessions is a major predictive factor in the success of a therapy (Burns & Auerbach, 1992; Burns & Spangler, 2000; Garland & Scott, 2002; Kazantzis, Deane, & Ronan, 2000; Kenardy & Adams, 1993), and thus computer systems that aid the patient with homework assignments are valuable. Furthermore, a computer program may motivate the patient by providing feedback on homework. A computer program can also offer therapists a record of homework compliance. Homework compliance has been found to be affected positively by therapists' reviewing of assigned homework (Bryant, Simons, & Thase, 1999).

Natural Language Processing in Therapy Applications

Systems designed for dialogic interaction in therapy have not been very successful in the clinical domain. In first attempts, like the famous ELIZA (Weizenbaum, 1965), researchers sought to model therapeutic dialogues, which can take an unpredictable course. Although computer programs can be designed to give meaningful replies such as "Tell me more about X," and prompt the user to analyze his/her situation, it is difficult to design them to flexibly respond to specific problems as they emerge in a dialogue or to strategically pursue a therapeutic goal. The components of cognitive therapy (or cognitive-behavioral therapy), however, have predictable steps. Thus, they may be a promising area for the application of natural language processing techniques.

Given the central importance of identifying dysfunctional thoughts, it seems promising to develop computer systems that can recognize such thoughts. If such a task can be performed accurately, then appropriate feedback can be selected on the basis of the users' thoughts. For example, a user may enter the thought "he thinks I'm stupid." The system would detect that this user is "mind reading," and it would select category-appropriate feedback—for example, prompting the user to recognize that he/she can't read someone's mind and to think about alternative things that the other individual may have been thinking that are neutral or even positive.

In the remainder of this paper, we describe a computer application that was developed to recognize dysfunctional thoughts in sentences entered by patients. In contrast to the cognitive therapy applications mentioned above, this program does not present a complex tutorial program. Instead, it is a module that identifies dysfunctional thoughts. In the programs described earlier, patients have to identify their own dysfunctional thoughts. For example, The Stress Manager prompts users to examine their thoughts and to indicate whether they had a particular type of dysfunctional thought. While such systems provide definitions of dysfunctional thoughts, they do not have the capacity to identify them independently of the user. To do this, a computer program would need to be able to process natural language to some extent.

When patients begin therapy, they are typically not aware of dysfunctional thoughts that they may have (Beck, 1967, 1976). Such thoughts may be difficult for patients to identify because they tend to be automatic and thus accepted without critical reflection. Also, the variety of dysfunctional thought types may be overwhelming to learn at first. Yet, it is important to distinguish among them because they differ in their effects and in the cognitive strategies that would facilitate restructuring. A computer system that identifies dysfunctional thoughts could train patients to recognize them. Users could first use it to identify dysfunctional thoughts for them, to sharpen their awareness of such thoughts. In later stages, users could compare the dysfunctional thoughts they have identified themselves with the system's detections. Also, patients may systematically overlook some types of irrational beliefs, particularly those that are deeply rooted and thus the most important to change. Fortunately, evidence suggests that cognitive therapy is effective even for patients with firmly rooted dysfunctional beliefs (Halford, Bernoth-Doolan, & Eadie, 2002).

A computer system that identifies dysfunctional thoughts could further be useful in the training of professionals in the recognition of maladaptive thought patterns. Sometimes, dysfunctional thoughts can be hidden in the subtleties of language. Of course, as we shall see, a computer system that processes natural language will tend to have more problems detecting such cases, which may limit the usefulness of such programs in training therapists.

A more ambitious goal of this work is to explore ways in which the computer could be useful beyond presentation of information, a role in which the computer essentially replaces printed materials. Once the computer is tuned to recognize dysfunctional thoughts, it can also become useful in a basic function of the therapist role: It can analyze patient speech for these thoughts and initiate steps to take toward their modification, such as instructions for restructuring the thoughts or the designing of experiments to test specific thoughts. That is, this program may be the first step toward more conversational computer programs that can provide guided interactions with patients.

Computer-Facilitated Cognitive Restructuring

Over the last 2 years, we have developed a computer program called COGNO to test whether it is possible to classify dysfunctional thoughts on the basis of linguistic markers in these thoughts. The question is whether a computer can take a dysfunctional thought and automatically classify it as, for example, mind reading or magnification. In contrast to other systems, this program attempts to analyze the user's explicit thoughts so that feedback on those thoughts can be provided online. While this is still a far cry from a natural dialogue, it presents

an important feasibility test for more sophisticated dialogue programs. Furthermore, the identification of specific types of dysfunctional thoughts is a central challenge for such systems to be useful in cognitive therapy. In our application, users of the system go through a brief tutorial on dysfunctional thoughts and then enter a description of an event, their emotion, and the thoughts that went through their heads at the time of the event. The program currently analyzes only the thoughts, searching for indicators of which dysfunctional thought is present. On the basis of this procedure, standard feedback with suggestions for restructuring is selected for each event description. In this paper, we focus on this "intelligent" component of the system and its performance.

System development. The problem we addressed was the distinction of various kinds of dysfunctional thoughts. That is, the system has not yet been trained to discriminate dysfunctional from nondysfunctional (i.e., adaptive) thoughts. Theoretically, the system should not return any classification when the thought does not have any dysfunctional characteristics, but we have not yet formally tested this. Instead, in our first system evaluation, we tested whether the system could decide what kind of dysfunctional thought was present. Considering the large number of dysfunctional thought categories, this presents a substantial challenge.

Training set. The system uses a battery of syntactic and lexical matching procedures to detect dysfunctional thoughts. To establish linguistic markers and rules, a training set of dysfunctional thoughts was collected from publications on cognitive therapy, including a large number of journal articles, handbooks, and training manuals in the area of cognitive therapy. A total of 188 unique examples were collected in this way. These thoughts were then grouped into the different thought categories. Two categories were excluded from COGNO (see below), so that altogether 149 thoughts were the basis for the system presented here. A text analysis was performed on all items to identify linguistic markers that were shared by exemplars of a given category and that distinguished between the categories.

Linguistic markers. The linguistic markers included a variety of language aspects such as tense, keywords, and syntactic frames. The complete list of markers is shown in Table 2. The list is complex, but considering the task it has to accomplish, it is surprisingly compact. In fact, some types of dysfunctional thoughts (such as "should" thoughts) are easily identified with the help of a few simple rules because the logical error is verbalized explicitly (i.e., *should*, *ought to*, *must*, etc.). In contrast, discrimination of other categories is nearly impossible because they overlap in structure and wording and express conceptually similar thoughts. For example, in mental filter and tunnel vision, the individual focuses on the negative aspects of the situation. Not surprisingly, the linguistic markers characterizing similar thought types overlap, so these two types of dysfunctional thoughts had to be omitted from this first version of COGNO. The system has markers and rules for the thought categories listed in Table 1.

Decision tree. A decision tree is a hierarchical system of rules according to which thoughts are searched for any

Table 2
Linguistic Markers Used by Current Dysfunctional Thought Classification System

Marker Type	Markers
Keywords by content	1 Success and achievement
	2 Failure
	3 Disaster
	4 Expressions of inevitability
	5 External/uncontrollable cause
Keywords by part of speech	6 Adverbs of completeness
	7 Adjectives of completeness
	8 Adverbs of extreme temporal frequency
	9 Adverbs of exclusivity
	10 Causal connectives
	11 Verbs of emotion
	12 Verbs of interpretation
	13 Verbs of perception
	14 Verbs of intent/cognition
	15 Verbs of interpersonal evaluation
	16 Verbs of communication
	17 Adjectives of emotion
	18 Adjectives of negative evaluation
	19 Nouns of negative evaluation
Syntactic frames	20 PROP–VB–neg.–PROP
	21 1st person singular–I am–neg. ADJ/NN
	22 3rd person sg OR pl–[AUX]–VB
	23 3rd person sg OR pl–[AUX]–be ADJ
Grammatical markers	24 Future tense
Frozen expressions	25 Negative evaluative statements

linguistic markers, and feedback is assigned on the basis of the matches. The decision tree in our system is not hierarchical because many linguistic markers are independent of each other. Accordingly, the markers need to be checked in parallel. However, feedback for some categories is only provided in the presence of a combination of several markers. For example, emotional reasoning typically involves an emotion verb, an adjective indicating negative emotion, and a separate clause beginning with a causal connective (e.g., *so*, *because*, *therefore*). Such markers are checked in a sequence. Feedback for emotional reasoning is provided to the user only if all three produce a match.

Performance on training set. The program was first tested on the thoughts that were used to establish its markers and rules. This test is only a preliminary check, since the system can be expected to perform well on these cases. The system correctly identified 115 out of 149 thoughts (this number excludes thoughts representing the categories that were not included in this first version). Thus, 77% of the dysfunctional thoughts were classified correctly. This is a good proportion, but there are quite a few thoughts (roughly one out of four) that COGNO is unable to classify correctly. Thus, our rule and marker system captures the majority of regularities in the training set, but the marker set should be extended.

Latent semantic analysis. We compared the system's performance with that of an alternative approach using latent semantic analysis (LSA; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) to evaluate its relative strengths. In recent years, researchers have used this text-based tool quite successfully for semantic text comparisons. LSA compares two documents of varying size using a high-dimensional semantic space in which each document is represented as a vector, with one vector element for each dimension of that semantic space. The semantic space is based on an extensive co-occurrence matrix of words by documents that is reduced and optimized using single value decomposition. LSA produces a cosine value ranging from 0 to 1 to indicate the semantic similarity of any two documents (e.g., two sentences). Given a large and relevant text base for the LSA matrix, LSA cosines reliably correlate with similarity ratings provided by humans for pairs of documents.

LSA is a promising approach to the present application because it addresses a potential problem of the largely keyword-based approach we have taken. Patients may use different words to express a particular thought. Some dysfunctional thoughts may be expressed in slightly different words than the ones covered by our system rules; others may be expressed with entirely different sentence structure and content. LSA was initially developed to address this problem. It was designed to enhance search engines in document retrieval tasks to retrieve relevant documents related to the words in a query (Deerwester et al., 1990). With a word-matching procedure, a search engine is likely to produce irrelevant documents in which the query words are used in a different sense, lowering the precision (the percentage of retrieved documents that are relevant to the search) of the search system. Additionally, it is likely to miss some relevant documents that use different expressions, which would lower the system's recall (what percentage of all the existing relevant documents the system retrieved). Comparing our system with LSA classifications of thoughts provides an interesting evaluative angle because it has strengths that the method of linguistic markers lacks.

LSA has been used with most success for larger units of texts, such as paragraphs and essays. In the present application, individual thoughts need to be evaluated. However, a few successful applications have been developed over the last few years in which LSA has been used to evaluate individual sentences. For example, LSA can reliably discriminate good and bad student answers in a tutor context (Graesser et al., 2000) and can identify different reading strategies of students in verbal protocols (Kurby et al., 2003; Magliano, Wiemer-Hastings, Millis, Muñoz, & McNamara, 2002). Thus, there is some indication that LSA can be useful in the present application.

METHOD

Materials

One hundred twelve new dysfunctional thoughts were collected from a book by an expert researcher and practitioner in the field of cognitive therapy (Burns, 1980). The dysfunctional thoughts identified by the system for each test case were compared with the classification provided by the author to estimate system performance. A perfect system should classify all cases exactly like the expert.

Measures

The system's performance on the test cases was evaluated using signal detection measures and standard measures in computational linguistics. Signal detection theory classifies responses as hits, misses, false alarms, and correct rejections. For a given Category C, a hit is a test item that is classified as C by both system and expert. A miss is an item that the expert, but not the system, classifies as C. A false alarm is made when the system, but not the expert, classifies a test item as C. Finally, a correct rejection is when an item is classified as not C (¬C) by system and expert. On the basis of these four scores, the following measures were calculated.

Signal detection measures: Hit rates, false alarm rates, and *d′*. The hit rate is the proportion of correctly classified test cases for a Category C. This measure is also referred to as "recall" in the information retrieval literature. The false alarm rate is the proportion of test cases that are not part of Category C (¬C) that were erroneously classified as belonging to C. This measure is also known as the "fallout rate." Both hit and false alarm rates vary between 0 and 1. A system is better the closer its hit rate is to 1, and the closer its false alarm rate is to 0. Finally, d' measures the extent to which the system can distinguish cases belonging to a Category C versus other categories (¬C). As the hit rate increases and the false alarm rate decreases, d' gets larger. It is equal to 0 when the system performs at chance level (i.e., hit and false alarm rates are .50).

Information retrieval measures. We also evaluated the system with standard measures used in the evaluation of natural language processing systems (see Manning & Schütze, 2002, for a good introduction), which are also based on signal detection measures. In addition to recall and fallout, which are identical to hit and false alarm rates, respectively, we report the precision, which is the proportion of test cases that the system classifies as C that in fact belong in Category C (i.e., the number of hits, divided by hits plus

false alarms). Precision scores vary between 0 and 1, with higher values indicating better performance. For 10 categories, a system performing at chance would have a precision of .1. We include precision because high recall can be achieved at the cost of precision. For example, a system has perfect recall for Category C if it flags all test cases as C, but it would produce the maximum number of false alarms possible. Thus, it would not be a good system. Ideally, a system balances recall and precision to minimize the annoyance of false alarms, without missing critical feedback.

***F* measure**. The *F* measure (Manning & Schütze, 2002; van Rijsbergen, 1979) combines recall and precision, allowing for easier system comparison. *F* allows for different weighting of recall and precision scores. It is calculated as

$$F = \frac{1}{\alpha \frac{1}{\mathrm{P}} + (1-\alpha)\frac{1}{\mathrm{R}}},$$

where α is the weight ranging from 0 to 1, R is the recall, and P is precision. *F* becomes larger the higher the hit rate and the lower the false alarm rate, with the highest value being 1. In our analyses, we set α at .50 for equal weighting of recall and precision because there is no a priori reason to bias the measure in favor of higher recall (i.e., more false alarms) or higher precision (i.e., more misses).

Summary Measures

Signal detection theory is usually applied to binary classification problems (e.g., signal vs. no signal). There are two procedures for calculating such measures for a multicategory system like the present one (Manning & Schütze, 2002). One procedure, called macro-averaging, calculates all measures separately for each category and then averages the measures across categories. The second approach, micro-averaging, sums up hits, false alarms, misses, and correct rejections across categories. Measures are then calculated collectively for this entire set. The results can differ when there are unequal numbers of test cases for the different categories. Macro-averaging weights each category equally regardless of the number of items, whereas micro-averaging gives more weight to larger categories. We compared results from both approaches since we had unequal numbers of test items for the categories.

RESULTS

The measures for the test set calculated for our classification system are shown in Table 3. Overall, the system's recall was .77 (or .74). That is, it detected a given dysfunctional thought in 7 to 8 out of 10 cases. This rate is exactly as high as the rate that was achieved for the actual training set. This is a promising result, since it suggests that the rules were not tuned to idiosyncratic linguistic characteristics of the training set, but generalize to new examples. The present set of markers covers the majority of dysfunctional thoughts. Still, it is clear from the measures that additions to the present system are required to make it usable in praxis, especially if used by patients.

System Discrimination

On average, d' was above 2, which indicates good discrimination. The d' scores obtained for the different categories vary considerably. The numbers in Table 3 suggest that the d' scores were higher for categories with very small test sets, which is probably a mathematical artifact. This is reflected in the difference in the average d' scores based on micro- and macro-averaging. d' was lower when categories were weighted by the number of test cases. This smaller d' probably estimates the system performance more accurately. This smaller score, $d' = 2.01$, is a satisfactory sensitivity score. The average *F* score is .58 or .61, respectively. While this leaves much room for improvement, the scores for many categories were considerably higher, suggesting that the weaknesses in the present system may represent local problems.

Analysis by Category

There were a few categories for which the system had high hit and low false alarm rates, which suggests good discriminative markers. This was true for cases of "should" thinking, labeling, emotional reasoning, and negative predictions. These categories all had precision scores of .61 and higher. Precision was particularly high for discounting the positive and emotional reasoning. These four categories also had the highest *F* measures.

In contrast, cases for magnification and discounting the positive had the lowest hit rates, suggesting that additional markers are needed to account for all the items.

Table 3
Signal Detection and Information Retrieval Measures for System Test Case Classification by Thought Category

Category	*N*	HR (= R)	FA Rate	d'	P	*F*
All-or-nothing	7	.86	.40	1.33	.13	.22
Negative predictions	21	.95	.14	2.72	.61	.74
Discounting the positive	15	.40	.01	2.07	.86	.55
Emotional reasoning	5	.80	.00	4.84	1.00	.89
Labeling	23	.91	.13	2.46	.64	.75
Magnification	10	.40	.13	0.87	.24	.30
Mind reading	4	.50	.01	4.00	.67	.57
Overgeneralization	15	.80	.12	2.01	.50	.62
Personalization	8	.75	.03	2.55	.50	.60
Should	4	1.00	.03	6.88	.73	.84
Macro-average	11	.74	.10	2.97	.59	.61
Micro-average	11	.77	.10	2.01	.46	.58

Note—*N*, number of cases; HR, hit rate; FA rate, false alarm rate; R, recall; P, precision; *F*, *F* measure.

Magnification and all-or-nothing thinking had very low precision scores. This suggests that their markers are not discriminative—which is also reflected in their low *F* scores.

Comparison with LSA

We conducted a feasibility test for LSA in this application. To estimate the extent to which it may be useful in dysfunctional thought detection, in particular for categories that are not covered well by linguistic markers by the present system, each test sentence was compared by LSA with the set of training sentences for each category. The entire set of training cases for the thought category was used as a comparison text to reduce the influence of idiosyncratic expressions and events. Whether LSA detected the particular dysfunctional thought was determined in two ways. First, the thought category that produced the highest cosine with the test thought was selected as the detected dysfunctional thought by LSA. This was the strict criterion since our system was permitted multiple classifications. To achieve a fairer comparison, we also included a lax criterion, in which we considered the LSA classification as a hit when the correct dysfunctional thought category produced one of the highest three cosines. The results are presented in Tables 4 and 5.

The average scores on all measures show that, overall, LSA did not perform as well as the rule system using markers. This was true both using a strict criterion (Table 4) or a lax criterion (Table 5). The difference between these two was reflected mostly in the changes in hit and false alarm rates, both of which, as may be expected, increased using the lax criterion. The d' scores suggest that whether a lax or a strict criterion provides the better results depends on how the individual categories are weighted. If the category numbers are taken into consideration (micro-averaging), the strict criterion is the better choice. Additionally, the precision scores for the lax criterion are often not substantially different from chance (P = .1), suggesting that under this criterion, many LSA hits are by chance.

An interesting question was whether LSA would outperform our system on categories that the rules did not handle well. If so, LSA could be used as a complementary system to increase overall performance. However, this was not the case. LSA performed best on emotional reasoning, personalization, and "should" thinking. Interestingly, two of these were also handled well by our system. It is possible that our rule system captures the same information that influences the LSA measures.

DISCUSSION

Computer programs utilizing natural language processing capabilities have been of limited use in therapy because the dialogue between therapist and patient is relatively unstructured and unpredictable. We have shown that a computer system can be set up to reliably discriminate at least some types of dysfunctional beliefs as they are reflected in patients' thoughts. In terms of discrimination and precision strength, the evaluation of our first version shows that this approach is promising.

Of course, this holds true mostly from a developer's point of view. Missing roughly 25% of thoughts presents a problem from the therapist's point of view. Our analysis can drive system improvements. For example, the results show that the system identified a majority of occurrences for many types of dysfunctional thoughts, suggesting that their markers are working quite reliably in the distinction of dysfunctional thoughts. This was especially true for negative predictions, emotional reasoning, and "should" thinking. In contrast, magnification thoughts were handled particularly badly by the current system. These data can be used to direct improvement efforts at the weakest parts of the system.

User Goals, Recall, and Precision

One interesting question concerns whether the present system should be designed for maximum recall or precision, or to keep both in balance. The advantage of high recall is that users will obtain feedback for most of their dys-

Table 4
Signal Detection and Information Retrieval Measures for Latent Semantic Analysis Classification of Test Cases by Thought Category, Following a Strict Criterion

Category	*N*	HR (= R)	FA Rate	*d'*	P	*F*
All-or-nothing	7	.00	.00	0.00		
Negative predictions	21	.10	.04	0.47	.33	.15
Discounting the positive	15	.31	.06	1.06	.45	.37
Emotional reasoning	5	.83	.07	2.42	.38	.53
Labeling	23	.67	.13	1.55	.55	.60
Magnification	10	.00	.07	−2.53	.00	
Mind reading	4	.40	.09	1.09	.17	.24
Overgeneralization	15	.50	.19	0.87	.29	.36
Personalization	8	.60	.02	−0.05	.60	.60
Should	4	.56	.03	2.03	.63	.59
Macro-average	11	.40	.07	0.69	.38	.40
Micro-average	11	.38	.07	1.17	.38	.38

Note—Some cells are empty because of divisors being equal to zero. *N*, number of cases; HR, hit rate; FA rate, false alarm rate; R, recall; P, precision; *F*, *F* measure.

Table 5
Signal Detection and Information Retrieval Measures for Latent Semantic Analysis Classification of Test Cases by Thought Category, Following a Lax Criterion

Category	*N*	HR (= R)	FA Rate	*d′*	P	*F*
All-or-nothing	7	.38	.22	0.47	1.07	.17
Negative predictions	21	.48	.21	0.75	.32	.38
Discounting the positive	15	.50	.17	0.95	.31	.38
Emotional reasoning	5	1.00	.43	4.17	.11	.19
Labeling	23	.79	.55	0.68	.26	.40
Magnification	10	.45	.35	0.26	.11	.18
Mind reading	4	.60	.28	0.83	.09	.15
Overgeneralization	15	.75	.59	0.45	.16	.27
Personalization	8	.60	.22	1.02	.10	.18
Should	4	.89	.05	2.92	.57	.70
Macro-average	11	.64	.31	1.25	.21	.30
Micro-average	11	.64	.30	0.87	.19	.29

Note—*N*, number of cases; HR, hit rate; FA rate, false alarm rate; R, recall; P, precision; *F*, *F* measure.

functional thoughts, providing them with a substantial amount of information. At the same time, a high rate of false alarms can be confusing and difficult for a patient to handle. A professional using the system for training, on the other hand, would be better able to disregard feedback that does not seem to fit. One way to empirically explore the false alarm problem is to have test users rate how well the feedback fits the thoughts they entered. Users would vary in the amount of their expertise in cognitive therapy.

The precision of our system may be slightly better than the rates presented in the data suggest. It is permissible for our system to report several dysfunctional thoughts for one given thought. This is useful, since some thoughts contain several types of "thought errors." For example, the thought "I know the jerk will be late! He always is!" is at once a negative prediction, labeling, and overgeneralization. Thus, the system needs to classify a thought in all possible ways. In our study, the system was evaluated only with respect to one major dysfunctional thought, so a small percentage of the "false alarms" may have been relevant reports.

Future Directions

COGNO can be improved in a variety of ways. One thing that we have not yet evaluated is how well the system discriminates dysfunctional and adaptive thoughts. We have conducted a pilot test with 19 thoughts (e.g., "I so regret having come here") that are negative but not distorted. Fifty-eight percent remained unclassified, correctly. This analysis showed that some markers produce a high rate of false alarms on such items, particularly the future tense marker (four out of the eight false alarms). Thus, a few markers will have to be modified or examined in combination with other markers once the system is modified for discrimination of dysfunctional and adaptive thoughts.

A larger corpus of dysfunctional thoughts than we have used in this study would make possible two further minor improvements to COGNO. First, the set of linguistic markers could be expanded to cover a wider range of expressions. Second, weights could be derived for the system's markers. An extensive corpus analysis could provide information about two critical characteristics of our markers: discriminative power and typicality. A marker is discriminative to the extent that it occurs only in exemplars for a small number of thought categories. Highly discriminative markers are more valuable because they increase the number of hits but not false alarms, thereby increasing *F* and *d′* scores. A marker is typical the more exemplars of a given thought category it occurs in. Highly typical markers are economical because they cover the majority of examples for a category of thoughts. If a category has no typical markers, it has to be represented by many markers, which are unlikely to generalize to new thoughts in this category. With discrimination and typicality scores, classification could be made contingent on the sum of the markers' weights exceeding a certain threshold. This would increase the system's precision.

Other aspects for improvement are more challenging. Most notably, there are ways to express dysfunctional thoughts of certain categories in implicit ways—ways that are not expressed by surface linguistic markers. For example, "As if I didn't know better" expresses a dysfunctional thought of the "should" category: the speaker failed an expectation to (not) perform a certain action. Similarly, a thought that disqualifies some positive event can be expressed without containing any of the markers we identified for these thoughts: "So I passed the exam. So what?" It is quite difficult to formulate rules to identify such thoughts. One possibility is to include so-called "frozen expressions" as markers, which contain complete thoughts like "so what?"

Use of Latent Semantic Analysis

One has to understand the meaning of the thought to realize that it expresses a failed self-expectation or a dismissal of an accomplishment. We discussed LSA as a possible remedy for this problem. In the high-dimensional

LSA space, documents (such as statements) are represented similarly even if they do not have any word overlap. The effect of singular value decomposition is that the dimensionality of the LSA space is reduced in an optimal way so that semantically similar expressions are represented similarly. As discussed, LSA did not outperform our system—particularly not in categories that our system does not perform well in. Further studies are required to estimate the extent to which this is a result of the stricter classification constraints that were applied to the LSA classifications.

LSA presents two challenges. The first is not directly evident from the results we presented in Tables 4 and 5. The cosines generated for the different categories were often different by only one decimal point (.01). A system may be programmed to select only the categories with the absolute highest cosine, but this may still be a number of categories. Especially categories identified correctly by the lax criterion were often one of five to seven categories with the highest matches. Thus, for many types of thoughts, LSA does not discriminate very well between categories.

The second issue is related to this problem. One cannot always assume (as was the case in this study) that a thought contains a dysfunctional aspect. In the application for which we are developing this system, the system needs to decide both whether a dysfunctional thought is present, and which type of dysfunctional thought it is. To decide the former, a criterion is required. Our system returns no feedback if no markers match the current expression. However, LSA returns a pattern of cosines in either case. So, while cosine patterns can solve the question of which of the given categories is the best match, the extent to which cosine values could be used to determine whether a dysfunctional thought is present is currently unclear. Often differences in cosine values are subtle, and finding a critical cosine value to report a dysfunctional thought may be impossible.

LSA may be useful in a different respect, though. Our system was developed and tested with exemplars that were presumably "polished" for publication—that is, most of them conveyed the type of error in a straightforward way. It is likely that when used with a wide range of users, the system would have to deal with more "muddy" thoughts. LSA may be better able to handle such thoughts and may thus adapt the system to a wider range of users. It may be particularly useful if used in combination with the markers.

After improvements, our tool could be useful for training individuals to recognize dysfunctional thoughts, for help with therapy homework assignments, and perhaps even as a stand-alone system to help users defeat patterns of self-impairing thoughts in nonclinical settings. Furthermore, it may be used to train professionals in the recognition of dysfunctional thoughts. We believe that this system is a promising first step toward a natural dialogue system in cognitive therapy. Our findings suggest that computers may become increasingly useful in therapeutic work in a variety of ways.

REFERENCES

Agras, W. S., Taylor, C. B., Feldman, D. E., Losch, M., & Burnett, K. F. (1990). Developing computer-assisted therapy for the treatment of obesity. *Behavior Therapy*, **21**, 99-109.

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. New York: Harper & Row.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. New York: Meridian.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford.

Blackburn, I.-M., & Eunson, K. M. (1989). A content analysis of thoughts and emotions elicited from depressed patients during cognitive therapy. *British Journal of Medical Psychology*, **62**, 23-33.

Bryant, M. J., Simons, A. D., & Thase, M. E. (1999). Therapist skill and patient variables in homework compliance: Controlling an uncontrolled variable in cognitive therapy outcome research. *Cognitive Therapy & Research*, **23**, 381-399.

Burns, D. D. (1980). *Feeling good*. New York: Avon.

Burns, D. D., & Auerbach, A. H. (1992). Does homework compliance enhance recovery from depression? *Psychiatric Annals*, **22**, 464-469.

Burns, D. D., & Spangler, D. L. (2000). Does psychotherapy homework lead to improvements in depression in cognitive-behavioral therapy or does improvement lead to increased homework compliance? *Journal of Consulting & Clinical Psychology*, **68**, 46-56.

Butler, A. C., & Beck, J. S. (2001). Cognitive therapy outcomes: A review of meta-analyses. *Tidsskrift for Norsk Psykologforening*, **38**, 698-706.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**, 391-407.

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting & Clinical Psychology*, **57**, 414-419.

Erdman, H. P., Klein, M. H., & Greist, J. H. (1985). Direct patient computer interviewing. *Journal of Consulting & Clinical Psychology*, **53**, 760-773.

Fennell, M. J., & Campbell, E. A. (1984). The Cognitions Questionnaire: Specific thinking errors in depression. *British Journal of Clinical Psychology*, **23**, 81-92.

Garland, A., & Scott, J. (2002). Using homework in therapy for depression. *Journal of Clinical Psychology*, **58**, 489-498.

Gloaguen, V., Cottraux, J., Cucherat, M., & Blackburn, I.-M. (1998). A meta-analysis of the effects of cognitive therapy in depressed patients. *Journal of Affective Disorders*, **49**, 59-72.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the Tutoring Research Group (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, **8**, 149-169.

Gruber, K., Moran, P. J., Roth, W. T., & Taylor, C. B. (2001). Computer-assisted cognitive behavioral group therapy for social phobia. *Behavior Therapy*, **32**, 155-165.

Halford, W. K., Bernoth-Doolan, S., & Eadie, K. (2002). Schemata as moderators of clinical effectiveness of a comprehensive cognitive behavioral program for patients with depression or anxiety disorders. *Behavior Modification*, **26**, 571-593.

Hart, R. R., & Goldstein, M. A. (1985). Computer-assisted psychological assessment. *Computers in Human Services*, **1**, 69-75.

Hollon, S. D., & Beck, A. T. (1979). Cognitive therapy of depression. In P. C. Kendall & S. D. Hollon (Eds.), *Cognitive-behavioral interventions: Theory, research, and procedures* (pp. 153-204). New York: Academic Press.

Irwin, W., & Bassham, G. (2003). Depression, informal fallacies, and cognitive therapy: The critical thinking cure? *Inquiry: Critical Thinking Across the Disciplines*, **21**, 15-21.

Janit, A., Wiemer-Hastings, K., & Cromer, S. (2004). *Mood changes through short-term cognitive restructuring in a nonclinical population*. Unpublished manuscript.

Kazantzis, N., Deane, F. P., & Ronan, K. R. (2000). Homework assignments in cognitive and behavioral therapy: A meta-analysis. *Clinical Psychology—Science & Practice*, **7**, 189-202.

Kenardy, J., & Adams, C. (1993). Computers in cognitive-behavior therapy. *Australian Psychologist*, **28**, 189-194.

Kumari, N., & Blackburn, I.-M. (1992). How specific are negative automatic thoughts to a depressed population? An exploratory study. *British Journal of Medical Psychology*, **65**, 167-176.

Kurby, C. A., Wiemer-Hastings, K., Ganduri, N., Magliano, J. P., Millis, K. K., & McNamara, D. S. (2003). Computerizing reading training: Evaluation of a latent semantic analysis space for science text. *Behavior Research Methods, Instruments, & Computers*, **35**, 244-250.

Magliano, J. P., Wiemer-Hastings, K., Millis, K. K., Muñoz, B. D., & McNamara, D. S. (2002). Using latent semantic analysis to assess reader strategies. *Behavior Research Methods, Instruments, & Computers*, **34**, 181-188.

Manning, C. D., & Schütze, H. (2002). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Newman, M. G. (1999). The clinical use of palmtop computers in the treatment of generalized anxiety disorder. *Cognitive & Behavioral Practice*, **6**, 222-234.

Newman, M. G., Consoli, A., & Taylor, C. B. (1997). Computers in assessment and cognitive behavioral treatment of clinical disorders: Anxiety as a case in point. *Behavior Therapy*, **28**, 211-235.

Persons, J. B., & Burns, D. D. (1985). Mechanisms of action of cognitive therapy: The relative contributions of technical and interpersonal interventions. *Cognitive Therapy & Research*, **9**, 539-551.

Proudfoot, J., Swain, S., Widmer, S., Watkins, E., Goldberg, D., Marks, I., Mann, A., & Gray, J. A. (2003). The development and beta-test of a computer-therapy program for anxiety and depression: Hurdles and lessons. *Computers in Human Behavior*, **19**, 277-289.

Sacco, W. P., & Beck, A. T. (1995). Cognitive theory and therapy. In E. E. Beckham & W. R. Leber (Eds.), *Handbook of depression* (2nd ed., pp. 329-351). New York: Guilford.

Selmi, P. M., Klein, M. H., Greist, J. H., Johnson, J. H., & Harris, W. G. (1982). An investigation of computer-assisted cognitive-behavior therapy in the treatment of depression. *Behavior Research Methods & Instrumentation*, **14**, 181-185.

Selmi, P. M., Klein, M. H., Greist, J. H., Sorrell, S. P., & Erdman, H. P. (1990). Computer-administered cognitive-behavioral therapy for depression. *American Journal of Psychiatry*, **147**, 51-56.

Strunk, D. R., & DeRubeis, R. J. (2001). Cognitive therapy for depression: A review of its efficacy. *Journal of Cognitive Psychotherapy*, **15**, 289-297.

Stuart, S., & LaRue, S. (1996). Computerized cognitive therapy: The interface between man and machine. *Journal of Cognitive Psychotherapy*, **10**, 181-191.

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.

Weizenbaum, J. (1965). ELIZA—A computer program for the study of natural language communication between man and machine. *Communication of the Association for Computer Machinery*, **9**, 36-45.

White, J., Jones, R., & McGarry, E. (2000). Cognitive behavioural computer therapy for the anxiety disorders: A pilot study. *Journal of Mental Health*, **9**, 505-516.

Wright, J. H., Wright, A. S., Salmon, P., Beck, A. T., Kuykendall, J., Goldsmith, L. J., & Zickel, M. B. (2002). Development and initial testing of a multimedia program for computer-assisted cognitive therapy. *American Journal of Psychotherapy*, **56**, 76-86.

(Manuscript received November 4, 2003;
revision accepted for publication March 26, 2004.)