In J. Kay (Ed.), *User Modeling: Proceedings of the Seventh International Conference, UM99*. Vienna, New York: Springer Wien New York, 1999. This paper available from http://www.cs.uni-sb.de/users/jameson/.

Interpreting Symptoms of Cognitive Load in Speech Input

André Berthold and Anthony Jameson*

Department of Computer Science, University of Saarbrücken, Germany

Abstract. Users of computing devices are increasingly likely to be subject to situationally determined distractions that produce exceptionally high cognitive load. The question arises of how a system can automatically interpret symptoms of such cognitive load in the user's behavior. This paper examines this question with respect to systems that process speech input. First, we synthesize results of previous experimental studies of the ways in which a speaker's cognitive load is reflected in features of speech. Then we present a conceptualization of these relationships in terms of Bayesian networks. For two examples of such symptoms—sentence fragments and articulation rate—we present results concerning the distribution of the symptoms in realistic assistance dialogs. Finally, using artificial data generated in accordance with the preceding analyses, we examine the ability of a Bayesian network to assess a user's cognitive load on the basis of limited observations involving these two symptoms.

1 The Challenge for User Modeling

When cosmonauts on the space station Mir communicate with ground control, their speech is monitored by psychologists for symptoms of stress (Arnold, 1997). The interpretation of the symptoms in turn influences the nature of the dialogs conducted with the cosmonauts.

Computer users do not in general stray quite as far from home as the Mir cosmonauts, nor are they subjected to the same sort of stress. But the mobility of modern computing devices has moved them ever further into the hustle and bustle of everyday life. Situational distractions can have major impact on the quality of interaction with a system—as anyone who has tried to jot down a person's address with a handheld device while standing on a street corner can testify. For user modeling research, situational distractions represent one more thing that a system can try to recognize and adapt to. Adaptation may involve, for example, a simplification of either the system's output or the required user input, in cases where situational distractions are suspected.

1.1 Scenario and Field Study

For concreteness, consider the example scenario handled by the dialog system READY (see, e.g., Jameson et al., 1999): Users are drivers whose cars need minor repairs; they request assistance from the system in natural language by phone. Our first step in studying this scenario was to get a concrete idea of the cognitive load induced by this situation and the ways in which it is

* This research is being supported by the German Science Foundation (DFG) in its Collaborative Research Center on Resource-Adaptive Cognitive Processes, SFB 378, Project B2, READY. The comments of the two anonymous reviewers strongly influenced the content of the final version.

manifested in the users' speech:[1] In a field study conducted on a winter night beside a fairly busy road, each of 8 subjects was given the task of identifying and repairing an intentionally created mechanical problem with a car. They communicated with a professional auto repairman via cellular phone. To get an idea of the information present in features of the subjects' speech, we analyzed the 8 dialogs in detail. For example, filled and silent pauses were measured and errors were classified.

In Sections 2 through 4, we will see how the data from this field study can be analyzed together with results of laboratory experiments of previous researchers so as to yield an empirical basis for a user modeling component for a dialog system. We will then check whether such a user modeling component, if given a sufficiently sound empirical basis, can make usefully accurate inferences on the basis of the limited data about a user that is available in this scenario.

1.2 Determinants of Cognitive Load

In this paper, the term *cognitive load* refers to the demands placed on a person's working memory by (a) the main task that she is currently performing, (b) any other task(s) she may be performing concurrently, and (c) distracting aspects of the situation in which she finds herself.

In the example scenario, we view the main task of the user ($\mathcal{U}$) as that of communicating with the mechanic (or a corresponding system $\mathcal{S}$). Concurrent tasks can involve looking for things, performing actions on the car, or communicating with other persons. Distracting aspects of the situation can include noises and events that interfere with one's concentration on task performance, as well as internal factors like emotional stress that have similar effects.

In the dynamic Bayesian networks that form the core of READY's user model, these types of influence on a user's continually changing cognitive load are modeled separately (see, e.g., Jameson et al., 1999; Schäfer and Weyrath, 1997). In this paper, we will simply consider the problem of assessing the total load currently placed on $\mathcal{U}$'s working memory, regardless of its origin. This load will be assumed to remain constant throughout the period during which it is being assessed.

2 Overview of Symptoms and Their Modeling

We reviewed literature from psycholinguistics and linguistics looking for evidence concerning the effects of cognitive load on features of speech. Table 1 gives a high-level summary of the results of this survey.[2]

Figure 1 shows how the relationships between these symptoms and cognitive load can be modeled with a Bayesian network.[3] To see the meaning of the variables, suppose that various factors have created a POTENTIAL WM LOAD for $\mathcal{U}$. If this load is too great for $\mathcal{U}$ to handle without

[1] The READY system also tries to recognize and adapt to the user's time pressure. For reasons of space, this variable will be mentioned only in passing in this paper.

[2] A much more detailed discussion of these results is given by Berthold (1998), along with references to the individual studies and results for less important features not listed here.

[3] For introductions to Bayesian networks, see, e.g., Russell and Norvig (1995) or Pearl (1988). An overview of their applications to user modeling is given by Jameson (1996).

Table 1. Summary of previous results concerning potential speech symptoms of cognitive load.

Symptoms involving output quality			Symptoms involving output rate		
Feature	Tendency[a]	Tally[b]	Feature	Tendency	Tally
Sentence fragments (number)	+	4/5	Articulation rate	−	7/7
False starts (number)	+	2/4	Speech rate	−	7/7
Syntax errors (number)	+	1/1	Onset latency (duration)	+	9/11
Self-repairs (number)	+, −, 0[c]	2, 1, 4	Silent pauses (number)	+	4/5
			Silent pauses (duration)	+	8/10
			Filled pauses (number)	+	4/6
			Filled pauses (duration)	+	1/2
			Repetitions (number)	+	5/6

[a] "+" means that the measure was generally found to increase under conditions of high cognitive load; "−" means the opposite.

[b] "m/n" means that of n relevant studies, m found the tendency indicated in the second column. (In most—but not all—cases the tendency was statistically significant.)

[c] Results concerning self-repairs show an inconsistent pattern.

difficulty, $\mathcal{U}$ may cope with the overload by reducing the speed of speech generation—for example, by pausing intermittently to think or to deal with distractions. (The extent to which $\mathcal{U}$ does this can be influenced by features of the task as well as by $\mathcal{U}$'s time pressure and preferences.) Any such speed reduction can be reflected in specific symptoms like the ones shown on the right in Table 1. Because of the slowing, the ACTUAL WM LOAD—which can be conceptualized as the amount of cognitive work that has to be done in a given unit of time—will be reduced.

On the other hand, $\mathcal{U}$ may for various reasons avoid slowing down, or may slow down only to a degree that is inadequate to reduce the ACTUAL WM LOAD to a normal level. In this case, the high ACTUAL WM LOAD is likely to be reflected in various types of defect in the utterances produced, such as the types listed in the left-hand side of Table 1 (cf. the left-hand side of Figure 1).[4]

So far, we are aware of only partial and indirect evidence in favor of the speed-accuracy tradeoff postulated in Figure 1. Concerning the relationships between the nodes for the individual symptoms and their parent nodes, useful empirical data can be extracted from the studies summarized in Table 1 and from our own field study that was sketched above. The next two sections will show how this can be done, using one example from each of the two broad categories of symptoms, starting with one that involves a decline in the quality of output.

3 Sentence Fragments as a Symptom

A *sentence fragment* can be defined as an incomplete syntactic structure *I* for which there exists a syntactic continuation *C* such that *I C* constitutes a well-formed sentence. After articulating *I*,

[4] Baber et al. (1996), while not explicitly postulating the relationships depicted in Figure 1, discuss a number of phenomena and relationships that are consistent with this account.

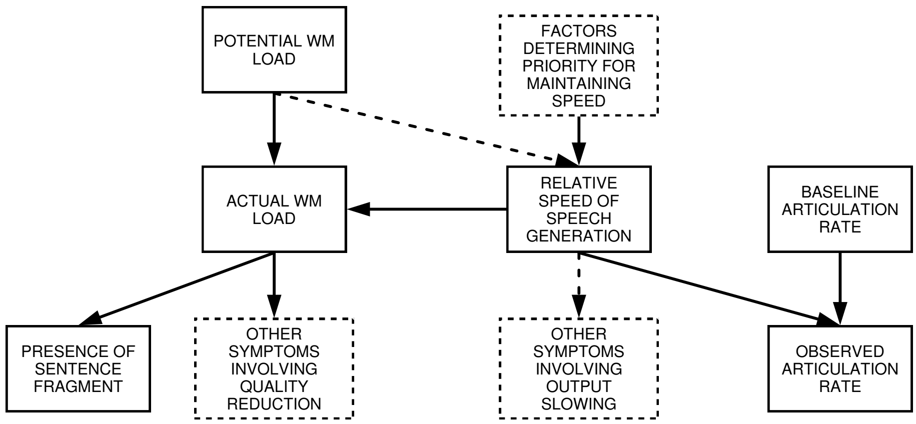

Figure 1. Simplified depiction of part of a Bayesian network for interpreting symptoms of cognitive load. (Each box with a solid border represents a node that corresponds to a single variable; each box with a dashed border denotes a group of variables that play a similar role in the network. Solid and dashed arrows denote positive and negative causal influences, respectively.)

the speaker either gives up the dialog turn or begins a new sentence *N*. Here are some examples from the field study:

Just a minute, I'll look. The cables ... [gives up turn]

Yes, that's ... uh, just keep repeating.

In many cases, the new sentence *N* begins with an alternative formulation of the content of *I*. In these cases, the sequence *I N* represents a particular type of self-repair called a *false start*. Some relevant empirical studies have looked specifically at false starts, while others have considered the broader class of sentence fragments. Since it's difficult for a system to determine automatically whether the material that follows a sentence fragment constitutes a self-repair, we will likewise ignore this distinction here and consider simply whether an utterance contains a sentence fragment.

Previous empirical results. Previous results for sentence fragments (including false starts) can be seen in the left-hand side of Table 1. The five studies in which a concurrent task was used to induce cognitive load produced the strongest effects: The concurrent task multiplied the frequency of sentence fragments by factors ranging from 1.52 to 5.50, with an average of 3.34.[5]

Distribution in the field study. Table 2 classifies the 54 sentence fragments that were found in the 628 dialog turns in our field study. The 15 fragments in the lower half of the table illustrate how sentence fragments can occur independently of cognitive load. The last category could presumably be recognized by the system and kept separate from the other categories—at least in cases where it was the system itself that interrupted the user.

[5] A single study by Roßnagel (1995a) that yielded similar factors in the opposite direction has yet to be explained.

Table 2. Frequency of six types of sentence fragments found in the field study dialogs.

Possibly due to high cognitive load:

- 24 turns consisting of (or ending with) a fragment
- 9 fragments followed by formulations with similar meaning
- 6 fragments followed by formulations with different meaning

Probably independent of cognitive load

- 7 sentences possibly aborted because of the arrival of new information or perceptions
- 1 sentence intended to be completed by dialog partner
- 7 sentences interrupted by the dialog partner

On the whole, though, it is not a trivial task to recognize sentence fragments automatically with a speech recognition system. The same is true of most of the other symptoms listed in Table 1.[6]

Modeling. In a Bayesian network such as that of Figure 1, observations of sentence fragments can be taken into account most straightforwardly with a single node that has two possible values: whether the most recent dialog turn of the user contained a sentence fragment or not. But what about the conditional probabilities that link this node with its parent node ACTUAL WM LOAD? Let us assume that the cognitive load induced in the five experiments that involved concurrent tasks corresponds roughly to the highest level of the variable ACTUAL WM LOAD. Assume further that the low-load condition of these experiments corresponds to the lowest level of this variable. Then the experimental results suggest the following constraints on the conditional probabilities relating ACTUAL WM LOAD (A) and PRESENCE OF SENTENCE FRAGMENT (F):

$$\frac{P(F = \text{Yes}|A = \text{Highest})}{P(F = \text{Yes}|A = \text{Lowest})} = 3.3 \qquad \frac{P(F = \text{No}|A = \text{Highest})}{P(F = \text{No}|A = \text{Lowest})} = .94$$

The second ratio, .94, reflects the fact that dialog turns that *do not* contain a sentence fragment are slightly less likely given a high level of ACTUAL WM LOAD. We found only one previous study (Roßnagel, 1995b) that yields data that can be used for the estimation of this ratio, but the exact value is actually unimportant: Given that sentence fragments occur in fewer than 10% of dialog turns, this ratio must be some number slightly less than 1.0. The consequence is that the observation of a single dialog turn *without* a sentence fragment will only slightly diminish $\mathcal{S}$'s estimate of $\mathcal{U}$'s cognitive load.

Before examining what sort of diagnostic performance these basic relationships can give rise to, let us examine a different type of symptom of cognitive load.

[6] Berthold (1998) discusses some of the problems involved and the possibilities offered by various approaches to speech recognition. The strategy pursued in the READY project is to determine which symptoms can play a useful role in a dialog system before making the considerable effort required to extract them automatically while using a speech recognizer. Accordingly, for system tests the properties of the input utterance are specified via a menu interface (see Jameson et al., 1999).

4 Articulation Rate as a Symptom

Among the symptoms that reflect the speaker's attempt to reduce output rate, the various types of pauses have been most thoroughly investigated (see Table 1). Though pauses also figure prominently in READY's modeling, we will look here at a less complex and less obvious symptom: the rate at which the speaker articulates syllables. To avoid overlap with the definition of pauses, we adopt the following definition:

$$\text{Articulation rate} = \frac{\text{Number of syllables articulated}}{\text{Total duration of articulated syllables}}$$

Filled pauses are left out of consideration, as are silent pauses whose length exceeds a certain threshold (here: 200 msec). The following translated example of an utterance produced by our mechanic illustrates this definition:

<uh> <P> *In the* <P> *inside under the steering wheel* <P> *to the left* <P> *there's a fuse box.*

Here, <P> stands for a silent pause; only the underlined material enters into the computation of articulation rate.[7]

Previous empirical results. Seven studies were found that measured articulation rate under conditions of varying cognitive load. As is indicated in Table 1, all of them found a tendency toward slower articulation given higher load. In the five studies that yielded specific data on average articulation rates, the rate reductions in the higher-load condition ranged from 8.8% to 19.7%, with an average of 13.6%. All of these studies induced high cognitive load by making the speaking task more difficult. We would expect the slowing to be more drastic in a condition involving a concurrent task, since this type of manipulation produced the strongest effects on sentence fragments and also in the studies on pauses.

Distribution in the field study. In a dialog situation, some dialog turns contain only a few syllables (e.g., *Yes, I can*). Measurement of articulation rate is problematic for such turns, since it would depend strongly on the properties of the syllables involved, on random variation, and on measurement error. On the basis of an initial analysis of the empirical distributions, we eliminated from consideration measurements of articulation rate for dialog turns involving 3 or fewer syllables.

The articulation rates for the 8 callers in the field study ranged from 6.3 to 7.7 syllables per second, with an average of 7.0 syllables/s. The utterances of each individual caller also varied somewhat in articulation rate, the standard deviations ranging from 1.0 to 2.1 syllables/s, with an average standard deviation of 1.35.

Modeling. In Figure 1, OBSERVED ARTICULATION RATE is viewed as being influenced by RELATIVE SPEED OF SPEECH GENERATION, but it also has a second parent node, BASELINE ARTICULATION RATE. This node is included because individual speakers differ systematically in their usual articulation rate, independently of any variations in cognitive load (cf. Goldman-Eisler, 1968). (The differences just cited in the average articulation rates of the 8 callers were presumably due both to

[7] By contrast, the studies counted in Table 1 for the symptom *speech rate* used a definition that was based on the total duration of each utterance.

stable individual differences and to random differences in the demands that were placed on the different callers.) Inclusion of this node in the network allows the system to learn about $\mathcal{U}$'s baseline rate in the course of a dialog so as gradually to become better at interpreting $\mathcal{U}$'s OBSERVED ARTICULATION RATE.

In sum, the potential diagnostic value of the variable OBSERVED ARTICULATION RATE lies in the tendency of speakers to slow their articulation by roughly 14% when subjected to fairly high cognitive load; but the diagnostic value may be diminished by other factors that influence articulation rate, such as individual baselines. So it is not obvious that this symptom can be of significant use for the assessment of a user's cognitive load. The next section will address this question with regard to both of the symptoms that we have discussed.

5 Assessing Potential Diagnostic Performance

Even if a network model is completely accurate, it may be of no use in practice for the modeling of individual users, because of the limitations of the available data in a dialog situation. As we have seen, the observable variables are at best noisy symptoms of the underlying variables of interest. Moreover, the number of relevant observations in a dialog may be small. To examine the data limitations in our example scenario with respect to the two symptoms discussed here, we performed the following steps:

1. Specification of the basic Bayesian network. We specified a Bayesian network with the structure shown in Figure 1 that fulfilled all of the constraints mentioned above.[8] To make possible a test simple enough to be discussed within the space limitations of this paper, we omitted all variables in the groups OTHER SYMPTOMS INVOLVING QUALITY REDUCTION and OTHER SYMPTOMS INVOLVING OUTPUT SLOWING. Moreover, the FACTORS DETERMINING PRIORITY FOR MAINTAINING SPEED were fixed at an intermediate level that reflected the assumption that users would attach roughly equal priority to output rate and output quality. We assume hypothetically for the rest of the analysis that this network is *entirely* accurate; in this way, problems arising from data limitations can be analyzed separately from those that are due to incorrect assumptions embodied in the network.

2. Definition of hypothetical users. We defined four groups of 15 hypothetical "users". Those in the first group were assumed to be experiencing somewhat below-average POTENTIAL WM LOAD (0.6 on our scale[9]); those in the fourth, very high POTENTIAL WM LOAD (1.8); and those in the second and third groups, intermediate levels (1.0 and 1.4, respectively). Within each of these groups, we defined 3 subgroups of 5 "users" with different levels of BASELINE ARTICULATION RATE: 6.75, 7.00, and 7.25 syllables/s, respectively. Recall that in our field study the average articulation rate of a speaker ranged from about 6.3 to about 7.7 syllables/s. Hence our hypothetical users do not include representatives of the extreme levels of BASELINE ARTICULATION RATE.

3. Generation of data for each user. For each such hypothetical user, we used the network to generate 10 "observations" of utterances that the user might produce in the course of a dialog. Each observation consisted of a pair of values, for the variables PRESENCE OF SENTENCE FRAGMENT and OBSERVED ARTICULATION RATE. For each user $\mathcal{U}$, we generated the observations by (a) instantiating the variables POTENTIAL WM LOAD and BASELINE ARTICULATION RATE according to the definition

[8] A machine-readable version of this example network is available from the authors.

[9] POTENTIAL WM LOAD is indexed on a scale from 0.0 to 2.0, where 1.0 corresponds to a load that the $\mathcal{U}$ in question could (just barely) handle without exhibiting any decline in the quality or speed of speech.

of that $\mathcal{U}$; (b) noting the network's resulting probability distributions for the two symptom variables; and (c) for each observation independently, using random numbers to generate values for these two variables on the basis of the probability distributions.[10] Since in our scenario about 40% of all utterances are too short to have their articulation rate measured meaningfully, for a random 40% of the utterances the OBSERVED ARTICULATION RATE was specified as "undefined".

4. Initialization of the network's prior beliefs. The network was then prepared to interpret the 10 observations of each user. For each user, we had the network start with the same a priori expectations about the unobservable variables POTENTIAL WM LOAD and BASELINE ARTICULATION RATE. These expectations corresponded to the actual distribution of these variables in the hypothetical sample (see above). In other words, we simulated a situation in which $\mathcal{S}$ has already accurately narrowed down its expectations with respect to these variables somewhat—a situation that could arise after the first few utterances in a dialog.

5. Interpretation of observations for each user. For each user independently, the 10 observations were interpreted one by one by the network, and $\mathcal{S}$'s assessments were updated accordingly. Figure 2 shows the development of $\mathcal{S}$'s assessments of the key variable POTENTIAL WM LOAD for the two groups of users with the lowest and highest actual levels of this variable, respectively.

On the positive side, we see that $\mathcal{S}$'s assessments do tend to move in the right direction: After 10 utterances there is hardly any overlap in the assessments for the two extreme groups.

At the same time, the results illustrate the reasons why a diagnostic network can fail to arrive at a precise and accurate assessment even when the data are completely consistent with its structure and probabilities:

1. The differences in the baseline articulation rates of the $\mathcal{U}$s tend to mask each $\mathcal{U}$'s actual cognitive load somewhat. In each graph, the slower-articulating $\mathcal{U}$s (represented by the gray lines) are assessed as suffering from greater cognitive load. In fact, this difference would persist even with a larger number of observations, until $\mathcal{S}$ encountered some observations that allowed a more precise assessment of BASELINE ARTICULATION RATE.

2. Because of the partly random variability in the data, $\mathcal{S}$'s assessment of a $\mathcal{U}$ often follows a zig-zag pattern instead of moving steadily toward the true value. In addition to the changes caused by the occasional sentence fragments (marked with a dot), this variability concerns the OBSERVED ARTICULATION RATE of individual utterances (not marked explicitly in the graphs).

3. Even in the whole sample of 10 utterances, a given $\mathcal{U}$'s speech may happen to show a pattern that is untypical of $\mathcal{U}$'s actual cognitive load, because of random variation (i.e., sampling error). For example, two of the $\mathcal{U}$s with low POTENTIAL WM LOAD happened to produce 3 sentence fragments in their 10 utterances, although the overall frequency of fragments even for the $\mathcal{U}$s with very high POTENTIAL WM LOAD is only about 10%.

6 Summary of Contributions and Current Work

The methodological contributions of this paper, in increasing order of novelty, are the following:

1. a way of synthesizing previously published experimental data so as to strengthen the empirical basis of a user modeling component;

[10] A similar method for generating hypothetical input data for a Bayesian network was applied, for example, by Henrion et al. (1996).

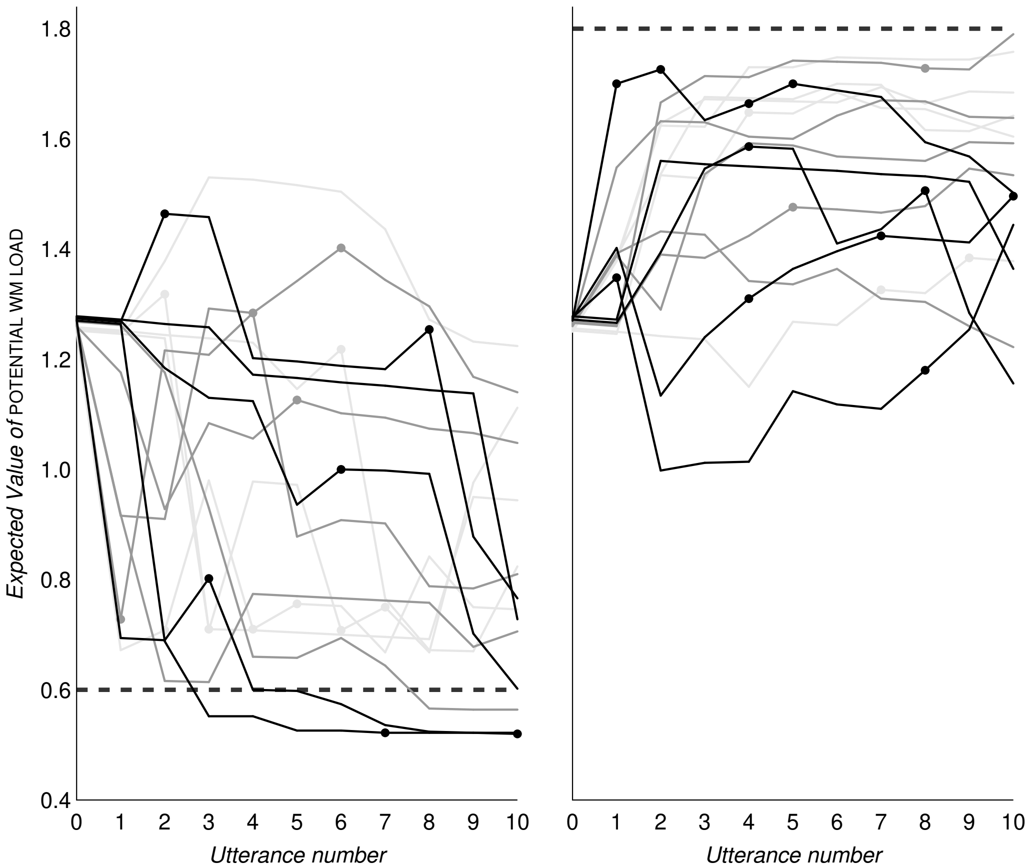

Figure 2. Test of the potential diagnostic utility of two symptoms of cognitive load. (Each jagged line traces, for one hypothetical user $\mathcal{U}$, the changes in the expected value of $\mathcal{S}$'s belief about $\mathcal{U}$'s POTENTIAL WM LOAD. The darker the line, the greater the BASELINE ARTICULATION RATE that was assumed for $\mathcal{U}$. Each observation that included a sentence fragment is marked with a dot. The horizontal dotted line in each graph shows the true value of POTENTIAL WM LOAD for the $\mathcal{U}$s in that graph.)

2. a way of combining such results with the results of detailed analyses of interactions in a given application domain so as to derive qualitative and quantitative constraints for a user modeling component; and
3. a method for analyzing the ways in which the diagnostic performance of a user modeling component is limited by the nature of the data available in the application scenario.

With regard to the particular problem of assessing cognitive load on the basis of speech input, the contributions are the following:

1. an overview of the most important indicators of cognitive load in speech input that have been identified so far;

2. a qualitative model of the relationships between these symptoms and theoretical variables which, though it requires specific testing, already has some degree of theoretical and empirical support;
3. an overview of the specific problems involved in the coding of sentence fragments and articulation rate;
4. several general trends concerning the diagnostic value of these two symptoms when realistically small amounts of input data are available.

The methodology is currently being applied to a different application scenario in which other interaction modalities in addition to speech are employed (Jameson, 1998). At the same time, our investigation of speech symptoms is continuing in the form of experiments whose data will be analyzed using learning algorithms for Bayesian networks with a view to arriving at a better empirical description of causal relationships such as those depicted in Figure 1.

References

Arnold, S. R. (1997). Wo das Perlhuhn trudelt. *Die Zeit* 33–34. 6 July 1997.

Baber, C., Mellor, B., Graham, R., Noyes, J. M., and Tunley, C. (1996). Workload and the use of automatic speech recognition: The effects of time and resource demands. *Speech Communication* 20:37–53.

Berthold, A. (1998). Repräsentation und Verarbeitung sprachlicher Indikatoren für kognitive Ressourcenbeschränkungen [Representation and processing of linguistic indicators of cognitive resource limitations]. Master's thesis, Department of Computer Science, University of Saarbrücken, Germany.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

Henrion, M., Pradhan, M., Favero, B. D., Huang, K., Provan, G., and O'Rourke, P. (1996). Why is diagnosis using belief networks insensitive to imprecision in probabilities? In Horvitz, E., and Jensen, F., eds., *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann. 307–314.

Jameson, A., Schäfer, R., Weis, T., Berthold, A., and Weyrath, T. (1999). Making systems sensitive to the user's time and working memory constraints. In Maybury, M. T., ed., *IUI99: International Conference on Intelligent User Interfaces*. New York: ACM. 79–86.

Jameson, A. (1996). Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling and User-Adapted Interaction* 5:193–251.

Jameson, A. (1998). Adapting to the user's time and working memory limitations: New directions of research. In Timm, U. J., and Rössel, M., eds., *ABIS-98, Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen*. Erlangen, Germany: FORWISS.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Roßnagel, C. (1995a). Kognitive Belastung und Hörerorientierung beim monologischen Instruieren [Cognitive load and listener-orientation in instruction monologs]. *Zeitschrift für Experimentelle Psychologie* 42:94–110.

Roßnagel, C. (1995b). Übung und Hörerorientierung beim monologischen Instruieren: Zur Differenzierung einer Grundannahme [Practice and listener-orientation in the delivery of instruction monologs: Differentiation of a basic assumption]. *Sprache & Kognition* 14:16–26.

Russell, S. J., and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice-Hall.

Schäfer, R., and Weyrath, T. (1997). Assessing temporally variable user properties with dynamic Bayesian networks. In Jameson, A., Paris, C., and Tasso, C., eds., *User Modeling: Proceedings of the Sixth International Conference, UM97*. Vienna, New York: Springer Wien New York. 377–388.