ABSTRACT
This position paper describes an approach to speech-driven user interfaces for automotive drivers. Drivers divide their attention between hands-free electronic devices and the necessity of driving safely. These devices are often controlled via spoken dialogue interfaces, which impose their own burden on driver attention. We argue for an approach that modulates the complexity of the user interface with ongoing driving conditions using psycholinguistic measures of language complexity, and we describe an overall system design in the context of research into language in divided-attention contexts. Our system design uses a given language complexity metric to calibrate the rate of syntactic and semantic information transfer between the driver and the synthetic speech interface in order to prioritise safe driving while permitting the driver to perform auxiliary tasks.

Categories and Subject Descriptors
J.7 [Computer Applications]: Computers in other systems—consumer products, real time; H.5.2 [Information systems]: Information interfaces and presentation—natural language

1. INTRODUCTION
Driving a car is a task that takes a non-trivial portion of the driver's attention and concentration, but technological advances, some now long-standing, permit drivers to perform both auxiliary and unrelated tasks. This presents challenges to the safety and effectiveness of driving to which legislative solutions are increasingly being applied: in particular, strongly enforced requirements for hands-free electronic devices while operating a vehicle.

An obvious solution that represents a compromise between the safety need for two-handed driving with eyes on the road and the driver's desire for increased productivity and control is the use of artificial spoken dialogue-based interfaces to present information and options. But these interfaces also themselves require some amount of attention from the driver, and hence pose potential impediments to the driver's response to road stimuli. Our challenge is to develop interfaces that relate the complexity of spoken system prompts to the amount of attention that the user can afford to pay to a non-driving task.

Recent results in experimental psycholinguistics suggest that we can indeed relate some component of attention directly to the features of the language being synthesised. There is an established and growing body of psycholinguistic research that provides numerical measures of linguistic complexity during incremental (or real-time) parsing derived from information theory and from the algorithmic complexity of hypothesised human parsers. There is evidence, from eye-tracking and other modes of experimentation, that these measures are correlated with some form of cognitive load.

To use these results in automotive UIs, we are establishing this research program: (a) study of the relationship between divided attention and different levels of linguistic representation (particularly semantic, syntactic, discourse); (b) design of a system to mediate in-car synthetic speech prompt production; and (c) measurement of the effects of in-vehicle dialogue system interaction.

We consider (a) in sections 2 and 3, wherein we introduce some relevant linguistic and psychological concepts and research. We illustrate (b) and (c) in section 4, wherein we sketch out a way forward for in-vehicle linguistic attention control. In section 5, we provide some concluding remarks that summarise our position on in-vehicle speech systems.

2. ATTENTION AND COMPREHENSION
There is some preliminary evidence for the effect of language on the ability to hold attention to a task from work on attention focus and switching. Experiments have shown that synthetic speech can impose different attention requirements on the user as well as different degrees of comprehension when compared to natural human speech. Delogu et al. (1998) showed that human subjects needed to pay more attention to synthetic speech in order to perform well on a comprehension task. We might hypothesise that we could account for this, for example, with reference to phonetic and phonological factors, but further experimentation appears to demonstrate otherwise. More recently, Swift et al. (2002) used eye-tracking measurements to show that there was a difference in the time it took to pay attention to an object
referred to in speech-synthesised instructions than it took in human-voiced instructions if the instruction used a definite article rather than an indefinite article. This suggests a more complex interaction between semantic factors and listener expectations in synthesised speech comprehension.

Both these research efforts used the same text stimulus via natural speech and synthetic speech. In applied contexts, however, the text is going to vary, and our effort is precisely to determine how much to vary it. If linguistic factors in synthetic speech do indeed have an effect on human attention switching, then by varying the generated language, we should be able to limit its deleterious effect on attention during in-vehicle multitasking. This requires that we be able to measure the cost of shifting attention between tasks, particularly when one task is a linguistic interaction task. Fortunately, there is work in unsynthesised speech settings that provides suggestive evidence for these costs.

Taube-Schiff and Segalowitz (2005) use an alternating-task experimental framework to show that changes in “grammaticized” words (conjunctions, prepositions, and so on) have an independent effect on the cognitive cost of shifting attention from one task to a different task1. Fang et al. (2009) use a language-directed 3D “treasure hunt” simulation with a conversational interface. They demonstrate that the distance between their subjects’ gaze fixations over time is mediated by the “smoothness” of discourse transitions from sentence to sentence, where discourse transition smoothness is defined in terms of Centering Theory (Grosz et al., 1995). Centering Theory ranks discourse “centres” partly by relative grammatical roles.

The cumulative effect of these levels of linguistic representation (phonological, syntactic, semantic, pragmatic) is potentially suggestive of an organising principle at work. Some of the sentence processing literature (Frank and Jaeger, 2008) points towards the existence of a limitation on human linguistic information processing. Human linguistic cognition attempts to hold the rate of information flow as close to constant as possible and deploys strategies within the various levels of representation in order to do so.

If that is the case, then we need to design systems that compensate by altering the information flow rate when there is a competing stimulus. Since the above literature has demonstrated that attention switching and focus are affected by linguistic information flow, we would expect these effects to be exacerbated when the other task (e.g. driving) is entirely unrelated. As the primary task consumes more driver attention, the quantity of linguistic information generated by the spoken UI should decrease. A more direct way of putting it would be to say that the interface should produce more concise, information-rich utterances when the driver’s attention is less taken by driving, and it should produce slower-paced utterances with information transfer more spread out over time when the driver needs to pay more attention to the traffic.

Then our challenge becomes one of finding the function that mediates this relationship.

3. COMPLEXITY AND COGNITIVE LOAD

The task of a dialogue system is to convey a certain set of facts to the user. However, there are various different ways for how the information can be formulated. User studies have shown that user satisfaction with dialogue systems is negatively correlated with task duration (Walker et al., 2001), suggesting that one should optimise for efficient communication. This is also supported by findings of Demberg et al. (2011), who compared two dialogue systems from the flight booking domain, one of which generated more complex linguistic structures, explicitly pointing out trade-offs between different options while the other generated less complex utterances. The more complex system led to more efficient user interaction, higher task success and improved user satisfaction when users could fully concentrate on the interaction with the dialogue system. When the two systems were evaluated in a dual task setting in a car, users however preferred the less complex system and also had a slightly increased number of minor driving errors (like crossing the line) when using the dialogue system that generated more complex utterances in the “difficult driving” condition (Hu et al., 2007). These findings call for a dialogue system that generates its utterances with suitable complexity in a given situation. In order to do this, we need to know which linguistic structures cause processing difficulty in humans.

A large body of research has shown that processing difficulty can occur at various linguistic levels, such as the structure of sentences, the semantic relationships between words and the discourse cues that link phrases or sentences.

3.1 Measures of sentence processing difficulty

One source of processing difficulty are unexpected linguistic events. A prominent measure of such processing difficulty is surprisal (Hale, 2001; Levy, 2008). Surprisal is defined as the negative log-probability of a word given its context. A highly-predictable word or sentence structure has low surprisal and causes little processing difficulty, while a word or
structure which is improbable given its context would cause increased processing difficulty. Surprisal can be calculated at various linguistic levels, such as word n-grams, syntactic structures\(^2\), or semantic representations.

Another factor that has been shown to affect human processing difficulty is the accessibility of previous material when integrating it with new material, as captured by dependency locality theory (DLT; Gibson, 2000). DLT measures the distance between grammatically dependent words in terms of intervening discourse-mediated items. Long distance dependencies thereby cause increased processing difficulty, in particular if people need to keep track of multiple open dependencies at the same time.

Demberg and Keller (2009) have recently developed a model of sentence processing called “Prediction Theory” which combines the expectation-based with the integration-based aspect of processing difficulty in a single theory.

3.2 Extensions to existing models
While computational models that assess processing difficulty related to syntax are available and have been tested (Demberg and Keller, 2008; Roark et al., 2009; Wu et al., 2010; Frank, 2010), there is significantly less work on how to integrate the difficulty measures with semantics (Mitchell et al., 2010), and, to the best of our knowledge, no psycholinguistic model that also includes the effect of discourse referents (like “even”, “although”, “therefore”). Evidence for the impact of such discourse referents on the effectiveness of human interaction with dialogue systems comes from a recent experiment by Winterboer et al. (2011), who found that the presence of discourse cues in dialogue system utterances helps users compare different options and select the optimal option (the setting was again interaction with a flight booking system). Another challenge thus lies in the development of a theory and model of discourse level processing difficulty.

4. COMPLEXITY MANAGEMENT
4.1 System design
The spoken dialogue interface is guided by the linguistic complexity metric (figure 1), some possibilities for which we described in section 3; it seeks to choose responses from its language generation system based on the magnitude of the metric. In turn, it evaluates the quality of user responses, which becomes a parameter of the metric. The quality of user responses is measured based on two overarching factors: speech fluency and task performance.

Similarly, driving behaviour and attentional requirements are also evaluated and factored into the complexity metric. A key insight is that only the spoken dialogue system can

be managed in order that it yield the appropriate amount of cognitive capacity to safe driving.

Initial parameters are set by data collected through experimentation. This last idea is crucial to the operation of the system. Most current applications of language technology are guided by machine learning algorithms based on text and speech corpora. We describe the conditions under which these must be collected in the following section.

4.2 Data collection and experimentation
In order to identify the situations and estimate the parameters under which linguistic complexity affects task performance, we need to construct a parallel corpus of recorded conversation and driving performance information. Then we can correlate the values of the complexity measures and see which ones are most predictive of either driving impairment or difficulty in performing the auxiliary speech task. As graphics technology has progressed, there are increasingly realistic and accessible driving simulator apparatus. It is possible to use these tools to measure performance on driving in terms of deviation from an ideal line in a course.

In fact, the major challenge in data collection is not the recording of driver activity, but the preparation of the linguistic data for statistical analysis. One such effort is Fors and Villing (2011); they describe a Swedish-language in-automotive data collection exercise for human-human dialogue. Their collection and analysis of Swedish conversation found that pause duration during conversations increased for the driver during periods of apparent high cognitive load, measured by the driver’s reaction time to a buzzer. Apart from the time and cost of transcribing the speech into text (something which automatic speech recognition may partially solve particularly for more widespread languages like English), Fors and Villing note that identifying the pause boundaries is very time consuming. Recent developments in data annotation such as crowdsourcing, however, now enable us to use unskilled Internet labour to annotate a variety of subtle distinctions (Sayeed et al., 2011); it is a matter of user interface design to present the task in such a way that errors and other features can be annotated by people who have no training in the task.

Once we have driving performance data and speech task performance data aligned over time for a sufficient number of subjects, it is then possible to explore the correlation between quantitative linguistic complexity and driving attention in context. This will allow us to set the parameters of the system in synthesising appropriate prompts.

5. CONCLUDING REMARKS
In this paper, we have brought together a number of elements that belong to a research agenda in in-vehicle spoken dialogue systems, and we have done so in a manner that makes psychologically plausible models of sentence processing relevant to automotive and other dialogue-based control and communication systems. A key component of this is attention and the effect of linguistic cognitive load. Multimodal tasks that involve communicating with an automated system while driving need to modulate the linguistic complexity of the interface with reference to current load and the need to retain a safe level attention on driving.
It happens that there is a large literature on attention and comprehension in the synthetic speech context and in the dialogue context. Different levels of linguistic representation affect or appear to be affected by tasks that demand attention; in particular, syntactic and semantic factors appear to have an effect on human performance in tasks that involve interaction with both synthetic and natural speech. There is also a large literature on computing psycholinguistic complexity and different levels of linguistic representation; divided-attention tasks such as driver multitasking permit this work to be leveraged in an applied context.

Given this scientific background, we have proposed an overall system design for in-vehicle spoken dialogue complexity management. We have also identified some of the design and data collection challenges that need to be overcome in pushing this agenda forward. As yet there is no publicly available corpus that relates transcribed speech data to user task performance and driving performance. However, the technical challenges to this, though significant, are not insurmountable. The final step is the design and construction of variable-complexity in-car dialogue systems.

References


