Linguistic Complexity and Cognitive Load in a Dual-Task Context

Nikolaos Engonopoulos

Universität des Saarlandes
Université de Lorraine

Erasmus Mundus MSc
in Language & Communication Technologies

MSc Thesis
supervised by
Dr. Vera Demberg and **Dr. Asad B. Sayeed**

Saarbrücken, September 2012

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, 28th September 2012

Nikolaos Engonopoulos

Abstract

In this work, we investigate the effect of linguistic complexity on cognitive load in a dual-task scenario, namely simultaneous driving and language use. To this end, we designed and implemented a psycholinguistic experiment where participants use a driving simulator while listening to spoken utterances and answering comprehension questions. On-line physiological measures of cognitive load, including the recently established Index of Cognitive Activity, as well as measures of performance in both tasks, have been collected. The resulting rich corpus of aligned fine-grained data streams can be used to test a vast array of different hypotheses about the relationship between performance, difficulty and cognitive load in dual tasks, at various levels of temporal resolution and linguistic structure. In this Master's thesis, we present the theoretical motivation and background, the methodological and technical aspects of the experiment, the resulting corpus and some first interesting results from the data analysis.

Acknowledgments

Although this thesis was the outcome of months of hard personal work of which I am proud, I would be ungrateful if I did not acknowledge the help and contributions of several people, without whom I would not be able to accomplish it.

First and foremost, I would like to thank from the bottom of my heart my two amazing, talented and hard-working supervisors, Dr. Vera Demberg and Dr. Asad Sayeed. This thesis would have never been possible without their constant guidance and active support, which is definitely not a given thing when it comes to a Master's thesis supervision. They assisted me in every step of this thesis, from high-level decisions to low-level details, including the creation of the linguistic stimuli for which my rudimentary German skills were not enough. Working with them and next to them has been a great pleasure and an incredible learning experience, and certainly the highlight of the 7 years I have spent in academia so far.

I would also like to thank my advisor from the University of Lorraine, Dr. Maxime Amblard, for his useful comments and advice and for his help during the first year of my Master's in Nancy.

I am grateful to the Erasmus Mundus LCT Consortium for the confidence they showed in me by funding my studies and for the unique opportunity they gave me to study in this great Master's program.

I would like to thank the Multimodal Computing and Interaction Cluster of Excellence (MMCI) for covering the costs of participant compensation and of part of the equipment used in the experiments.

Studying and working at Saarland University opens a huge window of opportunity for collaborating with a number of researchers from different fields; such collaborations were at the heart of this thesis. In particular, I would like to thank Dr. Christian Müller and the people from DFKI's automotive research group, for providing the driving simulator used in our experiments and for their helpful advice. Special thanks to Rafael Math for his constant support and quick replies to my questions and requests regarding the simulator. I would also like to thank Prof. Hubert Zimmer, Prof. Roland Brünken, Dr. Babette Park and Dr. Sarah Malone, for sharing their eye tracking experiences and for allowing us to access their lab and their copy of the EyeWorks software.

Many thanks go to Mike Bartels from EyeTracking, Inc. for his advice on how to use their software despite our non-standard setting.

I would also like to thank Prof. Andrew Kun from the University of New Hampshire for sharing his expertise with pupillometry and driving simulation experiments.

I would like to thank Prof. Matthew Crocker and all the people at the Psycholinguistics group of the Department of Computational Linguistics, for providing the lab, the eye tracker and most of the equipment used in this project and for offering useful feedback regarding my thesis project, when it was still in its infancy. Special thanks to Dr. Pirita Pyykkönen-Klauck, Fatemeh Torabi Asr and Evi Kiagia for helping me with my first steps in the art of eye tracking and for their useful tips and advice.

Many thanks to all the participants of the experiments, especially to the ones who kindly volunteered to invite others to participate, when I urgently needed to finish my experiments.

I would also like to thank my teachers, all the professors, researchers, staff of our department and my fellow students, who together make CoLi a great place to work, learn and do research.

Thanks to all my awesome friends in Saarbrücken, for being part of my life during this unforgettable year, for helping me cope with the stress and pressure of finishing this thesis, and for the lots of fun we had together. I hope our paths will cross again soon.

Thanks to my parents and to my sister; your love and constant support, during my whole life, brought me here in the first place and made me who I am.

Thanks to Angeliki, for being my test participant several times without complaining, for helping me in so many meaningful ways, and for making my life fun, beautiful and interesting.

Contents

CHAPTER 1

Introduction

1.1 Language in a dual-task context

A driver is listening to instructions from a navigation system while also talking with a passenger. An airline pilot is communicating with an air traffic controller. A worker is listening to radio. A hairdresser is talking with a customer. A commuter in a car is listening to a broadcast weather and traffic report. A technician is getting instructions from a supervisor. A taxi driver is receiving messages from a call center.

The cases above are just a few examples of a phenomenon which is ubiquitous in everyday life: *dual-task* scenarios, with language use as the secondary task. In such situations, people speak or listen to speech while engaging in a primary task where attention is critical to task performance, or in some cases to safety, thus making language use a potentially performance-degrading factor or a safety risk.

One particularly common primary task, where language is often used, is driving a vehicle. It is a cognitively demanding task, where constant attention and situation awareness, including fast processing of visual stimuli, is critical. It has been previously shown that engaging in secondary tasks involving language decreases driving performance and activity in driving-related areas of the brain (Just et al., 2008). One might argue that total abstinence from using language while performing safety-critical tasks would be a solution to the problem of language-caused distracted driving. For instance, several jurisdictions have banned the use of a hand-held mobile phone while driving, as a result of a number of studies a strong link between mobile phone use and increased car crashes in many countries (McCartt et al., 2006). According to a literature review study examining a vast range of secondary tasks (Young et al., 2007), there is evidence that virtually all non-driving tasks, including language-related ones, have a detrimental effect to driving performance and safety, although there are large differences in the magnitude of these effects across different tasks.

However, abstaining from all kinds of linguistic interaction while engaging in such a ubiquitous and time-consuming activity is arguably unrealistic. Furthermore, there are cases where accomplishing a linguistic task is helpful or even critical to primary task success, such as in the case of following navigation instructions or listening to weather reports. Even when the language task is not immediately related to the driving task, it can still have negligible or even positive effects on performance: for instance, listening to radio has been shown by some studies not to have a significant effect on primary task performance (Strayer and Johnston, 2001); under certain conditions, engaging in a secondary linguistic task can even have beneficial effects on driving performance (Atchley and Chan, 2011), for example by preventing under-stimulation and loss of vigilance, while conversation with a situation-aware passenger can direct the driver's attention to important objects or events on the road, thus reducing risks (Drews et al., 2004). Even in cases where the above mitigating circumstances do not hold, people still have important reasons to carry out language-related tasks efficiently, while keeping safety risks at a minimum.

It is indeed possible in many cases to achieve a balance between tasks, maintaining an acceptable level of engagement in the language task while still minimizing risks. For example, Crundall et al. (2005) show that in an experiment involving real driving, both drivers and passengers regulated their speech when faced with perceived difficult driving situations. Regulating speech might not be the only strategy to achieve the optimal balance; in a driving simulation experiment, Yannis et al. (2011) showed that drivers, when engaging in "simple" (as opposed to "complex") conversation, did not exhibit reduced reaction times or increased accident risk, which is hypothesized to be attributable to drivers compensating the conversation-caused distraction by reducing their speed.

1.2 In-vehicle spoken dialogue systems

An increasingly important category of secondary language tasks in which people tend to engage while driving is the use of automatic spoken language interfaces. Typical examples range from voice-controlled navigation systems to more general-purpose applications such as on-demand weather or traffic reports, to hands-free voice dialing or radio tuning, e-mail access systems, or automatic ticket/restaurant booking. It has been suggested that such voice-controlled interfaces might represent a better alternative to their manually controlled counterparts with respect to driving safety (Young et al., 2007). However, there is still evidence that using such a spoken interface remains a cognitively demanding task which interferes to some extent with driving performance (Lee et al., 2001).

The design of speech interfaces with minimal impact on driving safety is still an open engineering question (Hua and Ng, 2010). A possible direction to go in order to build such interfaces would be to learn from the strategies followed by their human counterparts, i.e. human situation-aware passengers. As we saw in the section 1.1, there is evidence that human speakers tend to modify the language they use depending on the real-time demands of the driving task. Building a dialogue system which would be able to use a similar self-regulation strategy would require at least two important components: a) a measure of difficulty of the driving task and b) the ability to modify the language used by the dialogue system in order to minimize the impact on the driver, depending on the detected difficulty of the driving task. A useful notion in the attempt to answer both of these questions is *cognitive load*, which can be thought of as a quantification of the demands imposed by mental tasks on the cognitive capacities of a person at a particular point in time. Although it is an abstract notion, several physiological measures have been proposed as metrics of cognitive load, such as pupillometric measures (measuring pupil dilations using eye trackers), including the recently introduced Index of Cognitive Activity (ICA, Marshall (2002)), and skin conductance level (SCL). We will elaborate on the notion of cognitive load and its existing metrics in Chapter 2.

In particular, it would be very useful for spoken dialogue system designers, if there were a way of knowing whether there are particular features of the language used which are more “difficult” to process, i.e. cause more cognitive load, or cause more interference with the driving task, depending on the driving difficulty. This would enable systems to use a continuous range of strategies, from using less and less demanding language, to pausing the dialogue altogether, depending on the amount of driving-induced cognitive load.

1.3 A role for linguistic complexity

In this attempt towards quantifying the cognitive load induced by language in dual tasks, another useful notion is the one of *linguistic complexity*, i.e. the amount of processing difficulty of language. In psycholinguistic research there has been extensive evidence that language comprehension and production does cause different amounts of difficulty on the language user, depending on the features of the language used, like prosody (Engelhardt et al., 2010) and structure (Demberg and Keller, 2008). As we will elaborate in Chapter 2, measures of linguistic complexity on various level of analysis (e.g. prosodic, syntactic, semantic) have been shown to correlate with measures used for measuring both generic and language-specific cognitive load. However, until now this relationship has primarily been assessed either in language-only experiments, or in dual-task scenarios where language comprehension is the primary task, while the secondary task is immediately related to language, e.g. "visual world" experiments (Tanenhaus et al., 1995). It is less clear what happens in dual-task situations where language is secondary and unrelated to the primary task. In particular, we would like to answer the following questions, among others:

- Does linguistic complexity still correlate with measures of cognitive load in a dual-task situation?

- Are changes in linguistic complexity capable of shifting attention away from the primary task or to cause measurable deterioration in primary task performance?

- Are these effects also measurable on a finer-grained temporal scale, such as during individual sentences, phrases or even words?

- Could we ultimately use this information in order to build dialogue systems which lower the complexity of their utterances on-line, according to the current difficulty or the driving task?

1.4 A novel dual-task experiment

The goal of this thesis project is to allow the investigation of the above formulated research questions. To achieve this, we designed and implemented a new dual-task experiment, manipulating both language and task difficulty and collecting fine-grained measurements of task performance, task difficulty and cognitive load. In our experiment, participants complete a driving simulation task, while simultaneously listening to spoken utterances and answering comprehension questions. The driving simulation task allows for different difficulty settings, while the linguistic stimuli are designed to allow comparison between linguistic syntactic structures which are known to cause different levels of processing difficulty in language-only settings. In addition, pupil size data are collected using an eye tracker, for calculating the Index of Cognitive Activity (ICA) and other pupillary measures of cognitive load, as well as skin conductance levels; these measures, to the best of our knowledge, have not been previously tried in detecting cognitive load caused by linguistic complexity. Moreover, exact timestamps of onset and duration are recorded for each of the words spoken, which makes it possible to pinpoint spikes in cognitive load caused by specific words or linguistic structures on a very fine-grained temporal scale. The outcome of these experiments is a rich annotated corpus of fine-grained measurements of task difficulty, task performance and cognitive load. This corpus allows for testing a large number of different hypotheses related to the interaction between driving, language and cognitive load in dual task scenarios across various temporal resolutions, as the initial results from our preliminary analyses already indicate, and may eventually contribute to the design of safer and more efficient spoken language interfaces.

1.5 Structure of the thesis

The rest of this document is structured as follows: we will provide some background on the notion of cognitive load, the measures used in this work and the particular issues related with language-induced cognitive load in Chapter 2; we will also present some of the dual-task literature related to our work in Chapter 3. The detailed presentation of our experimental setup and its technical aspects will follow in Chapter 4, while the resulting corpus and some initial results of the data analysis will be presented in Chapter 5. We will conclude and propose some future research directions in Chapter 6.

CHAPTER 2

Background

2.1 Cognitive load

In cognitive science and psychology, there has been a long-standing consensus, starting from the work of Miller (1956), that people have limited and measurable cognitive capacities for addressing mental tasks. The notion of *cognitive load* was introduced in an effort to quantify the strain induced by a mental task on these limited resources de Jong (2010).

A related finding is that performing one mental task interferes with the ability to perform other tasks (Wickens, 2002); this is especially important for multitasking situations, which are prevalent in real life. Dual-task scenarios such as simultaneous driving and participating in conversation, are often reproduced in experimental settings in order to investigate this kind of interference. There are many theories regarding the extent to which mental tasks interfere with each other and the exact nature of this interference, depending on the tasks in question.

One approach to answering the question of task interference and cognitive load is Multiple-resource theory (MRT (Wickens, 2008)). MRT hypothesizes four main dimensions of mental resources:

1. *stages* of processing (distinguishing between perceptual-cognitive vs action selection and execution resources)
2. *codes* of processing (distinguishing between spatial and verbal resources)
3. *modalities of perception* (distinguishing between auditory and visual processing, only for perceptual tasks)
4. *visual channels* (distinguishing between focal and ambient vision, only between visual tasks)

The main prediction of this theory, with respect to cognitive load, is that it expects increased interference when time-shared tasks have high demand for the same type of resources along some of these dimensions, resulting in mental overload and performance deterioration. On the other hand, two tasks which demand different types of resources along each of these dimension will exhibit "difficulty insensitivity" to a certain extent, i.e. an increase in difficulty of one task will not degrade the performance in the other. Performance will deteriorate when demand for one or both of these tasks along a particular dimension reaches an "overload" threshold. In the case of driving and linguistic comprehension, both tasks compete for perceptual resources along the stages of processing dimension, while they generally different resources along the other 3 dimensions. It would thus be expected that performance deterioration will result in cases of increased demand for the shared perceptual resources (e.g. particularly complex or surprising perceptual stimuli), while the deterioration will be even larger in case the linguistic task involves visual processing (e.g. reading the screen of an electronic device) or if understanding the content of the language involves elaborate spatial reasoning.

Cognitive load, like other notions in cognitive sciences, is an elusive concept which takes the shape of the measures used in an attempt to quantify it (Klingner, 2010). Various such measures of cognitive load have been proposed in the literature. One strong candidate is the P300 component of the event-related potential (ERP), measured via electroencephalography (EEG), whose amplitude has been shown to be sensitive to increased cognitive effort (Castro and Diaz, 2001). In particular, the P3b subcomponent of P300 has been shown to increase when improbable task-related stimuli occur, while the P3a ("novelty P3") increases with improbable non-task-related stimuli which cause involuntary attention shifts (Squires et al., 1975; Grillon et al., 1990). Therefore, under the assumption that the occurrence of less probable task-related stimuli is related to increased task difficulty, such as the sudden crossing of a child while driving, P300 can be used in practice as an indicator of higher task-related cognitive load.

To avoid the complications and costs of using EEG, less obtrusive psychophysiological proxies have also been proposed, which have been shown to vary with the occurrence of rare stimuli. In this study, we will focus on two of them: skin conductance level (SCL) and the pupillometric Index of Cognitive Activity (ICA) which measures the component of pupil dilation related to mental effort. We describe these measures in 2.1.1 and 2.1.2 respectively.

There are also other methods of quantifying cognitive load, apart from physiological proxies. Brunken et al. (2003) proposes a classification scheme for such methods along two dimensions: direct vs indirect and subjective vs objective. According to this scheme, subjective methods include self-reported mental effort (indirect) and self-reported stress or task difficulty (direct), while objective measures include physiological measures (indirect), brain imaging techniques (direct) and task performance (direct). In this work, we use physiological and task performance measures, which we will describe in more detail in chapters 4 and 5.

2.1.1 Skin Conductance Level

The sympathetic skin conductance level (SCL), also known as electrodermal activity (EDA (Dawson et al., 2000)) is a measure of the electrical conductance on the skin due to increased moisture level. Changes in moisture level are due to activity of the sweat glands, which are controlled by the sympathetic nervous system. Skin conductance amplitude usually changes with respect to its "neutral" (*tonic*) level in response to unexpected, significant or aversive stimuli. This change typically starts 1-2 seconds after the presented stimulus, with the peak level of skin conductance occurring a further 1-3 seconds after initiation, followed by a long period of returning to the normal amplitude (2-10 seconds for a 50% recovery to previous level). Such a spike in skin conductance level is also known as *phasic skin conductance response* or simply SCR. It has been shown, for instance, that SCRs are related to the preparation for action in reaction to critical vs non-criminal stimuli, in the "Concealed Information Test" used in criminal investigations (Matsuda et al., 2009).

SCL has also been used as a measure of cognitive load (Shi et al., 2007). In a dual task experiment with simulated driving and a secondary cognitive task, Mehler et al. (2009) found that skin conductance levels peaked in cases of mental overload caused by incrementally increasing secondary task difficulty, which was followed by a deterioration in the performance of the primary task. Son and Park (2011) found skin conductance levels, along with steering wheel reversals as a measure of task performance, to be good input features for an artificial neural network built to predict task difficulty, in comparison to other candidate features.

Skin conductance is a relatively unobtrusive measure, which is part of the reason why it is a widespread psychophysiological measure of cognitive and emotional load. A typical modern skin conductance sensor consists of two electrodes strapped on a person's fingers; the setup usually takes less than a minute. Its disadvantages include the rather slow response to critical stimuli and slow recovery rate, as well as the large differences in responses among participants and the occurrence of spontaneous spikes in skin conductance levels, known as "nonspecific SCRs" (NS-SCRs).

2.1.2 Pupillometry and the Index of Cognitive Activity

The second psychophysiological measure of cognitive load used in this study is pupillometry, i.e. the measuring of the size of the pupil. Pupillometry is a widely used measure of measure of cognitive load (Just et al., 2003; Engelhardt et al., 2010; Palinko et al., 2010). It has been shown that pupil dilation correlates with a pattern in the amplitude of P3 reflecting the locus coeruleus-noradrenergic system (LC-NE), which in turn is related to stress and attention (Laeng et al., 2010). In particular, pupil diameter has been shown to correlate with the firing of an LC neuron in monkeys.

Changes in pupil size induced by cognitive load typically amount to a difference of 20% relative to the typical pupil size. However, light conditions also affect pupil sizes, with brightness-induced changes being much larger than cognitively induced ones (up to 120% of typical pupil size) (Laeng et al., 2010). This creates a confounding effect when trying to measure pupil changes due to cognitive load in simulation environments, where objects differ in brightness between them. Palinko and Kun (2011) tried to address this problem in a static visual world scenario with eye tracking by developing a predictor of the brightness-induced changes in pupil size and subtracting it from the measured size, to obtain the cognitive load-induced pupil signal. However, their approach needs accurate eye-gaze information and brightness values for the areas of the screen where the participant is currently looking, which are quite difficult to obtain in a simulation setting, and unrealistic in the case of real driving experiments.

In response to the need for a pupillometric measure which would isolate the component of pupil dilations related to cognitive load, Marshall (2002) developed the *Index of Cognitive Activity* (ICA). This is a patented measure which applies fast Fourier transform to filter out slow, large light-induced changes and identify the occurrence of short, abrupt changes in pupil size which are caused by cognitive load. Its reported uses include measuring cognitive load in driving simulation tasks (Schwalm et al., 2008), distinguishing between different cognitive states in problem solving, simulated driving and visual search (Marshall, 2007) and distinguish between different levels of surgical skill (Richstone et al., 2010). ICA measurements have been shown to be relatively stable across several commonly used eye tracker models and using different sample rates ranging from 60 to 300 Hz[1] (Bartels and Marshall, 2012). To the best of our knowledge, ours is the first study to investigate the potential of the ICA as a measure of linguistically-induced cognitive load in a dual-task scenario.

2.2 Linguistic complexity and cognitive load

A central problem in both computational and experimental psycholinguistics is the quantification of the processing difficulty of language at various levels of linguistic structure. Several such measures of linguistic complexity have been introduced, with the goal of accounting for the differences in human language processing difficulty. In this section, we will present two measures stemming from distinct views on human language processing: on the one hand, probabilistic, information theoretical approaches; on the other hand, approaches based on computational modeling of memory and other resource availability constraints.

One measure of linguistic complexity which comes from the class of information theoretical views on linguistic processing is surprisal (Hale, 2001). From the perspective of incremental processing, surprisal is defined as the amount of information contributed by each word in an utterance, which in turn is the negative log-probability of a word w given its history h: $S(w) = -\log p(w|h)$. Different definitions of this history can be used, yielding different variations of surprisal (e.g. preceding n-grams, syntactic constituents, part-of-speech tags or even semantic arguments). The nature of this measure is such that it requires NLP methods to accurately estimate probabilities from large corpora, in order to approximate the linguistic knowledge (and subsequent expectations) of human speakers. (Roark et al., 2009) demonstrate a method to calculate lexical and syntactic surprisal using a broad-coverage probabilistic PCFG parser.

On the other hand, one theory that belongs to the class of memory-modeling approaches is Dependency Locality Theory (Gibson, 2000). This theory is based on a dependency approach to grammar and utilizes the notions of *storage cost* and *integration cost* to model processing difficulty. Storage cost refers to the

[1] We used 250 Hz in our study.

cost of memorizing multiple open-ended dependencies and increases with the number of such dependencies. Integration cost refers to the cost of incorporating new dependencies into existing structures and increases with the number of discourse referents between the two nodes of the dependency relation. Intuitively, DLT predicts that long-distance dependencies will usually be harder to process due to the large number of open dependencies and intervening discourse referents. More recently, (Demberg and Keller, 2009) have introduced Prediction Theory, which aims at unifying DLT with syntactic surprisal, by making use of a psycholinguistically-motivated version of tree-adjoining grammar (PLTAG).

When it comes to evaluating the psycholinguistic plausibility of such theories, various experimental methods have been proposed. Many of these methods are based on the measurement of reading times. The underlying hypothesis of using reading times as a correlate for processing difficulty is the 'eye-mind-link' which suggests that people fixate for longer at words where they are experiencing processing difficulty. (Tanenhaus et al., 2000). Self-paced reading is a common, low-cost method of measuring reading times: participants read each word of the sentence one at a time and hit a button when ready to move forward; the interval between button hits is considered as the reading time. However, this method is problematic as the participants usually cannot revisit previously appeared words. To remedy this problem, eye tracking is usually preferred; for example, Demberg and Keller (2008) show how reading times extracted from eye tracking data, using information such as fixation durations on text areas and total reading times, can be used to test the validity of theories of sentence processing such as surprisal theory and DLT. The problem with these methods is that they are specific to written texts and cannot be applied to spoken language comprehension. However, eye tracking in the broader sense still remains relevant in this case: "visual worlds" scenarios, for instance, where fixations on various visual objects are measured while the participant listens to spoken utterances, are commonly used to test theories of lexical access (Tanenhaus et al., 2000).

Closer to our work, pupillometry has been shown to be predictive of cognitive load, including load induced by linguistic comprehension. Just and Carpenter (1993) showed that pupil dilation (relative change to the mean) is a reliable indicator of increased sentence processing difficulty. Engelhardt et al. (2010) showed that pupil dilations reflected changes in linguistic difficulty resulting by manipulation of prosody and visual context. Standard pupillometry is however necessarily limited to simple visual tasks; as we explained in 2.1.2, relying on raw pupil sizes while engaging in tasks with varying luminosity, such as in a simulated driving context, is problematic due to the confounding effect of changes in pupil sizes induced by brightness differences. We thus use the ICA measure as a predictor of cognitive load, which to the best of our knowledge has not been previously explored as such in exploring the difference in cognitive load induced by different linguistic structures.

At this point, we have to note that although we do not use numerical measures of linguistic complexity in the analysis section (5.2) of this thesis, such measures can be readily applied to our corpus in order to obtain fine-grained complexity metrics, since we have recorded exact onsets and durations for each word heard during the linguistic task.

CHAPTER 3

Related Work

3.1 Cognitive load and task performance in dual tasks

There has been extensive research about the effect of language-related secondary tasks while driving, using a large variety of measures and experimental methodologies. In this chapter, we present some of these studies which are more relevant to our work.

Auditory language comprehension alone, without engaging in full dialogue, has been shown to decrease driving performance. Just et al. (2008) carried out a simulated driving experiment, with the secondary task being listening to utterances and judging whether they were true or false. A significant decrease in driving performance was found in the language condition; interestingly, the authors also used functional magnetic resonance imaging (fMRI) to find a decrease in brain activation related to spatial processing in the language comprehension condition, which indicates that auditory linguistic processing diverts brain resources from the driving task.

In another study, Palinko et al. (2010) conducted dual-task experiments, with participants simultaneously engaging in simulated driving and in two interleaved spoken dialogue-based games with a human passenger. They used pupil dilation metrics, namely mean pupil diameter change (MPDC) and mean pupil diameter change rate (MPDCR) as physiological measures of cognitive load. MPDC is the difference of pupil size from the mean over a certain time frame, while MPDCR is the average slope of the pupil size over the time frame. They found that driving performance was significantly different between the different language conditions. Their results also seem to indicate that pupillary measures are more sensitive to rapid changes of task difficulty within short time spans than driving performance measures, such as lane position variance or steering wheel angle.

However, there is evidence that there are significant differences in the ways in which different types of linguistic tasks affect driving performance. Lee et al. (2001) modified the "complexity" of a speech-based email system by manipulating the number of available options in the application, finding an increased cognitive load for driving while using the complex system vs the easy system, as well as using any version of the system compared to the driving-only condition. Yannis et al. (2011) also presented a driving simulation scenario, with a "simple" and "complex" dialogue condition; the conditions differed in the thematic content of the language task in each case (i.e. casual conversation as opposed to answering complex general-knowledge questions and mathematical puzzles); they found that participants exhibited significantly worse driving performance in the complex condition but not in the simple, when compared to the driving-only baseline.

3.2 Manipulating language task difficulty for reducing cognitive load

There is evidence that self-regulating speech in case of increased driving difficulty is a strategy employed by human speakers. Crundall et al. (2005), in real car experiments with simultaneous dialogue, demonstrated that in-vehicle dialogue with situation-aware passengers is significantly suppressed when compared to conversation with remote or blindfolded passengers. In defining the notion of "suppressed dialogue" they examined specific dialogue features, i.e. number of utterances, mean number of words per utterance and number of questions asked. There is also evidence that self-regulating strategies are indeed effective in minimizing the detrimental effects of dialogue in driving performance. Drews et al. (2004) carried out experiments with participants using a high-fidelity simulator and conversing with remote (via a hands-free device) vs in-vehicle passengers. Their results showed that when conversing with in-vehicle passengers, both drivers and passengers were able to regulate the dialogue when the driving situation demanded it, by lowering their speech rate; driving performance was also significantly better for in-vehicle vs remote dialogue. This work, in combination with Crundall et al. (2005), provides additional motivation for spoken interface designers to incorporate self-regulating mechanisms in their systems.

An interesting example of dialogue management according to detected cognitive load is the one presented by Villing (2009). She proposes a dialogue management system which uses different dialogue management strategies in case of detection of increased cognitive load, depending on whether this load is driving-induced or dialogue-induced. The increase in cognitive load is inferred indirectly through dysfluencies in the driver's speech, while a vehicle-state analyser is detecting situations which are known to be difficult for drivers; if such situations occur, the load is assumed to be driving-induced, otherwise it is assumed to have been caused by the dialogue. The different language strategies tested are to pause the dialogue until cognitive load is decreased, when the load is driving-induced, and to reformulate the utterance when it is dialogue-induced. The most important difference with our experimental approach is that she adopts a binary (high-low) scale to measure cognitive load, while our setup allows for much finer-grained psychophysiological measurements.

In dual task settings such as the above, the complexity of the language task itself is usually modeled in a very coarse-grained fashion, if at all. Drews et al. (2004) define a slightly more linguistically-oriented notion of "linguistic complexity" as the number of syllables per second; this is different to the type of structural linguistic complexity discussed in our work, as it is essentially a way to measure speech rate and does not concern syntactic or semantic features of the language used. Furthermore, a common feature of all the research works presented above is that, with the exception of Villing (2009), the effect of different language conditions is only examined at a coarse-grained level, for instance across different linguistic tasks. The novelty of our work is that the high temporal resolution of our measurements, in combination with the fact that we record the exact words spoken and their onsets, allows for higher precision in determining which features of the language are responsible for the high cognitive load. These features of our experimental corpus also allow for the application numerical linguistic complexity measures and determining their relationship with cognitive load and task performance across different complexity levels.

CHAPTER 4

Dual-task experiment

As described in the previous sections, our experiments involve language use in a dual-task setting. In particular, the participants have to complete a driving simulation task, while simultaneously accomplishing a speech comprehension task in German. Our final version of the corpus contains data from 24 participants, all native speakers of German. In the rest of this chapter, we will start by describing in detail the driving simulation task (4.1) and the language task (4.2); we will then continue with presenting the experimental script in 4.3, followed by a description of the technical aspects of the experiment, including description of the hardware and software components used, the laboratory layout and technical challenges in 4.4. In the last section (4.5) we provide information about the participants of the experiment.

4.1 Driving simulation task

A software-based steering simulation task on a desktop computer is used as a substitute for actual driving. The software is provided by DFKI [1]; in Figure 4.1 we provide a screenshot of the simulation environment.

The participant sees on a desktop computer screen a countryside road. The current position of the simulated vehicle is represented by a blue cylinder (*steering object*), which is allowed to move on a horizontal axis and which the participants control using a gaming steering wheel. A second, yellow cylinder (*target object*) is also moving on the same horizontal axis. At each given point in time, a random number generator "decides" if and for how long the target object will stay still, or randomly generates a new target position for the cylinder, towards which it starts moving at a pre-defined constant speed. The position of the

[1] `http://www.dfki.de/`

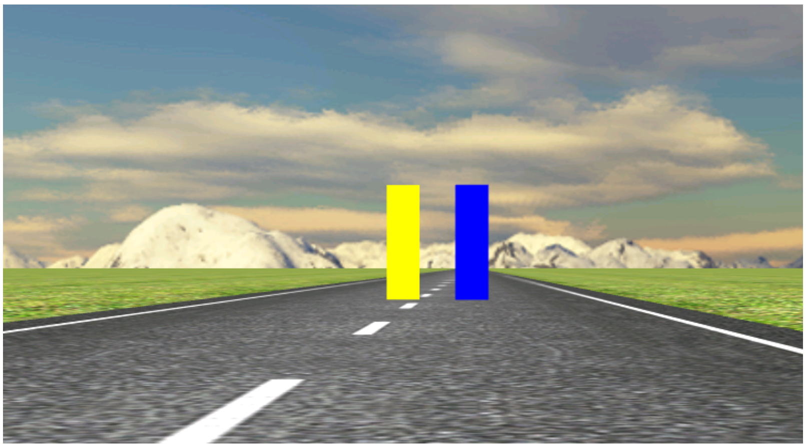

Figure 4.1: A screenshot of the modified OpenDS simulator used in the Experiments

steering object is determined proportionally to the angle of the driver's steering (although it cannot move further than the edges of the road). At the same time, the road and surroundings are moving at a constant speed, thus simulating the sensation of a vehicle moving forward. Except for the presence of the two moving cylinders, the background is very similar to a standard driving simulator; we have removed all potential distractions (signs, landmarks, speedometers) to minimize any confounding effects. For the duration of each recording session, the exact positions of the "vehicle bar" and the "target bar" are recorded at a sampling rate of about 26Hz and saved on a database.

The driver's goal in this task is to use the cover the target object with the steering one as much and as long as possible. As a result, our main measure of performance in this task is the *steering deviation*, i.e. the distance between the steering and target objects. We manipulate the difficulty of the driving task by changing the value of two variables: the speed at which the road and surroundings are moving and the maximum horizontal speed of the target object. We thus have two experimental conditions for the driving task: *Driving-easy* (De) and *Driving-difficult* (Dd). The two conditions induced significantly different levels of steering deviation on the participants on average, as we will show in 5.2.

4.2 Speech comprehension task

The spoken comprehension task consists in listening to a sentence containing a relative clause, followed by two thematically related 'filler' sentences and a comprehension question. The question is always polar (yes-no) and can be either directly related to the content of the relative clause (50% of the stimuli) or to the filler sentences. All sentences and questions are in German and are synthesized prior to the experiment using the MARY text-to-speech synthesis system (Schröder et al., 2008) developed by DFKI. The participants answer the question with *ja* ("yes") or *nein* ("no"). Their answer is recorded for later use in the data analysis.

All of our items, fillers and questions are in German. The items we are using are based on the ones created by Bader and Meng (1999). The filler sentences and questions were created by our research group, with the aim of providing a short, but coherent and narrative for each item, which would be as engaging as possible to the participant. These consist of 40 sentences containing a relative clause, each of which exists in two versions: one with an object relative clause (ORC), with high syntactic complexity, and one with subject relative clause (SRC), with lower syntactic complexity. The only difference between the object and subject relative clause is the number of the auxiliary verb form, which is singular ("hat"-$have_{3.SG}$) vs plural ("haben"-$have_{3.PL}$). The items are specifically designed to be grammatical and meaningful in both cases. For approximately half the stimuli, the comprehension question is in passive voice. An example of a stimuli pair, followed by the fillers and question, is the following:

> Die Lehrerin, die einige Eltern wegen einer solchen Kleinigkeit angerufen **[haben / hat]**, hat nun eine Elternversammlung einberufen. Diese Elternversammlung wurde schlecht geplant. Deshalb war die Teilnahme gering. Wurde die Lehrerin wegen einer solchen Kleinigkeit angerufen?
>
> *(The teacher$_{FEM}$ **[who called some parents / whom some parents called]** because of such a trivial issue, has now called a parents' meeting. This meeting was badly planned. Therefore participation was low. Was the teacher$_{FEM}$ called because of such a trivial issue?)*

As mentioned above, we synthesized all our stimuli using the MARY TTS system provided by DFKI. We manually checked all synthesized sound files for naturalness and corrected the synthesis in case of mispronunciations or erroneous intonation, either by providing the correct phonetic transcriptions to the TTS system or by giving specific prosody guidelines. One issue when comparing between different conditions in spoken stimuli is that different words have different spoken durations; in our case, the wordform "hat" has usually a shorter duration than "haben" when spoken. This might create problems when comparing physiological measures between the the conditions. We thus manipulated the duration of the pause marking the phrase boundary which immediately follows the auxiliary, to make sure that the critical regions in both the ORC and the SRC condition (defined as the time between the onset of the auxiliary to the onset of the next word) have exactly the same durations. This is important to allow for comparison of our task performance and cognitive load measures between the critical regions of different stimuli.

There were several reasons behind the choice to use synthesized speech, as opposed to pre-recorded natural speech. First of all, we wanted to be able to manipulate features of the produced speech, such as the duration of the critical regions, which would be impossible to do with a human speaker. Secondly, human speakers could potentially alter prosodic features of their speech depending on the amount of information carried by a word[2]; by using synthesized speech, we ensure that prosodic features of the utterances are almost identical, apart from the critical region. A further reason is that spoken interfaces rely more on synthesized speech than on pre-recorded utterances, especially as long as dialogue systems are moving towards a level of sophistication which allows automatic generation of arbitrary utterances. Since our ultimate goal is to eventually be able to improve such spoken interfaces, it is natural that we are primary interested in phenomena which can be detected in synthesized speech.

In order to assign the stimuli to the participants and determine the presentation order, we adopted the following scheme. The 80 total items are randomly split into two lists, A and B, under the constraint that each of those contains exactly 40 ORC and 40 SRC-type stimuli. This is to ensure that each participant will hear exactly one version of item. To ensure that each stimulus will be heard an equal number of times in the *Driving-easy* and *Driving-difficult* conditions, we further split A and B into two sublists, Ae, Ad and Be, Bd respectively, of 20 items each, such that each of those sublists has 10 SRC and 10 ORC. Therefore, half the participants are presented with 20 stimuli from Ae for *Driving-easy* and 20 from Ad for *Driving-difficult*; similarly the other half are assigned 20 stimuli from Be and 20 stimuli from Bd. The stimuli lists are randomized for each participants to minimize order effects. We ran experiments with 24 participants, thus obtaining 6 participants for each combination of item, clause type and driving difficulty condition.

[2] For instance, there is evidence that in conversational speech, words with a higher syntactic surprisal tend to have a longer duration (Demberg et al., 2012)

4.3 Experimental script

We will now describe in detail the experimental procedure followed during the experiments. At the beginning of the experiment, the participants fill in a consent form and read the instructions. After that, the experimenter attaches the skin conductance sensors to the middle and index fingers of the right hand of the participant, then places the eye tracker on the participant's head and performs the calibration. The installation and calibration of the measurement equipment takes 3-6 minutes on average, which is considerably less than for other methods of measuring cognitive load (e.g. EEG); this is a major advantage for an extensive experiment such as ours. The calibration is followed by a short training phase of around 3 minutes, which includes 1.5 minutes of driving on the easy setting without language, followed by 3 training items and items being played which are of similar construction but unrelated to our actual stimuli. Participants are given the option to continue training with the simulator after answering the final training question; in practice none of the participants made use of this option.

After training the main experiment starts. There are 4 recording phases, each of which lasts about 6 minutes. Each phase is composed of a driving-only phase of 2 minutes, followed by a driving + language phase of approximately 4 minutes, during which 10 of the items are played, each followed by the respective comprehension question. The participant speaks the answer, which is recorded by the experimenter using a response pad. In the first and the third phase, the driving difficulty is set to "easy", while in the second and fourth phase it is set to "difficult". Between the phases, participants are asked whether they want to take a short break. Most participants made use of this option once or twice during the experiments, while a few participants made no breaks at all. In case of a break, the eye tracker and skin conductance sensors are removed, while refreshments and snacks are also available. The equipment is then re-installed and re-calibrated. Even in case the participants do not choose the break option, the experimenter waits approximately 1 minute before moving on to the next phase, but without removing the measurement devices.

The full experimental script, with estimated durations for each phase, can be seen in figure 4.2.

Participant consent form, written instructions (2 min)

Eye tracker & SCL Setup, Calibration (3-5 min)

Training phase: De (1.5 min)

Training phase: De+L (1.5 min)

Recording Phase 1 (6 min)

- **1.1 Easy driving** (2 min)
- **1.2 Easy driving with language** (4 min)

Break, re-calibration (1-5 min)

Recording Phase 2 (6 min)

- **2.1. Difficult driving** (2 min)
- **2.2 Difficult driving with language** (4 min)

Break, re-calibration (1-5 min)

Recording Phase 3 (6 min)

- **3.1 Easy driving** (2 min)
- **3.2 Easy driving with language** (4 min)

Break, re-calibration (1-5 min)

Recording Phase 4 (6 min)

- **4.1 Difficult driving** (2 min)
- **4.2 Difficult driving with language** (4 min)

Questionnaire about participant information, payment (3 min)

Total Duration: approx. 45-55 min

Figure 4.2: Experimental script with approximate durations for each step.

4.4 Technical setup

4.4.1 Main hardware and software components

We will now describe the main hardware and software components of our experiments.

Hardware components

Host PC A Dell desktop computer, running MS-DOS, on which the EyeLink software is installed, which controls the EyeLink II eye tracker and collects eye samples.

Display PC A Dell desktop computer, running Microsoft Windows XP, on which the Experiment Builder software. It is connected to the Host PC via parallel port for communicating with the EyeLink software and guiding the sample recording and the calibration of the eye tracker.

"Simulation" PC A Dell desktop computer, running Microsoft Windows XP, connected to the Display PC via the parallel port. It is running the OpenDS simulator (see below), together with in-house developed software for collecting samples from the skin conductance sensors via Bluetooth and for sending parallel port signals to the Display PC for synchronization purposes.

EyeLink II eye tracker A head-mounted eye tracker produced by SR Research, with pupil size measuring capability and a maximum sampling rate of 500 Hz. We chose binocular sampling at 250 Hz, because of data format compatibility issues with the Eyeworks Analyze software we use for the ICA cognitive load measure (see below).

NeXus Skin Conductance Sensor A skin conductance sensor produced by Mind Media B.V., composed of two electrodes which are attached to the fingers of the participant. The sensor is connected to the **NeXus 10** physiological recording system, which is sending the collected samples to the Simulator computer via Bluetooth.

Driving Force GT A gaming steering wheel produced by Logitech.

Creative SoundBlaster Audigy 2 sound card An external sound card allowing 20 ms playback latency.

Dell Speakers 2 desktop loudspeakers with a subwoofer.

Samsung SyncMaster 245B monitor A 24-inch desktop monitor with input switch ability.

Cedrus RB-834 response pad A specialized response pad for coding participant answers.

Software components

OpenDS simulator A driving simulator provided by DFKI, modified for our purposes (see section 4.1). It records the position of the steering and target objects at the maximum possible frame rate of the graphics card (around 25 fps) and saves the recorded samples on a MySQL database on the same machine.

EyeLink software A software platform provided by SR Research for controlling the EyeLink II eye tracker.

Experiment Builder A visual experiment design tool developed by SR Research. It allows detailed control of the experiment flow, by exchanging signals with the EyeLink software on the Display PC, performing the calibration of the eye tracker and controlling the playback of the sound files. Signals and actions are triggered by events, which can either be

EyeWorks Analyze A software package provided by EyeTracking, Inc. which contains a module for calculating the patented Index of Cognitive Activity measure, using eye tracking samples as input.

MARY TTS An open-source text-to-speech synthesis system provided by DFKI, which we used to synthesize our stimuli and record word onsets in the sound files.

Apart from the above mentioned off-the-shelf software, we also developed our own source code for:

- reading the skin conductance samples from the Bluetooth receiver
- synchronizing the simulation by sending synchronization signals through the parallel port
- sending queries to the MARY TTS server process in order to synthesize our stimuli in batch and get the exact word onsets and durations

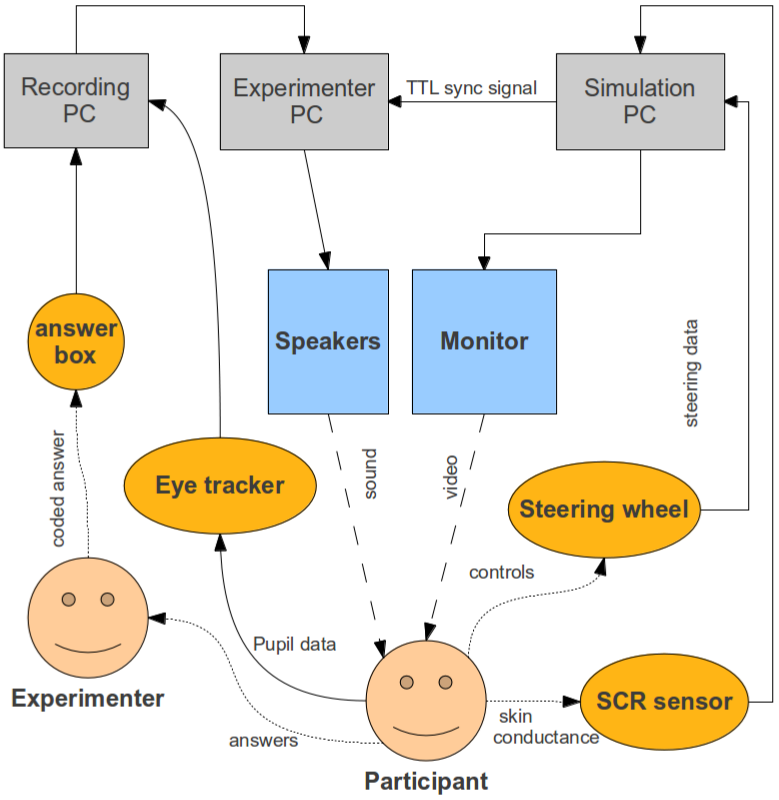

Figure 4.3: Hardware layout of our laboratory, including data streams during the experiments. Orange color = data collection device, Gray color = desktop computers, Blue color = output devices)

- aligning and post-processing the collected data from all different sources
- creating different versions of the data files depending on the desired granularity, aggregation method (e.g. mean), normalization method for physiological measures etc.

In the following section we describe the exact layout of these hardware and software components in the laboratory and how they interact during the experiment.

4.4.2 Hardware layout

In figure 4.3, we schematically illustrate the hardware layout and connections between the various components which are necessary to run our experiment. The participant gets input from the monitors (simulation video) and the speakers (linguistic stimuli), which are connected to the Simulation and Display PC respectively. The participant answers each question orally, and the experimenter codes the answers by pressing the respective button on the response pad which is connected to the Host PC. The skin conductance sensors are attached on the fingers of the participants and send recorded samples to the Simulation PC, while the eye tracker is attached to the head of the participant and sends data to the Host PC. The Simulation PC sends a signal via the parallel port to the Display PC at the start of each simulation phase, which allows for synchronization of the timestamps in both systems (also see 4.4.3). At the end of the experiment, the EyeLink data file, containing all eye samples and events recorded by the Host PC, is transferred to the Display PC.

4.4.3 Technical challenges

The implementation of the experiments described above presents important technical challenges, as one might expect from the sheer number of components and interactions between them. As is apparent in the configuration schema presented above, the experimental data is processed and stored in two different computers: data coming from the eye tracker and the response pad is processed in the Display PC, while data coming from the steering wheel and the SCL sensor is processed by the Simulation PC. Using two separate machines was necessary, given the high computational requirements of both the driving simulator and the Experiment Builder software. This creates the need for continuous synchronization of the timestamps of the two machines, in order to allow for accurate alignment of the collected data. Since high temporal precision is important, synchronization of the two data streams using more modern data ports (e.g. USB) was not appropriate due to the large transmission overhead, while other technical limitations precluded the usage of Network Time Protocol (NTP). For that reason, we opted for TTL (transistor-transistor logic) signaling via the parallel port, which is a much faster solution. A process on the Simulation PC sends a TTL signal to the Experimenter PC, while recording the exact system time which the signal was sent. The Experimenter PC then uses the TTL signal as a reference point in time for all the data which it records. The process is repeated at the start of each data collection phase to ensure precision.

Another problem is the head-mounted eye tracker losing track of the pupil due to excessive head movement of participants. Although the EyeLink II eye tracker allows for head movements up to 30° relative to the monitor, the fact that the participants have to use the steering wheel makes head movements more likely than in conventional eye tracking experiments. We addressed this problem by keeping each recording phase short (6 minutes) and re-calibrating the eye tracker in case there was substantial head movement between them.

Despite taking these measures, we still had to discard data from 6 experiments, for which there was substantial data loss for various reasons, such as when the participants wore too thick glasses, which caused high amount of light reflection on the glass and loss of the outline of the pupil. Problems with pupil detection also occurred in cases where participants wore cosmetic products (e.g. eyeliner) and at the same time exhibited a tendency to squint their eyes. Data loss also resulted in some cases from experimenter errors (e.g. an erroneous setting of the eye tracker or of the SCL device). In each of those cases, we recruited additional participants and repeated the experiments with the same stimuli lists.

4.5 Participants

The final corpus contains data from 24 participants, all of them native speakers of German, aged 20-34 years old. 10 of the participants were female and 14 were men. Participants were recruited via university mailing lists, printed advertisements on campus bulletin boards and through personal acquaintance. None of the participants had any knowledge of the exact research goal of the experiment, had any experience with the particular driving simulator or had heard any of the linguistic stimuli before. The majority of the participants were students or employees at Saarland University. All participants were compensated with the amount of 8 euros for their participation. As previously described in 4.4.3, data from 6 of the initial experiments had to be discarded and repeated with newly recruited participants, thus bringing the total number of people who participated in the experiment to 30.

CHAPTER 5

Results

5.1 Corpus description

The resulting corpus contains data from 24 experiments, with the total duration of the recorded samples summing up to about 12 hours. Table 5.1 describes in detail the variables contained in the corpus, along with their original source. A further categorization of the variables can be made into candidate measures of task difficulty (e.g. target acceleration for driving, relative clause type and voice of question for language), task performance (e.g. steering deviation for driving, answer accuracy for language) and physiological proxies for cognitive load. Anonymized participant information (e.g. age, gender) is also recorded. It is worth noting at this point that the list is by no means exhaustive as to what measures of task difficulty or performance can be used in future analyses of the data. For example, since we have the exact words played at each specific point in time, we can apply existing measures of linguistic complexity, e.g. syntactic surprisal or integration costs, as measures of difficulty of the language task. Similarly, we could calculate other measures of driving task performance (e.g. number of steering wheel reversals per time unit), since we have the exact positions of the steering and target object of the database for each recording sample. It is also possible to calculate other measures of cognitive load (e.g. mean pupil diameter change or blink rate (Palinko et al., 2010)), based on the raw eye tracker data. All of these additional measures can be extracted using the already existing corpus.

One issue when analyzing these data is the different sampling rates of the various data sources. To address that, we developed source code to first expand all data sources to the maximum sample rate (250 Hz) by repeating values when appropriate, then aggregate each of these measures over arbitrary pe-

Source	Variable	Type	Sample rate
Participant Information	ID	Categorical	1 per experiment
	Age	Integer	1 per experiment
	Gender	{f, m}	1 per experiment
	Occupation	Categorical	1 per experiment
	Videogame experience	{none, little, high}	1 per experiment
	Driving Licence	{yes, no}	1 per experiment
	Languages spoken	List	1 per experiment
Driving task	Steering Deviation	Real, non-negative	~26 Hz
	Target Velocity	Real, non-negative	~26 Hz
	Target Acceleration	Real	~26 Hz
	Difficulty Setting	{easy, difficult}	1 per experiment & recording phase
Linguistic task	Relative Clause Type	{subject, object}	1 per linguistic item
	Question Type	{related, unrelated}	1 per linguistic item
	Question Voice	{passive, active}	1 per linguistic item
	Correct answer	{yes, no}	1 per linguistic item
	Participant answer	{yes, no}	1 per linguistic item & participant
	Current word	String	250 Hz
	Current sound file	String	250 Hz
Physiological measures	SCL	Real	16 Hz
	Pupil Area (both eyes)	Real, Non-negative	250 Hz
	ICA event (both eyes)	{yes, no}	250 Hz
	Scaled ICA (both eyes)	Real, [0,1)	1 per aggregation unit
Time	Absolute time in ms	Integer	250 Hz
	Time in ms since start of recording phase	Integer	250 Hz

Table 5.1: Variables contained in the corpus

riods of time, depending on the type of analysis which is desired. In the initial analysis results presented in 5.2, we use aggregation over 1000 ms intervals for most types of regression analysis, except for the analysis within the critical region where we use aggregated values over 200 ms intervals. The method of calculation of the aggregated values depends on the type of each variable: for numerical values, we use the mean value of all samples within the time interval, while for categorical or binary value types (except ICA) we use the majority value.

In the particular case of the ICA measure, a different aggregation method is used: we created a scaled version of the variable, which is calculated for each time interval using the formula:

$$ICA_{scaled} = tanh(\frac{N}{E}) \tag{5.1}$$

where N is the number of detected ICA events within the time interval (i.e. the number of "yes" values of the ICA Event variable) and E is a constant integer representing the "typical" number of ICA events, for the part of the time interval for which there are available samples, i.e. excluding blinks or otherwise missing samples. We set the value of this constant to 30 ICA events for a fully sampled second[1]; the E, for an arbitrary interval where there are S recorded samples, can thus be calculated as

$$E = \frac{30 \times S}{SampleRate} \tag{5.2}$$

For instance, for a 30 available samples at a 250 Hz sample rate, the number of expected ICA events is $\frac{30 \times 30}{250} = 3.6$. The ratio of detected to expected ICA events is transformed to a $[0, 1)$ range via the hyperbolic tangent function.

5.1.1 Normalization of skin conductance levels

Tonic ("neutral") skin conductance values are known to vary across different people, or even across different psychological states of the same person (Dawson et al., 2000). It is therefore impossible to directly compare SCL measurements across participants, much less to average values between different experiments. In addition to that, our experience with pilot experiments indicated that neutral SCL values also varied depending on how tight the electrodes were

[1] This is also the value used by the EyeWorks ICA module

attached to the fingers of the participants. Since between phases the electrodes were usually removed and then re-attached, but remained attached for the whole duration of each phase, we assumed this factor to be constant for the duration each phase. We therefore normalized SCL values by subtracting the mean value of each phase (for a particular participant) and divided by three standard deviations. More sophisticated analyses which will take into account the specific behavior of the SCL measure, such as identifying skin conductance *responses* instead of relying on raw conductance levels, are also possible, since the raw SCL samples are also contained in the final corpus.

5.2 Data analysis

We will now present the results of some initial data analyses on our corpus. For running the analyses and producing the graphs shown below, we used the `R` mathematical suite. In particular, for the implementation of the regression models we used the `lme4` package for Linear Mixed Effect models (LME (Pinheiro and Bates, 2000)). LME models are a class of generalized regression models which allow for the inclusion of random as well as fixed effects. In our case, we used random effects in order to account for individual differences between participants.

In the rest of this section, we use the following star notation to indicate significance levels (i.e. probabilities of Type-I errors) obtained through t-test, unless indicated otherwise: *** = p<0.001, ** = p<0.01, * = p<0.05, no stars = no significance.

5.2.1 Correlation between physiological cognitive load measures

The first question we tried to answer was whether our main measures of cognitive load, namely ICA and skin conductance levels, correlate with each other. For this purpose, we built one mixed effects model for each eye's ICA measure, with the normalized SCL as the fixed effect, a random intercept per participant ID, to account for the individual differences in the ICA values between participants, as well as a random slope per participant ID, to account for the case where different participants exhibit different degrees of dependence between the two measures. The coefficients of the model are shown in table 5.2 (left eye ICA and SCL) and 5.3 (right eye ICA and SCL). We found a significant positive effect of SCL on the ICA measure for both eyes. The correlation between SCL

Predictor	Coef	t-value	Sig
INTERCEPT	0.781992	78.80	***
SCL_N	-0.01862	4.52	***

Table 5.2: Mixed effects model for left ICA with normalized SCL as a predictor; a random intercept and random slope was introduced for participants

Predictor	Coef	t-value	Sig
INTERCEPT	0.799327	77.39	***
SCR_N	0.016087	3.19	**

Table 5.3: LME model for right ICA with normalized SCL as a predictor and random intercept for participants

and ICA is also visualized in Figure 5.1, where we plot the measures against each other. The time interval for the aggregation in this analysis was 1000 ms.

One should be cautious about using raw (or even normalized) SCL values as a measure of cognitive load per se; as we described in section 2.1.1, the skin conductance response (SCR) to critical stimuli or mental overload is characterized by phasic activity, which appears as a peak in SCL occurring 2-6 seconds after perception of the stimulus, followed by a slow recovery towards the tonic level. The amplitude of peaks varies across participants, while the occurrence of non-significant SCRs is also an issue. Analysis of normalized skin conductance levels averaged across participants can only provide a very rough measure of the combined effect of frequency and amplitude of stimuli-related or non-significant SCRs across participants. That said, the regression analysis presented here is a useful sanity check for our measures, even if more sophisticated analyses, taking into account the nature of the SCL measure, are necessary for drawing any further conclusions about its potential to predict language-induced or driving-induced cognitive load in this particular context.

5.2.2 Driving performance and language

The next hypothesis we tested was whether our task performance measure in the driving task, i.e. the steering deviation, is sensitive not only to the driving task difficulty, but also to the presence of language. In figure 5.2 we have plotted the mean deviation for each of the difficulty settings (easy and difficult driving), with and without the secondary linguistic task. This figure illustrates an obvious difference between steering deviation in the easy and

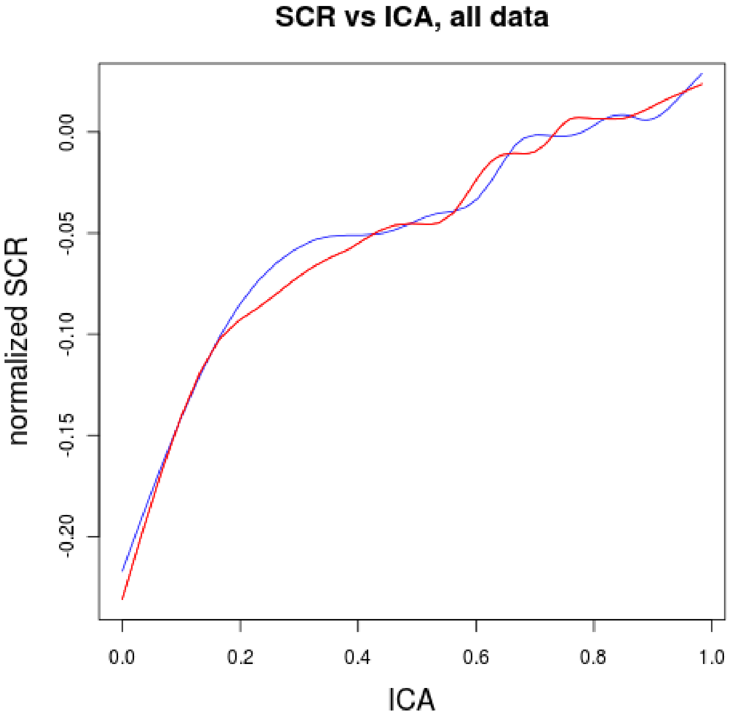

Figure 5.1: The ICA measure plotted against normalized SCL values for the same timestamps. Blue line = left eye ICA, Red line = right eye ICA.

Predictor	Coef	t-value	Sig
INTERCEPT	0.353917	32.50	***
PLAYING=YES	0.032568	3.79	***

Table 5.4: LME model for steering deviation, on the subset of the data with easy driving setting, with the presence of language as a predictor; a random intercept and random slope was introduced for participants

difficult driving condition. However, there are also smaller differences between the presence and absence of language. To test whether these differences are significant, we split the data between the easy and difficult settings and we built two LME models, one for each subset, with steering deviation as the response variable and the presence of language as the predictor; again, we introduced random intercept and random slope per participant to account for individual differences. We found out that there was indeed a significant positive effect of the presence of language on the steering deviation, which means that the driving performance worsens in the driving with language condition as opposed to simple driving, under both difficulty setting. The results of the regression analysis are presented in tables 5.4 and 5.5.

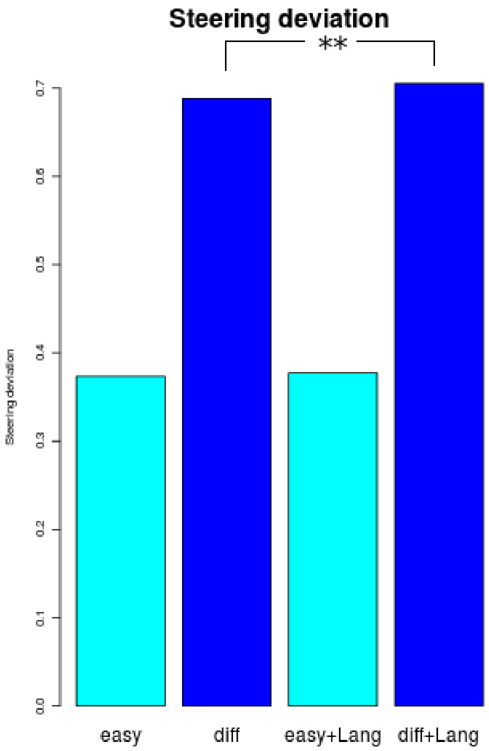

Figure 5.2: Mean steering deviation per difficulty setting, with and without language

Predictor	Coef	t-value	Sig
INTERCEPT	0.65903	35.20	***
PLAYING=YES	0.05823	3.20	**

Table 5.5: LME model for steering deviation, on the subset of the data with difficult driving setting, with the presence of language as a predictor; a random intercept and random slope was introduced for participants

5.2.3 Cognitive load and language task difficulty

One of the main novelties of our work is that we have used linguistic stimuli which differ in a *critical region* and are otherwise identical; since we have the exact word onsets, this allows us to isolate fine-grained changes in cognitive load which are related to a particular syntactic phenomenon. In particular, the critical region in each of our stimuli is defined as the duration between the onset of the auxiliary verb of the relative clause (*hat*/*haben*) and the onset of the next word. The form of the auxiliary differentiates between a subject relative clause and an object relative clause, the latter being shown to generally induce more processing difficulty in single-task psycholinguistic experiments (Bader and Meng, 1999).

Given this fact, we decided to look in more detail into cognitive load measurements within the critical region. In particular, we created a 200 ms mean-aggregated version of our corpus (see 5.1 for an explanation of the aggregated process) and isolated the subset of the data which fell within the duration of the critical region, which by definition includes the spoken word and the immediately following short pause (which marks the phrase boundary). The duration of this critical region is 650 ms in both conditions, which we imposed by manipulating the duration of the phrase boundary pause. On this subset of the data, we built two LME models (one for each eye) with the ICA measure as the response variable and the relative clause type as the fixed effect, while also introducing a random effect per participant.

The results of this analysis are shown in Table 5.6 for the left eye and in Table 5.7 for the right eye. We can see that there is a negative effect for the subject relative clause type in both cases, although only the result for the right eye is significant. The interpretation of the coefficient is that subject relative clauses tend to occur with smaller values of ICA than object relative clauses. This result is a first piece of evidence that the ICA measure might be sensitive to differences in linguistic complexity on a very fine-grained temporal scale. The fact that for the left eye the effect is less strong and does not reach significance, but is still in the same direction as for the right eye, may potentially mean that running the experiment with more participants could be useful in reducing noise and drawing more stable conclusions.

Predictor	Coef	t-value	Sig
INTERCEPT	0.76137	45.42	***
RC-TYPE (SUBJ)	-0.01862	-1.44	

Table 5.6: LME model for left ICA, with RC type as a fixed effect and random intercept per subject and random slope for RC type.

Predictor	Coef	t-value	Sig
INTERCEPT	0.78016	49.04	***
RC-TYPE (SUBJ)	-0.02651	-2.46	*

Table 5.7: LME model for right ICA, with RC type as a fixed effect and random intercept per subject and random slope for RC type.

5.2.4 Performance in the language task

One last link that we wanted to investigate was the one between performance in the linguistic task (i.e. answer accuracy) and the difficulty of driving and language tasks. Again, we built a LME model with the answer accuracy as the response factor and driving task difficulty, relative clause type and the voice (passive vs active) of the question as fixed effects, with a random intercept per participant and random slope for each of the effects. The resulting coefficients are presented in Table 5.8. Interestingly enough, no significant effect of the driving difficulty or the relative clause type was found; however, a significant negative effect on answer accuracy was found for passive voice, which means that there are significantly more wrong answers to passive voice questions than for active voice ones. The results indicate that linguistic comprehension, at least in this particular task, is robust enough to remain unaffected by driving difficulty and that grammatical voice is a more serious burden to language comprehension than relative clause type.

To shed more light on the ways in which grammatical voice affects answer accuracy, we plot in Figure 5.3 the frequency of types of errors made by par-

Predictor	Coef	t-value	Sig
INTERCEPT	1.83726	67.12	***
DRIVINGDIFFICULTY (EASY)	-0.02521	-1.05	
RC-TYPE (OBJ)	0.03554	0.98	
VOICE (PASSIVE)	-0.17278	-8.15	***

Table 5.8: LME model for answer accuracy explained by relative clause type

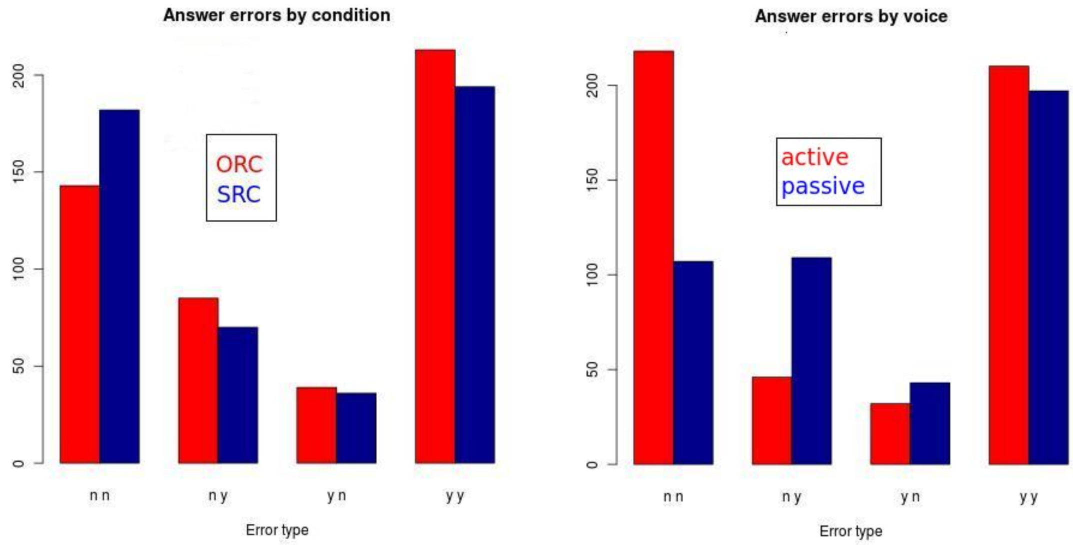

Figure 5.3: Plot showing the type of answer errors made by participants, depending on the relative clause type (left) and the grammatical voice of the question (right)

ticipants depending on the voice and the relative clause type. A first general remark is that participants generally made more errors of type no→yes than yes→no. As expected from the regression analysis, there are no big differences between error types in the subject or object relative clause condition. However, when we look at the error types by voice, it becomes apparent that the effect of voice almost exclusively influences questions which have a "no" correct answer. We hypothesize that this might be due to people relying more on lexical information when they failed to resolve the passive grammatical structure, thus answering "yes" whenever they hear overlapping lexical items in the question and the relative clause, resulting in more no→yes errors than yes→no.

CHAPTER 6

Conclusions

The goal of this work was to construct an experimental corpus which would allow the investigation of the interaction between fine-grained linguistic complexity, cognitive load and task difficulty in a dual-task context. In this thesis, we presented a novel dual-task experiment involving driving and language comprehension, and the resulting corpus which was specifically developed to address this goal.

We began this thesis by introducing the reader to the challenges presented by dual tasks involving language, with a particular focus on driving while using spoken dialogue interfaces and the associated effects on performance and driving safety. We proceeded by explaining how knowledge about language complexity would allow the design of better spoken language interfaces, which could balance linguistic task efficiency with driving safety and performance. With this motivation in mind, we provided background on two notions which are central to addressing our final research goals: cognitive load and linguistic complexity. We also presented two physiological measures of cognitive load used in this work, the pupillometric Index of Cognitive Activity (ICA) and the skin conductance level (SCL). We then placed our research in the context of recent dual task literature, which revealed crucial novel aspects of our work, among which our focus on structurally-defined linguistic complexity and their potential impact on cognitive load on a various range of temporal resolutions, down to the level of single words.

After providing sufficient motivation and theoretical background, we presented our new dual-task experiment, in which participants used a driving simulator while listening to synthesized utterances and answering polar comprehension questions. The utterances were designed in pairs to allow contrasting analysis between linguistic phenomena which are known in psycholinguistic research for causing different amounts of human processing difficulty.

The difficulty of both driving and language was manipulated, to allow the investigation of the interaction between different levels of difficulty of both tasks. We provided detailed information for both tasks, followed by a thorough presentation of the experimental script, the numerous hardware and software components used and the technical challenges associated with the experiment.

We then proceeded by presenting the resulting corpus, which contains online measurements of physiological data, i.e. pupil sizes and skin conductance levels, exact word onsets, measures of task difficulty and performance, for both language and driving. We also presented pre-processing and normalization steps, including the extraction of the scaled ICA measure.

Finally, we presented some initial results from an exploratory analysis of the data. We found a correlation between our two main measures of cognitive load on a large scale, as well as a significant effect of language on driving performance. Most importantly, though, we found a significant effect of a particular structural linguistic phenomenon on the ICA measure of the right eye, within a critical region of around 650 ms where the phenomenon occurred. This result may simultaneously constitute the first piece of evidence linking ICA to structural linguistic complexity at any scale and also the first result showing a fine-grained effect of a linguistic phenomenon on cognitive load in a dual task.

In sum, the results of the initial analysis add evidence to the potential of using our corpus for testing a multitude of hypotheses related to language and cognitive load. A few of these potential future uses of our corpus, as well as ideas for related future research directions are presented in the following section.

6.1 Future Work

As we already discussed in the previous section, and indicated in various points of this document, there are several ways in which the full potential of our experimental corpus can be exploited. A first step would be to try different combinations of predictors and response variables other than the ones we have already started showing in this thesis and test different hypotheses.

Further post-processing steps, such as more sophisticated normalization and noise elimination steps, might be necessary to obtain more reliable measures of cognitive load. It may be necessary to examine delayed effects of task difficulty on our physiological measures; this is especially the case with respect to SCL, where we have to take into account the particularities of the measure as reported in the literature in order to hope for more accurate cognitive load measurements. Our pupillometric measures of cognitive load do not have to be limited to the ICA; other metrics such as mean pupil diameter change rate (MPDCR) and blink rate can also be extracted from the raw eye tracker data which are part of the corpus.

A further immediate step, especially if we take into account the linguistic focus of our project, is to apply quantified measures of linguistic complexity, such as syntactic surprisal or integration costs, on the language used in our data. Since we have exact word onsets and durations, this would give us quantified, fine-grained measure of linguistic complexity which we could test as additional predictors of cognitive load. More sophisticated measures of difficulty can also be extracted for the driving task, such as steering wheel reversals or variance of steering intensity; this is possible since we record the exact position of the steering and target objects at a high sampling rate. We could finally examine the behavior of our measures with respect to individual participants, as well as the effect of various participant data, such as age, gender or videogame experience.

All of the research work mentioned above can be extracted from the readily available corpus. However, it may eventually be necessary to run the experiment with more participants, in order to either reproduce our results or obtain more statistical power and reduce noise effects. The detailed experimental script and our extensive related documentation will allow for faithful reproduction of the initial experimental conditions. It is also necessary to compare our findings with single task experiments with language only; such experiments are already underway by our research group.

Finally, once we have stable and reproducible results regarding the relationship between linguistic complexity, cognitive load and driving performance, we could design similar experiments for more realistic scenarios, namely using full-fledged dialogue systems, high-fidelity driving simulators or eventually real driving. Testing our research findings in such environments would bring us even closer to reaching our ultimate long-term goal: to use linguistic complexity to design better and safer automotive spoken interfaces.

Bibliography

Atchley, P. and Chan, M. (2011). Potential benefits and costs of concurrent task engagement to maintain vigilance a driving simulator investigation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(1):3–12.

Bader, M. and Meng, M. (1999). Subject-object ambiguities in german embedded clauses: An across-the-board comparison. *Journal of Psycholinguistic Research*, 28(2):121–143.

Bartels, M. and Marshall, S. (2012). Measuring cognitive workload across different eye tracking hardware platforms. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 161–164. ACM.

Brunken, R., Plass, J., and Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1):53–61.

Castro, A. and Diaz, F. (2001). Effect of the relevance and position of the target stimuli on p300 and reaction time. *International Journal of Psychophysiology*, 41(1):43–52.

Crundall, D., Bains, M., Chapman, P., and Underwood, G. (2005). Regulating conversation during driving: A problem for mobile telephones? *Transportation Research Part F*, 8(3):197–211.

Dawson, M., Schell, A., and Filion, D. (2000). The electrodermal system. *Handbook of psychophysiology*, 2:200–223.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, 38(2):105–134.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193 – 210.

Demberg, V. and Keller, F. (2009). Computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of* *the 31st Annual Conference of the Cognitive Science Society*, pages 1888–1893, Amsterdam. Cognitive Science Society.

Demberg, V., Sayeed, A., Gorinski, P., and Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Drews, F. A., Pasupathi, M., and Strayer, D. L. (2004). Passenger and cellphone conversations in simulated driving. In *PROCEEDINGS of the HUMAN FACTORS AND ERGONOMICS SOCIETY 48th ANNUAL MEETING*.

Engelhardt, P. E., Ferreira, F., and Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, 63:639–645.

Gibson, E. (2000). *Image, language, brain.*, chapter The dependency locality theory: A distance-based theory of linguistic complexity., pages 95–126. Miyashita, Y., Marantz, A., & O'Neil, W.

Grillon, C., Courchesne, E., Ameli, R., Elmasian, R., and Braff, D. (1990). Effects of rare non-target stimuli on brain electrophysiological activity and performance. *International Journal of Psychophysiology*, 9(3):257 – 267.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hua, Z. and Ng, W. (2010). Speech recognition interface design for in-vehicle system. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 29–33. ACM.

Just, M. and Carpenter, P. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(2):310.

Just, M., Carpenter, P., and Miyake, A. (2003). Neuroindices of cognitive workload: neuroimaging, pupillometric and event-related potential studies of brain work. *Theoretical Issues in Ergonomics Science*, 4(1-2):56–88.

Just, M., Keller, T., and Cynkar, J. (2008). A decrease in brain activation associated with driving when listening to someone speak. *Brain Research*.

Klingner, J. (2010). *Measuring Cognitive Load During Visual Tasks by Combining Pupillometry and Eye Tracking*. PhD thesis, Stanford University Computer Science Department.

Laeng, B., Sirois, S., and Gredeback, G. (2010). Pupillometry - a window to the preconscious? *Perspectives on Psychological Science*, 7(1):18–27.

Lee, J., Caven, B., Haake, S., and Brown, T. (2001). Speech-based interaction with in-vehicle computers: the effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors*, 43(4):631–640.

Marshall, S. (2002). The index of cognitive activity: measuring cognitive workload. In *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants, 2002.*

Marshall, S. (2007). Identifying cognitive state from eye metrics. *Aviation, space, and environmental medicine*, 78(Supplement 1):B165–B175.

Matsuda, I., Nittono, H., Hirota, A., Ogawa, T., and Takasawa, N. (2009). Event-related brain potentials during the standard autonomic-based concealed information test. *International Journal of Psychophysiology*, 74(1):58 – 68.

McCartt, A., Hellinga, L., and Bratiman, K. (2006). Cell phones and driving: review of research. *Traffic Injury Prevention*, 7(2):89–106.

Mehler, B., Reimer, B., Coughlin, J., and Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, 2138(-1):6–12.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2).

Palinko, O. and Kun, A. (2011). Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. *Driving Assessment*.

Palinko, O., Kun, A., Shyrokov, A., and Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *ETRA*.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Statistics and computing series. Springer-Verlag.

Richstone, L., Schwartz, M., Seideman, C., Cadeddu, J., Marshall, S., and Kavoussi, L. (2010). Eye metrics as an objective assessment of surgical skill. *Annals of surgery*, 252(1):177–182.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Schröder, M., Charfuelan, M., Pammi, S., and Türk, O. (2008). The MARY TTS entry in the Blizzard Challenge 2008. In *Proc. Blizzard Challenge*. Citeseer.

Schwalm, M., Keinath, A., and Zimmer, H. (2008). Pupillometry as a method for measuring mental workload within a simulated driving task. *Human Factors for assistance and automation*, pages 1–13.

Shi, Y., Ruiz, N., Taib, R., Choi, E., and Chen, F. (2007). Galvanic skin response (gsr) as an index of cognitive load. In *CHI '07 extended abstracts on Human factors in computing systems*, CHI EA '07, pages 2651–2656, New York, NY, USA. ACM.

Son, J. and Park, M. (2011). Estimating cognitive load complexity using performance and physiological data in a driving simulator. In *AutomotiveUI'11, November 29-December 2, 2011, Salzburg, Austria Adjunct Proceedings*.

Squires, N. K., Squires, K. K., and Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology*, 38:387–401.

Strayer, D. and Johnston, W. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological science*, 12(6):462–466.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., and Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, 29:557–580.

Villing, J. (2009). In-vehicle dialogue management - towards distinguishing between different types of workload. In *SIGDIAL Workshop on Discourse and Dialogue (SIGDIAL 2009)*.

Wickens, C. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2):159–177.

Wickens, C. D. (2008). Multiple resources and mental workload. *HUMAN FACTORS*, 50:449–455.

Yannis, G., Papadimitriou, E., Bairamis, C., and Sklias, V. (2011). Is it risky to talk, eat or smoke while driving? findings from a driving simulator experiment. In *Proceedings of the 3rd International Conference on Road Safety and Simulation, Indianapolis, USA, September 2011*.

Young, K., Regan, M., and Hammer, M. (2007). Driver distraction: A review of the literature. *Distracted driving. Sydney, NSW: Australasian College of Road Safety*, pages 379–405.