# Measuring linguistically-induced cognitive load during driving using the ConTRe task

Vera Demberg
Cluster of Excellence, MMCI
Saarland University
Saarbrücken, Germany
vera@coli.uni-saarland.de

Asad Sayeed
Cluster of Excellence, MMCI
Saarland University
Saarbrücken, Germany
asayeed@coli.uni-saarland.de

Angela Mahr
German Research Center for Artificial Intelligence
Saarbrücken, Germany
angela.mahr@dfki.de

Christian Müller
German Research Center for Artificial Intelligence
Saarbrücken, Germany
christian.mueller@dfki.de

## ABSTRACT

This paper shows that fine-grained linguistic complexity has measurable effects on cognitive load with consequences for the design of in-car spoken dialogue systems. We used synthesized German sentences with grammatical ambiguities to test the additional workload caused by human sentence processing during driving. For the driving task, we used the Continuous Tracking and Reaction (ConTRe) task, which we believe is suitable for the measurement of the fine-grained effects of linguistically-related workload phenomena in automotive environments, as it provides millisecond-level driving deviation measurements on a continuous course. We applied the task in an eye-tracking environment, using a pupillometric measure of cognitive workload called the Index of Cognitive Activity (ICA).

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

## General Terms

Methods

## Keywords

simulated driving, index of cognitive activity, pupillometry, cognitive load, tracking task, steering, language processing, relative clause, ambiguity


*AutomotiveUI '13,* October 28 - 30 2013, Eindhoven, Netherlands

http://dx.doi.org/10.1145/2516540.2516546

## 1. INTRODUCTION

As the development of hands-free and in-car user interfaces continues apace, engineers and regulators are increasingly confronted with the problem of assessing the burden of attention of these systems in the context of driver multitasking. As these newer technologies become more capable of performing complex user tasks, the need for a way to design interfaces that actively manage the level of cognitive workload they require increases [5].

A prerequisite to designing such interfaces is to be able to measure the effect of secondary tasks on the indispensable primary driving task. Examples of secondary tasks involving processing and responding to complex information include driving directions, restaurant and flight bookings, calendar management, etc. The rationale behind this is the observation that systems increasingly rely on vocal/auditory interaction (e.g. Apple's Siri), which is particularly significant in the driving environment where the primary task already consumes the user's visual resources.

In this work, we used synthesized German sentences with grammatical ambiguities to test the additional workload due to human sentence processing during driving. We show that fine-grained linguistic complexity has measurable effects on cognitive load with consequences for the design of such in-car spoken dialogue systems.

For our driving task, we used the Continuous Tracking and Reaction (ConTRe) task [15], which we believe is suitable for the measurement of the fine-grained effects of linguistically-related workload phenomena in automotive environments as it provides millisecond-level driving deviation measurements on a continuous course. We applied the task in an eye-tracking environment, using a pupillometric measure of cognitive workload called the Index of Cognitive Activity (ICA) [16], which is a measure derived from pupil diameter that has been shown to reflect changes in mental workload at a sub-second level. Our experiment shows a relationship between ICA level, driving difficulty, and driving performance, the presence of a language task having a tandem effect on both steering deviation and ICA.

## 2. THE CONTRE TASK

Driving tasks in simulators need to be selected according to the specific requirements of the experiment at hand. Depending on the underlying scientific or engineering questions, they need to be either more realistic or more controlled. For example, if we want to measure the driver's strategy of avoiding traffic jams on her daily commute to work, we need to design a highly realistic scene in the exact city or area. However, such a task comes with so many degrees of freedom that fine-grained measurement of performance is rather difficult. The ConTRe task [15] is designed to be on the opposite side of this spectrum: it is highly controlled and therefore less natural. ConTRe is derived from two well-known psychological tasks which have been used in numerous dual-task experiments: 1) the tracking task (also "perceptuomotor tracking task"), letting the subjects trace a moving object on the computer screen with a pointing device and continuously measuring the deviation (distance of pointer to object); 2) the reaction task, requiring the subjects to react on discrete visual stimuli by pressing a button [23]. Tracking is realized as steering in ConTRe. Turning the steering wheel simultaneously influences the lateral position of the vehicle/viewpoint and the lateral position of a blue vertical bar, the 'steering bar' (pointer). This bar is located at a constant longitudinal distance of 30 meters ahead and will always be at the same lateral position as the driver's point of view, i.e. at the center of the screen. A second vertical bar, the 'reference bar' (object), is colored yellow and is located at the same longitudinal distance as the steering bar. The reference bar moves autonomously to random lateral positions on the road holding there for 2 seconds before moving on.

The movements of the reference bar are neither controlled nor predictable by the user. The only exception is that lateral movement of both bars is limited by the solid side markings of the traveled way, which lie 8 meters apart. This prevents the driver from leaving the road. The driver's task is to control the lateral position of the blue bar via the steering wheel, keeping it overlapping with the reference bar as much as possible. As the maximum speed of the steering bar is about twice as high as that of the reference bar, half of the maximum steering wheel displacement is sufficient to keep the steering bar following the reference bar at the same speed. In ConTRe, reaction is furthermore implemented as a signal light placed on top of the reference bar, showing one of two lights from time to time. The upper red light requires an immediate brake reaction with the brake pedal, whereas the lower green light indicates that an immediate accelerator pedal action is necessary. However, in the experiment reported here, only the steering part of the task was used, as we will explain in Section 3.

The car moves autonomously at a constant speed along a predefined route on a unidirectional straight road consisting of two lanes per direction. Figure 1 shows the setup of the ConTRe steering task used in our experiment. Even though motion in the ConTRe task feels rather like a video clip, this task of continuous manual control based on continuous visual input and pedal responses upon discrete visual events effectively corresponds to a task where the user has to follow a curvy road. Even more important, it is absolutely controlled and leads to less user-dependent variability in the interpretation of instructions. Furthermore, the task allows for manipulation of task difficulty at run time. For this experiment, we created two difficulty settings (easy, difficult, for parameters, see Table 1. During the task, deviation from perfect bar overlap (blue bar covers yellow bar) and other relevant data are continuously recorded in a database.

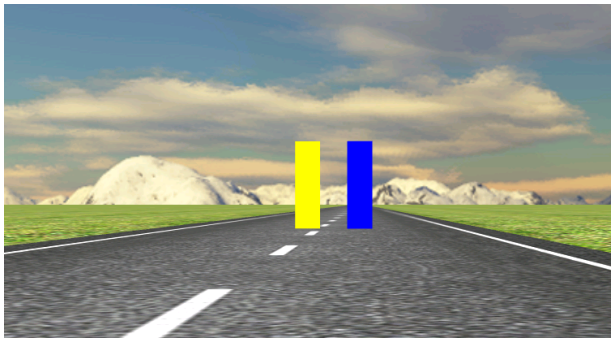


Figure 1: Screen shot of the driving task scene (ConTRe steering only).

| | easy | difficult |
|---|---|---|
| speed of reference bar (m/sec) | 1.0 | 2.5 |
| max. speed of steering bar (m/sec) | 2.0 | 4.0 |
| longitudinal speed (km/h) | 40 | 70 |

Table 1: Levels of difficulty

The ConTRe task is a driving task "plug-in" for OpenDS[1].

## 3. EXPERIMENTAL DESIGN

Our experiments involved language use in a dual-task setting. The participants were required to complete a driving simulation task, described above, while simultaneously accomplishing a speech comprehension task in German. Our experimental results contain data from 24 participants (aged 20-34; 10 female; all native speakers of German).

The spoken comprehension task consisted of listening to a sentence containing a relative clause, followed by two thematically related 'filler' sentences and a comprehension question, which we asked in order to make sure that participants were listing to our stimuli. The question was always polar (yes-no) and could be either directly related to the content of the relative clause (50% of the stimuli) or to the filler sentences. All sentences and questions were in German and were synthesized prior to the experiment using the MARY text-to-speech synthesis system [20]. We used synthesized speech (as opposed to recorded natural speech) in order to manipulate features of the produced speech, e.g., the duration of the critical regions, which would be impossible to do with a human speaker.

At the beginning of the experiment, the participants filled in a consent form and read the instructions. After that, the experimenter placed the eye tracker on the participant's head and performed the required calibration. The calibration was followed by a short training phase of around 3 minutes, which included 1.5 minutes of driving on the easy setting without language, followed by 3 training items and items which were of similar construction but unrelated to our actual stimuli.

After training, the main experiment started. There were

[1]Download available at http://www.opends.eu

4 recording phases, each of which lasted about 6 minutes. Each phase was 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 were played, each followed by the respective comprehension question. Participant answers were recorded by the experimenter using a response pad. In the first and the third phase, the driving difficulty was set to "easy", while in the second and fourth phase it was set to "difficult".

We used German locally ambiguous subject relative clauses (SRC) vs. object relative clauses (ORC) based on the materials by [2]. The object relative clause is known to be much harder to process than the subject relative clause. In (1), we see one example of our materials.

(1) *Die Nachbarin, [**die**$_{sg,\ nom/acc}$ einige$_{pl,\ nom/acc}$ der Mieter auf Schadensersatz verklagt **hat**$_{sg}$/ **haben**$_{pl}$]$_{RC}$, traf sich gestern mit Angelika.*
"The neighbor, [whom some of the tenants sued for damages / who sued some of the tenants for damages]$_{RC}$, met Angelika yesterday."

When reading such a sentence, people will usually interpret the relative pronoun *die* as the subject of the relative clause, and the following noun phrase *einige der Mieter* as the object. This interpretation is compatible with the embedded verb *hat*, which has singular marking, at the end of the relative clause. If, however, participants encounter the verb *haben*, which has plural marking, they will encounter some processing difficulty: in order to make sense of the relative clause, they need to re-interpret the relative pronoun *die* as the object of the relative clause and the following noun phrase *einige der Mieter* as its subject; such a reinterpretation is known to cause difficulties. (Note that the sentences are all grammatical, as the nouns are chosen such that they are ambiguous between nominative and accusative case marking.)

# 4. BEHAVIORAL MEASURES

Our first behavioral measure, steering deviation, directly reflects the performance on the ConTRe task. Steering deviation was calculated as the distance between the reference bar and the steering bar. We furthermore calculated derivative measures such as steering deviation acceleration, i.e., how quickly steering deviation increases or decreases.

During the driving task, we collected pupil size measurements using the head-mounted EyeLink II eyetracker at 250Hz on both eyes. From these recordings, we could calculate changes in overall pupil dilation (large pupil dilation is known to be associated with cognitive load, see e.g., [10, 9]), blink rate (cognitive load has been related to more blinks, see e.g., [18]), as well as the frequency of rapid small dilations of the pupil (Index of Cognitive Activity, ICA). The ICA[2] has been suggested as a robust and fast measure of cognitive load which has previously been evaluated on a small range of tasks including digit span tasks, language comprehension tasks and a simulated driving task [17, 4]. Compared to pupil dilation, the ICA has the advantage of being able to disentangle activation and inhibition patterns for reaction to light and reaction to cognitive activity.

[2]The method is patented, and the analysis program has to be licensed from EyeTracking, Inc., San Diego, CA. For details see [16].

# 5. DATA ANALYSIS METHODS

We calculated time-series analyses, spline models, and linear mixed effects regression models using R (packages mgcv and lme4). The data streams collected during the experiment were time series of different measurements at some interval (e.g., every 100msec). *Autocorrelation* (AC) refers to the correlation of a time series with its own past and future values, revealing whether it changes very dynamically (low AC) or is "persistent" (high AC). AC analyses can also reveal periodicity in a time series.

We can also measure how one time series is correlated with another time series by shifting their alignment by increasing intervals. This analysis is also referred to as the *cross-correlation* (CC) of two time-series. CC analyses are particularly useful for exploratory analyses, for example, when we do not yet know at what latency to expect participants' steering movements as a reaction to the reference bar or at what latency to expect an effect on other cognitive measures such as the pupil size or ICA measure (see Section 4).

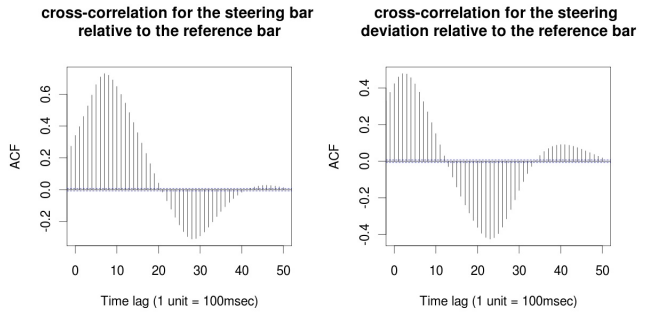
*Spline models* fit a smooth curve to a set of noisy observations using piece-wise polynomial functions with a high degree of smoothness at the points where the polynomial pieces connect. Spline plots are useful for getting a visual impression of the shape of a function or time series.

We also used *linear mixed effects models* (LME, [19]) to test whether our linguistic manipulations are predictive of driving performance and our cognitive measures. These models can be seen as a generalization of linear regression models that allow inclusion of random factors (such as participants or items) as well as fixed factors (e.g., reference bar velocity). When reporting mixed models, we give the estimates of the coefficients $\beta$ of the included fixed factors; these can be interpreted as the weights of the factors in the model (though only coefficients of factors on the same scale can be compared directly). In addition, each coefficient is associated with a standard error (SE), which expresses amount of variation in the estimate of that coefficient, and a $t$-value, which indicates whether the coefficient is significantly different from zero. For the model as a whole, we can measure the log-likelihood $ll$, which is an indicator of how well the model fits the data. Two models can be compared by testing whether their log-likelihood values are significantly different.

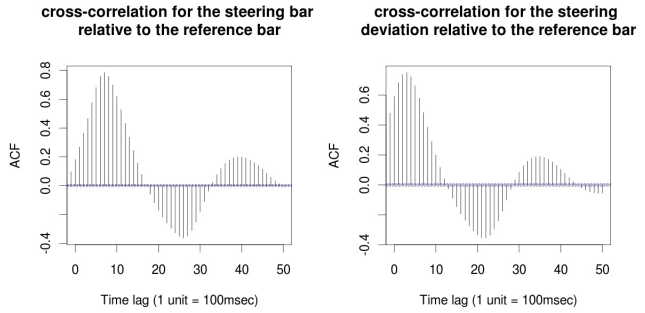
In the LME models reported in this paper, we treat the participant as a random factor, which means that our models contain an intercept term for each participant. In a model with steering deviation as a response variable, the random factor for participant allows the model to represent how well each the individual steers. Furthermore, we include a random slope for the predictor of interest (e.g., our linguistic manipulation), essentially accounting for idiosyncrasies of a participant with respect to the predictor of interest, such that only the part of the variance that is common to all participants and can be attributed to the main effect of the predictor. For models that test the effect of our linguistic manipulation, we furthermore include a random intercept and random slopes for items.

# 6. BEHAVIORAL EFFECT OF DRIVING TASK

Figure 2a shows the cross-correlation of the velocity of the reference bar and the steering bar at different time lags. We can see that the speed of the reference bar was most

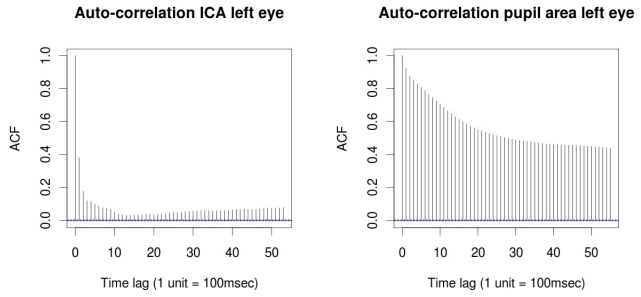



(a) easy driving (b) easy driving



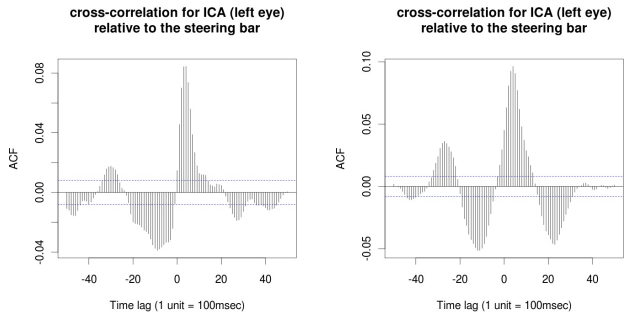

(c) difficult driving (d) difficult driving

Figure 2: Correlation analysis in time series for steering bar velocity (a,c) and steering deviation (b,d) with respect to reference bar velocity.




(a) ICA (left eye) (b) pupil area (left eye)

Figure 3: Auto-correlation function (ACF) for ICA and pupil area time series. (The 95% confidence interval is so close to 0 that it is hardly visible in the figures.)




(a) easy driving (b) difficult driving

Figure 4: Cross-correlation analysis for ICA and steering bar velocity. We find the same patterns for the ICA of the right eye.

strongly correlated with the speed of the steering bar at a time lag of about 800 msec. Hence, the average time it took the subjects to get the steering bar up to the speed of the reference bar (or accordingly, slow it down) was 800 msec.

The correlation is highly significant; the 95% confidence interval is indicated by the double dashed line close to 0—all values outside the area between the dashed lines are significantly different from 0 (no correlation) at $p < 0.05$. The overall periodicity in Figures 2 is due to the periodicity of the reference bar movement. Figure 2b shows the cross-correlation between the speed of the reference bar and the steering deviation. The speed of the reference bar was most strongly correlated with the steering deviation at around 400msec. This indicates that the subjects reacted to the movement of the target bar with a latency of about 400msec on average. Note that the high correlation at 400msec latency cannot be explained by the periodicity of the reference bar movement, which would yield a longer latency; additionally, we would then see different correlation latencies for the easy vs. difficult driving conditions. Figures 2c and 2d show analogous relationships for the difficult driving setting. We can see from the cross-correlation plots that the reaction times were identical between the two driving conditions.

Figure 3 displays the auto-correlation analyses for the ICA and for pupil area. The fast-declining self-correlation of the ICA (Figure 3a) demonstrates an important and advantageous property of the ICA over pupil area: it is highly dynamic in that the ICA value measured at a certain point in time is largely independent of the ICA measured a few hundred msecs earlier. As Figure 3b shows, pupil area has a much stronger self-correlation and is thus not a similarly dynamic measure. Figure 4 shows that there was a significant positive correlation at a time lag of about 500msec. Hence, when the driver moved the steering wheel, we saw a reaction in terms of the frequency of rapid dilations in the eye about 500msec later. Figure 4b shows that this relationship was more prominent in the difficult driving condition. On the other hand, when we run a cross-correlation analysis for *pupil area* and steering bar velocity, we find no significant positive correlation.

We conclude that the ConTRe steering task invokes a measurable cognitive load, which we can pick up with the ICA measure but not with traditional pupil dilation measures. The cross-correlation analysis furthermore indicates the delay with which to expect an effect on the ICA. The delay peaked at approximately 500msec with respect to the steering bar and with a lag of approximately 1sec with respect to the reference bar stimulus.

We also collected information about participants' age, video gaming experience and gender. As all belonged to the same age group (20–34 years), we did not find any significant effect of age on steering performance. Video gaming experience was not found to be a significant predictor of steering performance. Gender=male is a significant negative predictor of steering deviation ($\beta = -0.055; sdev = 0.0217; t = -2.55; p < 0.5$). We furthermore analyzed the reaction la-

| | $\beta$ | sdev | t val | signif |
|---|---|---|---|---|
| (Intercept) | 0.660 | 0.0196 | 33.64 | *** |
| easy driving | -0.305 | 0.0150 | -20.33 | *** |
| language | 0.056 | 0.0112 | 5.00 | *** |
| easy:lang | -0.024 | 0.0035 | -6.98 | *** |

Table 2: The language task leads to decreased steering performance.

tencies for steering on an individual bases using a cross-correlation analysis for each individual and determining the delay at which the largest correlation was measured. Reaction time varied with a correlation maximum between 300 and 500msec delay ($mean = 379msec, sdev = 58msec$). Gender was a significant predictor of reaction times, with males having shorter time delays for maximal correlation between reference bar and steering deviation, while video gaming experience and age were not significant predictors. Maximum correlation between reference bar and steering bar ($avgcor = 0.8, sdev = 0.055$) was reached with a delay of 700msec to 1 sec ($mean = 8125msec, sdev = 850msec$).

The individual differences were larger with respect driving task effect on the ICA measure. For one third of the individuals, we do not find a significant correlation between the ICA and the steering task at any time lag. These individual differences were not explained by age, gender or video gaming experience. We find, however, that there was a correlation between steering performance and the size of the ICA effect; the correlation of steering bar movement and the ICA was larger for those individuals who showed largest correlations for the steering bar velocity vs. target bar velocity and had the smallest steering deviations. This means that our measure of cognitive load works best for those people that performed best at the steering task.

# 7. SECONDARY TASK (LANGUAGE COMPREHENSION)

Table 2 shows main effects of driving difficulty and the linguistic task on steering performance: Using linear mixed effects models with a random intercept and random slopes by subject, we found a large significant main effect of driving difficulty ($\beta = 0.3; t = 20.33; p < 0.001$), showing that steering was less accurate when driving was more difficult. This reveals that the difficulty manipulation setting in the steering task was effective. We also found a significant positive main effect of whether we are in a language phase ($\beta = -0.05; t = -5.00; p < 0.001$; steering is worse when people are listening to language, see also Table 2), as well as a significant interaction between driving difficulty and the language phase, indicating that the effect of language was more burdensome in the difficult driving condition ($\beta = -0.024; t = -6.98; p < 0.001$). To confirm whether the effect of language is significant in both driving conditions, we also split the data into two subsets, easy driving and difficult driving, and found that the effect of language was statistically significant in both linear mixed effects models.

Each experimental phase consisted of 2 minutes of single-task driving, followed by 4 minutes of driving with a simultaneous linguistic task. Spline plots in Figure 5 show that all of our behavioral measures pick up on the dual task condition. The plots in Figure 5(a,c,d) show smoothed splines and their 95% confidence intervals aggregated by phase. Plot 5(a)

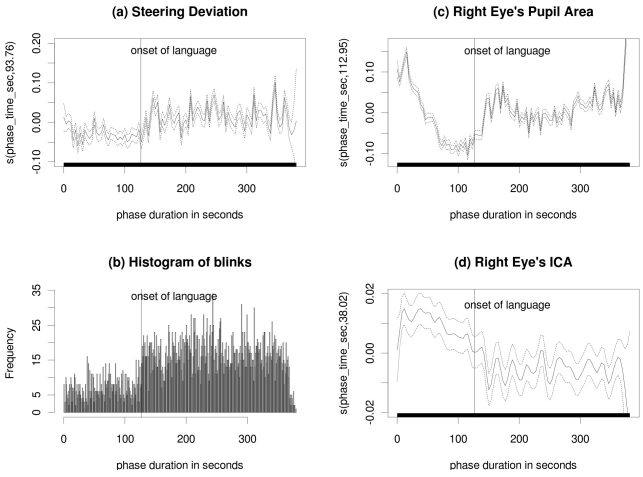



Figure 5: Presence of the language task is reflected in behavioral measures as well as steering performance.

shows that the level of steering deviation was much higher during the dual task phase than during the single task phase in accordance with the data in Table 2.

For pupil area (Figure 5(c)), we see that the pupil was initially large, but it contracted as the participant got used to the task. When the dual task condition started, the pupil dilated significantly and remained at a higher dilation level than in the single task condition. We furthermore recorded a higher blink rate during the dual task condition, shown in the histogram in Figure 5(b), consistent with the literature [18]. Finally, Figure 5(d) shows that the ICA levels were consistently *lower* during the dual task setting than in the single task condition. It is possible that the lower ICA level is due to "downsampling" both tasks (we also see that performance in driving is much worse than in a single task setting). It also shows that the ICA as a measure is not equivalent to overall pupil size. Interestingly, we also observe 10 distinct peaks in the ICA during the dual task period of the experiment, which we find correspond exactly to our 10 linguistic stimuli. We will investigate the relationship between the ICA and our linguistic manipulations in more detail below.

We calculated a mixed effects regression model for the critical region (200-1200msec after onset of *hat / haben*) with steering deviation as a response variable. The time period 200-1200msec was chosen to start at the point where the sound of *hat* starts differing from *haben*, which includes the next two words of the linguistic input, during which people may still be processing the information from the ORC. Equivalent results are obtained for similar time windows within the first 2 seconds following *hat/haben*. We found significant main effects of the binary predictor *easy driving* (which is set to 1 in the easy driving condition and 0 otherwise) relative to the difficult driving setting, the speed of the reference bar `target velocity`, and `phase time` (i.e., how far along we are within an experimental phase). The predictor variable `target velocity` was shifted by 400msec with respect to steering deviation, so that we correlate with the speed of the target bar 400msec earlier, since we know from cross-correlation analysis that the steering deviation most strongly reflects movements in the target bar that happened

Table 3: Mixed effects regression analysis with steering deviation as response variable, for region of 200ms till 1200ms after onset of the critical region.

| | $\beta$ | sdev | t-val | signif |
|---|---|---|---|---|
| (Intercept) | 3.680e-01 | 3.868e-02 | 9.51 | *** |
| time since onset | 1.402e-05 | 2.183e-05 | 0.64 | |
| subject RC | 4.779e-02 | 3.229e-02 | 1.48 | |
| phase time | 1.247e-07 | 5.221e-08 | 2.39 | * |
| easy driving | -2.478e-01 | 7.150e-03 | -34.67 | *** |
| target velocity | 3.586e-01 | 3.879e-03 | 92.44 | *** |
| timeOnset:SRC | -5.915e-05 | 2.219e-05 | -2.67 | ** |

400msec earlier (Figures 2b and 2d). The model includes random effects for participant and item as well as random slopes of relative clause type and time since onset of the disambiguating region as random slopes under both participant and item. These explanatory variables all exhibit the expected effects: steering deviation was significantly smaller in the easy driving setting than in the difficult driving setting ($\beta = -0.247; sd = 0.00715; t = -34.67; p < 0.0001$), the speed of the reference bar is a significant positive predictor of steering deviation ($\beta = -0.3586; sd = 0.00389; t = 92.44; p < 0.0001$), and there is a small effect of phase time ($\beta = 1.247e^{-7}; sd = 5.221e^{-8}; t = 2.39; p < 0.05$), presumably reflecting some effect of fatigue. Interestingly, we also find a significant interaction ($\beta = -5.915e^{-5}; sd = 2.219e^{-5}; t = -2.67; p < 0.01$) of relative clause type and the time gone by since the onset of *hat / haben*. This interaction means that steering deviation was getting smaller following the less difficult word *hat* compared to the more difficult *haben*. In summary, we find evidence for an effect of our linguistic manipulation on steering: steering performance got worse during the time period following *haben*, the word that designates the relative clause to be an object relative clause.

The finding that steering deviation was high during the disambiguating region and decreased in the easy condition, may also be a hint for a learning effect during the experiment, i.e., participants paying increasing attention to the disambiguating region (despite the use of fillers (67%) in the experiment). Indications for learning and increased attention on the disambiguating region also came from people's self-report, answer accuracy on the comprehension questions (correctness did not differ between questions about subject and object relative clauses), and ICA effects on the disambiguating region (discussed below).

If people did indeed pay extra attention to the disambiguating region, we should also see an effect of larger steering deviation during the disambiguating region compared to the regions before and after that critical region. To test this, we compare steering accuracy at the time of the disambiguating region (0msec to 650msec after the onset of *hat / haben*) with steering accuracy during the two seconds before onset and after offset of *hat / haben*. We use linear mixed effects regression modeling with a random intercept for participant and a random slope for our predictor `critical region` under participant. `critical region` is a binary predictor which is 1 for all measurements during the *hat / haben* region and 0 otherwise.

Further predictors in the model include the continuous predictor `phase time`, i.e., indicating how far along the subjects were in the experiment, and the continuous predictor `reference bar velocity`. We furthermore include as a continuous predictor `target bar acceleration`, a measure derived from target bar speed, which we also shift by 400msec[3], and binary predictor `easy driving`. We find that steering deviation was significantly larger during the disambiguating region ($\beta = 0.0139; t = 2.88; p < 0.01$) than before or after; see also Table 4. This supports our hypothesis that focussing attention on the linguistic task during the disambiguating region led to decreased steering performance. The other predictors also behave as expected: steering deviation increased during the course of the experimental phase, the velocity of the target bar is a highly significant positive predictor of steering deviation, and there was significantly less steering deviation during the easy driving condition compared to difficult driving. This analysis thus shows that steering performance was worse during the disambiguating region of the relative clause, when people presumably paid additional attention to the linguistic task.

Table 4: Mixed effects regression analysis with steering deviation as response variable, for region of 2s before the onset till 2s after end of the critical region.

| | $\beta$ | t-value | signif |
|---|---|---|---|
| (Intercept) | 3.562e-01 | 17.07 | *** |
| phase time | 8.459e-08 | 3.44 | *** |
| target velocity | 3.832e-01 | 205.08 | *** |
| critical region | 1.396e-02 | 2.88 | ** |
| easy driving | -2.248e-01 | -64.91 | *** |
| target acceleration | -2.680e-02 | -5.90 | *** |

Finally, we test whether the ICA is sensitive to fine-grained linguistic complexity effects. We isolate the subset of the data which fell within the 1800msec following the onset of the critical region *hat / haben*. We chose an interval of 1800 msec to capture the duration of the disambiguating region (650msec) and expected delay of effect in the ICA measure, peaking at about 1 second after stimulus, cf. Fig. 4. The duration of this critical region at *hat / haben* was 650 ms in both conditions, which we imposed by manipulating the duration of the phrase boundary pause during synthesis. On this subset of the data, we build two LME models (one for each eye) with the ICA measure as the response variable, random intercept for participants, and the relative clause type (`subject RC`) as a fixed effect. Additional predictors in the model are `phase time`, steering velocity shifted by 400msec (based on cross-correlation results in Figure 4b) `steering velocity`, and `steering acceleration` (also with a 400msec lag). Furthermore, we include the variable `time wrt. onset`, which is a continuous variable encoding the time gone by since the onset of the critical region *hat / haben*. Our models also include the random slope of relative clause type under participant.

The results of this analysis are shown in Table 5. We can see that there was a negative effect for the SRC type both for the left ICA and the right ICA, although only the result for the left eye is significant ($\beta = -0.0354; t = -2.12; p < 0.05$); the larger effect on the left eye is consistent with single task language processing tasks findings [4] and may reflect hemispheral differences in the brain regions related to activation of the muscles that control pupil size. The interpretation

[3]The predictor target bar acceleration was also initially included in the model shown in Table 3, but did not significantly improve model fit.

Table 5: Mixed effects regression analysis with left and right ICA as response variable, 200–1800msec after critical region onset. (Critical region duration: 0-650msec)

| | left ICA | | right ICA | |
|---|---|---|---|---|
| | $\beta$ | t-value | $\beta$ | t-value |
| (Intercept) | 0.7504 | 35.71 *** | 0.736 | 37.82 *** |
| subject RC | -0.0354 | -2.12 * | | |
| phase time | $-1.16\times10^{-7}$ | -2.59 * | | |
| time wrt. onset | $-2.78\times10^{-5}$ | -6.38 *** | $-1.84\times10^{-5}$ | -4.36 *** |
| steering veloc | 0.0257 | 5.37 *** | 0.0226 | 4.88 *** |
| steering accel | 0.0108 | 2.00 * | | |
| SRC:phase time | $1.34\times10^{-7}$ | 2.12 * | | |

of the negative coefficient is that SRCs tend to occur with smaller values of ICA than ORCs, as expected based on single task results [4]. This result provides evidence that the ICA is picking up on our manipulation even in the dual task setting and is in line with our findings of effects on steering deviation. We also see a learning effect, however, evident in the significant interaction of relative clause type and phase time `SRC:phase time`. As the experimental phase proceeded, the ICA difference between subject and object relative clause conditions got smaller.

# 8. DISCUSSION AND RELATED WORK

We have presented a number of results from an exploratory analysis of an experiment in simulated driving and language. As the paradigms used are novel and little explored, future work will have to test replication of the effects identified here. The analyses we used allowed us to measure the characteristics of the ConTRe steering task (section 6). We found that expected user behavior in terms of reaction times is borne out by the task, including comparable reaction times for the easy and difficult driving settings. This strengthens the case for the steering task as a way of measuring cognitive load.

We then proceeded to evaluate pupillometric measures; the observation that the self-correlation of ICA over time is lower than that of pupil area allows us to suggest that ICA is the more dynamic and hence more suitable measure for continuous tasks. This leaves us the question of whether or not the ICA actually reflected the workload of the task, and subject to the reaction times of the experimental participants, we found that it did.

With this background, we were then able to show not only the predictiveness of language input in steering deviation and pupillometric measures under both our conditions but also the effect of moment-to-moment grammatical processing; we could identify a disambiguating region in locally ambiguous German relative clauses with effects both on driving performance and pupil behavior. This leaves us with an important question: what is it about language that interacts with overall mental workload to produce these effects?

To our knowledge, this question has not been closely examined until very recently. There is indeed a rich literature on language use while driving a car, but it focuses on a much more coarse-grained level. For instance, there is work that shows that speaking on the telephone has a negative effect on driving performance [12, 13].

We would argue that the construction of automotive user interfaces needs to take into account the moment-to-moment burden of linguistic cognition in the design of new systems, particularly hands-free systems based on spoken dialogue. We know from earlier studies that conversations with remote conversational partners during driving (be it via mobile phone or via hands-free device) has a negative effect on driving performance, while conversations with an in-car passenger are less problematic [21, 6]. It appears that passengers adapt their conversation to the traffic situation, leaving the driver more resources to deal with demands of the driving task when driving becomes difficult [6, 3, 22]. By contrast, remote conversational partners cannot adapt their speech, so that the driver may reach the point of high cognitive load more easily and thus commit driving errors. However, these lines of research have not taken into account how the fine-grained details of linguistic complexity affect cognitive load and driving task performance.

Where else can we turn for guidance on linguistic complexity and workload? There are multiple models that explain dual-task cognitive load [1, 24, 11].

Specific to language, there is a very rich literature on linguistic processing difficulty in single tasks using brain imaging, ERPs, and reading time studies, as well as a number of dual task experiments generally showing that performance on the linguistic task deteriorates with increased complexity in the other task [10]. This study expands on this literature in testing different methods for assessing cognitive load and the effect of one particular linguistic structure—incrementally ambiguous relative clauses—on driving performance in a simplified but controllable and continuous driving task.

Further insight comes from a study by [12], who conducted a dual task fMRI study where people were driving a simulator while performing a language task (judging sentences to be true or false). Driving is a task that appears to draw on separate areas of the cerebral cortex from the areas occupied by language. They found that both the driving-associated cortical areas and the language-associated cortical areas are activated simultaneously, confirming that this type of dual-task arrangement can be used to experiment with cognitive load in simultaneous attention environments. The results also showed however that the overall neural activation levels of two simultaneous tasks were less than the sum of each tasks' activation individually. Furthermore, [12] observed a degradation in the performance of the driving task given the dual task presentation.

So what is it about grammatical constructions like subject and object relative clause disambiguation that might lead to cognitive load? Various psycholinguistic models have been developed in recent years, which can explain many of the attested effects of linguistically induced processing difficulty (e.g., surprisal [8, 14], DLT [7]). If we can relate these measures in a manner time-locked to driving performance or to in-automotive eye-tracker data, it will potentially point the way towards spoken dialogue systems that dynamically respond to the user's mental burden by adopting more or less compressed information delivery varying with driving needs.

# 9. CONCLUSIONS

In this paper, we showed that fine-grained linguistic complexity has measurable effects on cognitive load. We designed the tasks in our experiment to require continuous attention. The language task clearly affected performance on the primary steering task: we saw the effect of the secondary task in all of our measures. Furthermore, we found effects of

linguistic ambiguity and complexity in our measures of cognitive load: during the disambiguating region, we observed significantly higher steering deviation, which indicated that people are allocating more mental resources to the linguistic task, hence impeding steering performance. We also found evidence for a measurable effect of linguistic complexity in our pupillometric measure ICA: the ICA was significantly higher during the disambiguating region and the following second for the ORC condition compared to SRC.

This experiment has provided early support for both the ConTRe steering task and the ICA as useful measures for the assessment of language-induced cognitive load. We have shown several aspects of the suitability of the ConTRe steering task for measuring the fine-grained effects of cognitive load in a smooth, continuous driving environment. A key characteristic of the task is that is uses a continuous measure instead of turns or abrupt stimuli, making it suitable for use with tasks that require careful synchronization, such as measuring the workload that accrues to particular grammatical turns. It extends the assortment of solutions available for measuring driver distraction in simulator environments and was created to compensate certain drawbacks of other driving tasks. In summary, we successfully demonstrated that a more sensitive task is able to reveal more subtle effects on driving performance. A more fine-grained evaluation of driving performance enhances the possibilities to investigate cognitive workload and the effect of secondary tasks on cognitive workload.

## 10. REFERENCES


[1] A. Baddeley. Working memory: looking back and looking forward. *Nature Reviews: Neuroscience*, 4, 2003.

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

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

[4] V. Demberg, E. Kiagia, and A. Sayeed. Language and cognitive load in a dual task environment. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (CogSci-13)*, 2013.

[5] V. Demberg and A. Sayeed. Linguistic cognitive load: implications for automotive uis. In *Cognitive load and in-vehicle human-machine interaction, workshop at AutomotiveUI 2011*, 2011.

[6] F. A. Drews, M. Pasupathi, and D. L. Strayer. Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, 14(4):392, 2008.

[7] E. Gibson. Dependency locality theory: A distance-dased theory of linguistic complexity. In A. Marantz, Y. Miyashita, and W. O'Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT Press, Cambridge, MA, 2000.

[8] J. Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 159–166, Pittsburgh, PA, 2001.

[9] J. Hyönä, J. Tommola, and A. Alaja. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, 48(3):598–612, 1995.

[10] M. A. Just and P. A. Carpenter. The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian journal of experimental psychology*, 47(2):310–339, 1993.

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

[12] M. A. Just, T. A. Keller, and J. Cynkar. A decrease in brain activation associated with driving when listening to someone speak. *Brain research*, 1205:70—80, 2008.

[13] T. T. Kubose, K. Bock, G. S. Dell, S. M. Garney, A. F. Kramer, and J. Mayhugh. The effects of speech production and speech comprehension on simulated driving performance. *Applied cognitive psychology*, 20(1):43–63, 2006.

[14] R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.

[15] A. Mahr, M. Feld, M. Moniri, and R. Math. The ConTRe (Continuous Tracking and Reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. In *Automotive User Interfaces and Interactive Vehicular Applications*, pages 88–91, 2012.

[16] S. Marshall. U.s. patent no. 6,090,051, 2000.

[17] S. Marshall. The index of cognitive activity: Measuring cognitive workload. In *Human Factors and Power Plants, 2002. Proceedings of the 7th Conference on*, pages 7–5. IEEE, 2002.

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

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

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

[21] D. Strayer, F. Drews, and W. Johnston. Cell phone-induced failures of visual attention during simulated driving. *Journal of experimental psychology: Applied*, 9(1):23, 2003.

[22] J. Villing. In-vehicle dialogue management - towards distinguishing between different types of workload. In *Proceedings of SimPE, fourth workshop on speech in mobile and pervasive environments*, 2009.

[23] P. Viviane and P. Mounoud. Perceptuomotor compatibility in pursuit tracking of two-dimensional movements. *Journal of Motor Behavior*, 22(3):407–443, 1990.

[24] C. D. Wickens. Multiple resources and mental workload. *Human factors*, 50(3):449–455, 2008.