# Paper Proposal : Steps towards syntactic and semantic safety guarantees for voice assistants in autonomous vehicles

Warrick Macmillan
Marco, Matthew, Katya, ...

# 1 Abstract

We introduce a grammar for a controlled natural language (CNL) to give imperative commands to a voice assistant for a self-driving car to verify its behavior. We specifically seek to give the user of our system assurances against substitution-based attacks, whereby synonyms can be given the same meaning by imposing these conditions in how the parse trees (in our case, abstract syntax trees (ASTs)) are formed. In addition, we map our trees to a semantic form in an Agda implementation of linear temporal logic (LTL), which has many applications in the specification and verification of the behavior of robotics systems, particularly those with neural network components. We see that simple, formal systems provide a useful model for verifying systems with more a greater breadth of "knowledge" capable of more complex interactions.

# 2 Introduction

While the evaluation of machine learning systems provides assurances using different scores and metrics on different tasks assures one they may on average perform better than humans at certain tasks, the advent of adversarial attacks [5] with the intention of deceiving such a system by a hostile actor leads the system designer to desire, and possible require additional verification about the system's behavior. In the context of natural language processing (NLP), where data sources rely on strings of text, these attacks can focus an array of features from spellings of individual words to rearranging entire sentences []. So-called synonym attacks, which adversarially target the system at the lexical level, can cause traditional NLP models to [...] [].

In the context of designing a voice assistant for an autonomous vehicle, whereby one can give commands like "turn right after the woman with the big dog", we desire that the intensional belief a user has about her utterance is consistent with the extensional behavior of the vehicle. This can be done through an intermediary mapping to a formal semantic representation. Ensuring that the syntactic content of a voice director's (well-formed) utterance maps predictably to the logical form is important from the verificationist perspective : one wants to maximize the "syntactic completeness" of the system [3].

Aside from the user experience being compromised by a system which has been adversarially afflicted, there is also a possibility of physical danger for the passenger and other people in the vicinity. As voice directed robots have many possible points of failure, we focus on two types of verification for our system. Rather than focus on breadth of language coverage, which ML language models excel at of due to their reliance on statistical modeling and tons of data, our system is narrowly focused as a proof-of-concept, from which it could either be extended by hand, or different components modified using other techniques and tools.

# 3 Current Landscape

## 3.1 Voice assistants for autonomous vehicles

The public company Cerence [] is already designing voice assistants for autonomous vehicles, for which it has a large software stack between the voice processing to actual control of current automative components. In addition to its technologies, many of which aren't accessible to external researchers due to intellectual property restrictions, Cerence has contracts with large automakers [..]. It is therefore natural to inquire, what a small team with varied backgrounds and not nearly the same expertise nor experience within the technological team at Cerence can provide.

First, we believe that the focus on verification, insofar as we envision it, is unlikely to be of current concern at Cerence due to the fact that their products are still being developed, and the primary goal of producing a working product is likely to precedence over preventing non-existent hostile actors.

Additionally, it is going to have to be determined by [verification of self-driving cars generally : software, hardware, behavior in a real environment, etc]

### 3.2 Natural Language and Robots, generally

### 3.3 Semantic Representations of NL

Model logics, specifically those dealing with time like LTL, CTL, STL, ...

> Overall, the typical approach followed by these studies can be summarized as follows: given an input English utterance, preprocess it to extract syntactical information, which may include part of speech tagging, dependency parsing, semantic role labelling, and so on. Then, enrich the input with these pieces of information. Finally, run an attribute grammar-based parser, or rely on some hand-made rules, to derive a translation into a target logical format. A notable exception is the work of [89], where a fully-supervised learning setting is considered. [1]

### 3.4 Foundation Models

## 4 Work

### 4.1 GF Grammar

## 5 TODO

### 5.1 Grammar modulo wordnet

### 5.2 LTL in Agda

Along with colleagues from Singapore Management University, we have begun an Agda implementation [4] of LTL which will serve as the semantic space for our parsed utterances. Our method, uses a deep embedding, as opposed to the shallow embedding in [2], although the temporal encoding of paths as streams was directly adapted from this paper.

This implementation will hopefully allow us to prove decidability of LTL in a relatively straightforward manner. Other than the assurance that our implementation is correct, we hope this will allow us to feed the formula into some SAT or SMT solver so-as to actually allow verification of the behavior of a vehicle with respect to an utterance.

[TODO : Help from Matthew?]

### 5.3 AST -> Agda

### 5.4 ML training/verification stuff

Help from Marco, Nathalia if interested?

## References


[1] Andrea Brunello, Angelo Montanari, and Mark Reynolds. "Synthesis of LTL Formulas from Natural Language Texts: State of the Art and Research Directions". In: *26th International Symposium on Temporal Representation and Reasoning (TIME 2019)*. Ed. by Johann Gamper, Sophie Pinchinat, and Guido Sciavicco. Vol. 147. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, 17:1–17:19.

[2] Solange Coupet-Grimal. "An Axiomatization of Linear Temporal Logic in the Calculus of Inductive Constructions". In: *Journal of Logic and Computation* 13.6 (2003), pp. 801–813.

[3] Warrick Macmillan. "On the Grammar of Proof". MA thesis. University of Gothenburg, 2021, p. 90.

[4] Warrick Macmillan and Andreas Kallberg. *LTL in Agda*. https://github.com/wmacmil/LTL-Agda. 2021.

[5] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2014.