# Human Language for Artificial Intelligence

David R. Mortensen and Lori Levin

Language Technologies Institute
Carnegie Mellon University

August 26, 2019

# The Course

Human Langugae for Artificial Intelligence (hereafter HL4AI) is an introduction to the language sciences geared toward graduate students in computer science and artificial intelligence fields. It surveys linguistics and the other language sciences broadly and, while it cannot cover everything, it attempts to present a comprehensive picture of the aspects of language that impact the fields of human language technologies and artificial intelligence.

# The instructors

Both instructors of this course are COMPUTATIONAL LINGUISTS. That means we work at the intersection of computer science and linguistics.

- **David R. Mortensen**
    - Systems Scientist
    - dmortens@cs.cmu.edu
    - Phonology, morphology, typology, historical linguistics
- **Lori Levin**
    - Research Professor
    - lsl@cs.cmu.edu
    - Syntax, semantics, typology, data resources

# Reasons to take this course

1. You are a computer scientist and you're interested in incorporating linguistic knowledge into AI research and practice
2. You are a language technologist and you believe that having more knowledge about the structure and function of language will make you better at speech and language processing
3. You are a linguist interested in how language intersects with computation
4. You are an intelligent layperson who is comfortable with computers and programming and wants to apply these to perspectives to language

# The *in vivo* Turing Test: a *Gedankenexperiment*

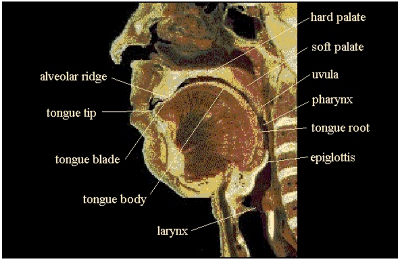
- Many people are familar with the Turing Test:
  - Can an agent convince a human that it is human by its interactions over a text terminal...
  - Through its human-like use of emoticons
- We think this test is too easy.
- Imagine instead a test in which a robot must convince a human interlocutor that it is human **in person**.
- Imagine that it must do so by speaking Tamil
- This would involve many non-linguistic and paralinguistic elements
  - Physical appearance
  - Motor functions
  - Gesture
  - Facial expression
  - World knowledge
- **However, some of the most challenging aspects of building such a robot involve speech and language**

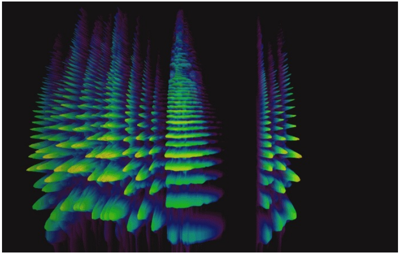
# The question

What must such a robot know about language in order to pass the test?

# Pronunciation

The robot must be able to produce speech that sounds like that produced by human speakers of Tamil. This includes modeling the subtle, physically-induced variations that occur in speech sounds when they overlap with other sounds. **Map abstract representation to acoustic signal.**





The robot must also be able to take an acoustic signal as input and map this on to a more abstract representation that generalized over the variants present in the signal.

# Pronunciation I

This is **phonetics**

# Pronunciation

There are more dramatic varitions in how speech sounds are realized when meaningful units are put together to form words and words are put together to form sentences. These variations are part of the rules of the language, rather than following from physical laws. They are especially important in a language like Tamil, where words have a lot of internal structure.

©Study.com

# Pronunciation II

This is **phonology**

# Words

A word in Tamil may have many forms, depending on context. Take the verb *kalakku*.

**Present**

| | |
|---|---|
| நான் | கலக்குகிறேன் |
| நீ | கலக்குகிறாய் |
| அவன் | கலக்குகிறான் |
| நாம் | கலக்குகிறோம் |
| நீர் | கலக்குகிறீர் |
| அவர் | கலக்குகிறார் |

**Past**

| | |
|---|---|
| நான் | கலக்கினேன் |
| நீ | கலக்கினாய் |
| அவன் | கலக்கினான் |
| நாம் | கலக்கினோம் |
| நீர் | கலக்கினீர் |
| அவர் | கலக்கினார் |

**Future**

| | |
|---|---|
| நான் | கலக்குவேன் |
| நீ | கலக்குவாய் |
| அவன் | கலக்குவான் |
| நாம் | கலக்குவோம் |
| நீர் | கலக்குவீர் |
| அவர் | கலக்குவார் |

**Imperative**

| | |
|---|---|
| நான் | கலக்கேன் |
| நீ | கலக்காய் |
| அவன் | கலக்கான் |
| நாம் | கலக்கோம் |
| நீர் | கலக்கீர் |
| அவர் | கலக்கார் |

The robot must know how to deal with new words.

# Words

This is **morphology**

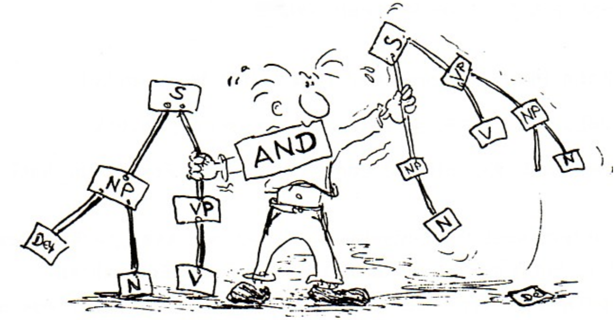
# Form and meaning

Language without meaning is meaningless. Language users employ various strategies to express meanings as parts of words, words, phrases, and sentences. But meaning is much “bigger” than language—a lossy coding scheme. Our robot must know how to “push” the vastness of meaning through the funnel of grammar.

# Form and meaning

This is **morphosyntax**

# Phrase and sentence structure



The robot must know how combine words into phrases and phrases into coherent sentences. Part of this is semantic, and part of it is grammatical. It also much have a way of representing this structure, be it symbolic or continuous.

# Phrase and sentence structure

This is **syntax**

# Meaning

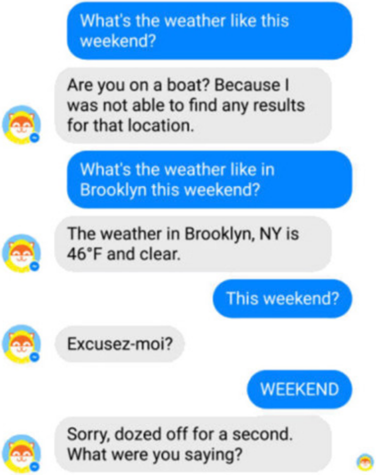
Behind everything in language is meaning, and our robot must have a way of representing it—a way of relating world knowledge and communicative intent to linguistic form.

# Meaning

This is **semantics**

# Language in context

If our robot is to convince people it is human, it needs to know more than how to make grammatical, meaningful sentences. It needs to be able to chain them together into coherent discourses (especially **dialogs**). It must also understand that utterances like “Can you pass the salt?” means “Please pass the salt,” not “Are you able to pass the salt?”



What's the weather like this weekend?

Are you on a boat? Because I was not able to find any results for that location.

What's the weather like in Brooklyn this weekend?

The weather in Brooklyn, NY is 46°F and clear.

This weekend?

Excusez-moi?

WEEKEND

Sorry, dozed off for a second. What were you saying?

# Language in context

This is **pragmatics** and **discourse**

# Language in social context

Finally, to be convincing, our robot needs to know that language has **social meaning** as well as **propositional meaning.** It has to be able to use language in ways that construct a social identity for itself and that convincingly interact with the social identities of its interlocutors. For Tamil, this means using HONORIFIC forms of pronouns appropriately, depending on whom it is talking to.

# Language in social context

This is **sociolinguistics**

# What this course can and cannot do

- This course cannot make you an expert in any of these fields
- However, this course will introduce you to **all** of these fields
- **Goals**: at the end of this course you will:
  - Know what you don't know about language
  - Know consciously (some of) what you didn't know you knew about language
  - Be empowered to seek out information on language that is relevant to your work, either through other courses or through self-study
  - Know the true meaning of fun

Questions?