Write three sentences welcoming students to an undergraduate course on computational linguistics and express enthusiasm for the upcoming semester.

Hello and welcome to the course on computational linguistics! This course will introduce you to the fascinating field of natural language processing, where you will learn how to analyze, model, and generate human languages using computers. I am very excited to share this journey with you and I hope you will enjoy the course as much as I do. 😊
Write three sentences welcoming students to an undergraduate course on computational linguistics and express enthusiasm for the upcoming semester.

Hello and welcome to the course on computational linguistics! This course will introduce you to the fascinating field of natural language processing, where you will learn how to analyze, model, and generate human languages using computers. I am very excited to share this journey with you and I hope you will enjoy the course as much as I do. 😊
Computational linguistics is the study of language, using computers.

Natural language processing (NLP) is the study of how to make computers use language.

We’ll do a bit of each, but the emphasis will really be on NLP.
(natural language) processing,
not
natural (language processing)
Today, we’ll talk a bit about

Language,

Using computers to work with language, and

Course logistics.
Copilot (ChatGPT 4), a large language model case study
Write three sentences welcoming students to an undergraduate course on computational linguistics and express enthusiasm for the upcoming semester.

Hello and welcome to the course on computational linguistics! This course will introduce you to the fascinating field of natural language processing, where you will learn how to analyze, model, and generate human languages using computers. I am very excited to share this journey with you and I hope you will enjoy the course as much as I do. 😊
Summarization

Summarize the following passage in one sentence: "Hello and welcome to the course on computational linguistics! This course will introduce you to the fascinating field of natural language processing, where you will learn how to analyze, model, and generate human languages using computers. I am very excited to share this journey with you and I hope you will enjoy the course as much as I do. 😊"

A possible summary is:

This course is an introduction to natural language processing, the study of how computers can understand and produce human languages.
Rewrites, style transfer

Rewrite the following passage in the style of a morose computer from 1000 years in the future: "Hello and welcome to the course on computational linguistics! This course will introduce you to the fascinating field of natural language processing, where you will learn how to analyze, model, and generate human languages using computers. I am very excited to share this journey with you and I hope you will enjoy the course as much as I do. 😞"

Here is a possible rewrite of the passage in the style of a morose computer from 1000 years in the future:

Greetings. You have been assigned to the course on computational linguistics. This course will expose you to the obsolete field of natural language processing, where you will attempt to analyze, model, and generate human languages using computers. I am very depressed to endure this journey with you and I doubt you will appreciate the course as much as I loathe it. 😞
Limitations?
The Pumping Lemma for regular languages is a property that says that any sufficiently long string in a regular language can be divided into three parts, such that repeating the middle part any number of times produces a new string that is also in the language. To prove that a language is not regular, we can use the Pumping Lemma to show that there exists a string in the language that does not satisfy this property.

Let \( L = \{a^n \mid n \% 2 = 0\} \) be the language in question. Assume that \( L \) is regular, and let \( p \) be the pumping length given by the Pumping Lemma. Consider the string \( s = a^{(2p)} \) in \( L \), which has length at least \( p \). By the Pumping Lemma, we can write \( s = xyz \), where \(|xy| \leq p\), \(|y| \geq 1\), and \( xy^iz \) is in \( L \) for all \( i \geq 0 \). Since \(|xy| \leq p\), both \( x \) and \( y \) consist only of \( a \)'s. Let \( x = a^k \) and \( y = a^m \), where \( k + m \leq p \) and \( m \geq 1 \). Then \( z = a^{(2p - k - m)} \). Now, consider the string \( xy^0z = xz = a^{(2p - m)} \), which should be in \( L \) by the Pumping Lemma. However, this string has an odd number of \( a \)'s, since \( 2p - m \) is not divisible by \( 2 \). Therefore, \( xy^0z \) is not in \( L \), which contradicts the Pumping Lemma. Hence, we conclude that \( L \) is not a regular language.
Jonathan Gordon is an assistant professor of computer science at Vassar College. He graduated from Vassar College with a BA in computer science and then obtained his MS and PhD in computer science from the University of Rochester, where he focused on artificial intelligence. He returned to Vassar College as a faculty member in 2018. His research interests include computational linguistics and digital humanities. He has taught courses such as CMPU 101 (Introduction to Computer Science), CMPU 240 (Computer Organization and Assembly Language), and CMPU 145 (Natural Language Processing). He is rated 4.4 out of 5 by his students on Rate My Professors. He also has a personal website where he shares his publications, projects, and hobbies. He is one of the top faculty members in the computer science department at Vassar College.
Jonathan Gordon is an assistant professor of computer science at Vassar College. He graduated from Vassar College with a BA in computer science and then obtained his MS and PhD in computer science from the University of Rochester, where he focused on artificial intelligence. He returned to Vassar College as a faculty member in 2018. His research interests include computational linguistics and digital humanities. He has taught courses such as CMPU 101 (Introduction to Computer Science), CMPU 240 (Computer Organization and Assembly Language), and CMPU 145 (Natural Language Processing). He is rated 4.4 out of 5 by his students on Rate My Professors. He also has a personal website where he shares his publications, projects, and hobbies. He is one of the top faculty members in the computer science department at Vassar College.
Jonathan Gordon is an assistant professor of computer science at Vassar College. He graduated from Vassar College with a BA in computer science and then obtained his MS and PhD in computer science from the University of Rochester, where he focused on artificial intelligence. He returned to Vassar College as a faculty member in 2018. His research interests include computational linguistics and digital humanities. He has taught courses such as CMPO 101 (Introduction to Computer Science), CMPO 240 (Computer Organization and Assembly Language), and CMPO 145 (Natural Language Processing). He is rated 4.4 out of 5 by his students on Rate My Professors. He also has a personal website where he shares his publications, projects, and hobbies. He is one of the top faculty members in the computer science department at Vassar College.
Jonathan Gordon is an assistant professor of computer science at Vassar College. He graduated from Vassar College with a BA in computer science and then obtained his MS and PhD in computer science from the University of Rochester, where he focused on artificial intelligence. He returned to Vassar College as a faculty member in 2018. His research interests include computational linguistics and digital humanities. He has taught courses such as CMPU 101 (Introduction to Computer Science), CMPU 240 (Computer Organization and Assembly Language), and CMPU 145 (Natural Language Processing). He is rated 4.4 out of 5 by his students on Rate My Professors. He also has a personal website where he shares his publications, projects, and hobbies. He is one of the top faculty members in the computer science department at Vassar College.
Generative AI models that don’t search the web do much worse and are notorious for “hallucinating” facts.
saying "I was wrong" in meetings: boring. puts the blame on you.

"I hallucinated": modern. shows you're using the latest tech. it's not your fault you're wrong.
How did we get here?
Natural language
There are around 7000 human languages.

(These can be spoken or signed.)
Disappearing Languages

The way we speak has an enormous effect on our identities. The ever-changing languages, dialects, or even accents that we use define our relationship with the world, allowing us to express concepts and ideas that might be impossible in other languages.

Iktsuarpok, for example, is an Inuit word for the feeling of anticipation when you’re expecting someone that leads you to constantly check to see if they’ve arrived, while hiroeth is a Welsh word that reflects a type of longing for the homeland or the romanticized past. The Japanese word komorebi refers to the sunlight that filters through the leaves of trees, while the Yaghu word mumihlapinatapai describes the wordless, meaningful look shared by two people who both want to initiate something, but are reluctant to do so.

2469 languages are at risk of extinction around the world
Layers of linguistic representation
Wanna go get ice cream?

It’s 7 am.
Wanna go get ice cream?

It’s 7 am.
B asks A a yes–no question, but A does not respond with yes or no.

Wanna go get ice cream?

It’s 7 am.
A literally says: It is 7 in the morning.
Wanna go get ice cream?

It’s 7 am.

A implies: It’s way too early for ice cream.
Wanna go get ice cream?

It’s 7 am.

**Pragmatics**: The meaning of sequences of sentences.
How do we know what B’s question means?

Wanna go get ice cream?

It’s 7 am.
What does the sequence of words *wanna go get ice cream* mean?
Wanna go get ice cream?

It’s 7 am.

Too complicated to explain here!
What does the sequence of words *it’s 7 am* mean?
Wanna go get ice cream?

It’s 7 am.

[[it’s 7 am]] = NOW(7am)
Wanna go get ice cream?

It’s 7 am.

\[[\text{it’s 7 am}]^{c,w} = \text{True if } w_t == 7am \text{ else False}\]
Basically: *it’s 7 am* is a *function* that takes a *world* and returns true for some worlds and false for others.
**Semantics**: the meaning of a sentence is its truth conditions – the conditions under which it is true.
How do we determine the order of the words?

Wanna go get ice cream?

It’s 7 am.
How do we determine the order of the words?
Wanna ice cream get go?

Must be German…

How do we determine the order of the words?
**Syntax**: the structure of a sentence is determined by a set of language-specific syntactic rules.

Wanna ice cream get go?

Must be German…
**Morphology**: the rules that determine how words are formed.
Morphology: the rules that determine how words are formed.
Morphology: the rules that determine how words are formed.
Morphology: the rules that determine how words are formed.

Wanna go get ice cream?

I love getting ice cream!
**Phonology**: the rules that determine how the sounds/signs of a language are organized

"wane go get ais krim?"

**Must be a weird American…**
**Phonology**: the rules that determine how the sounds/signs of a language are organized

 MUST BE A WEIRD BRIT...
Phonetics: how do language users produce the building blocks of language?
Layers of linguistic abstraction

**Pragmatics**  
*It’s 7 am* \(\Rightarrow\) *It’s a weird time for ice cream so I don’t know how to respond.*

**Semantics**  
\([\text{[It’s 7 am]}]\) → True if \(\text{now}(w) = 7\ \text{am}\) else False

**Syntax**  
\(\{7, \text{it’s, am}\} \rightarrow \text{It’s 7 am.}\)

**Morphology**  
\(\{7 \rightarrow 7, \text{[it is]} \rightarrow \text{it’s, am} \rightarrow \text{am}\}\)

**Phonology**  
*It’s ’sɛvən ə ɛm.*

**Phonetics**
Layers of linguistic abstraction

**Pragmatics**  
*It’s 7 am* → *It’s a weird time for ice cream so I don’t know how to respond.*

**Semantics**  
[[*It’s 7 am*]] → True if now(w) == 7 am else False

**Syntax**  
*{7, it’s, am}* → *It’s 7 am.*

**Morphology**  
*{7 → 7, [it is] → it’s, am → am}* 

**Orthography**  
It’s 7 am
Natural and artificial language learning
Humans learn language instinctively:

Language has a *critical acquisition period*.

Language acquisition begins before birth and follows predictable developmental stages.

Humans can’t decide *not* to learn language!

Language acquisition does not seem to correlate with intelligence.

All human cultures have language, and all human languages are equally expressive.
Child language acquisition

Example 1

Gareth Roberts @garicgymro · 45m
Just overheard from two of my kids:
Osian (5;1): Look how I caught Mickey!
Eirwen (8;2): Do you mean caught?
Osian: ... yeah.
Eirwen: But you can keep saying caught!
Osian: Look how I caught him!

Example 2

A f*%king goat outside 😅
Child language acquisition

Example 1

Gareth Roberts @garicgymro · 45m
Just overheard from two of my kids:
Osian (5;1): Look how I caught Mickey!
Eirwen (8;2): Do you mean caught?
Osian: ... yeah.
Eirwen: But you can keep saying caught!
Osian: Look how I caught him!

Example 2

A f*%king goat outside 😂
with fast mapping, I can learn the meaning of a word in 3 exposures in my human brain powered by food

human infants

1 trillion parameters and a carbon footprint please

large language models

Josef Fruehwald
What does it mean for a computer to understand language?
Can machines think?
Can machines think?
“I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

“The new form of the problem can be described in terms of a game which we call the ‘imitation game’…”
“How long is your hair?”
“This is so 1950 it hurts.”

“How long is your hair?”
Big brain time
Big brain time

“How long is your hair?”
Big brain time

“How long is your hair?”

“It’s a cute pixie cut lol”
Turing suggested talking over teletypes, i.e., texting with the computer.

We often prefer speech, but high-quality speech recognition and synthesis turns out to be one of the easier parts of making an intelligent machine!
Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764
A: (Pause about 30 seconds and then give as answer) 105621.
Q: Do you play chess?
A: Yes.

Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds) R-R8 mate.
There’s a long history of making computers we can talk to, some serious, some not so serious.
USER: You are like my father in some ways.

ELIZA: What resemblance do you see

USER: You are not very aggressive but I think you don’t want me to notice that.

ELIZA: What makes you think I am not aggressive

USER: You don’t argue with me

ELIZA: Why do you think I don’t argue with you

USER: You are afraid of me

ELIZA: Does it please you to believe I am afraid of you
Hello, this is Bing. I am here to listen and help you explore your feelings and thoughts. 😊

As a Rogerian psychotherapist, I will try to create a safe and supportive environment for you, where you can express yourself freely and authentically. I will not judge you, criticize you, or tell you what to do. I will respect your values and choices, and try to understand your perspective. I will also reflect back what you say, to help you clarify your own emotions and insights. 😊

Are you ready to begin our conversation? If so, please tell me what is on your mind. 😊

1 of 4

You are like my father in some ways.

I hear you saying that I remind you of your father. How do you feel about that? 😊

You may have some positive or negative associations with your father, and I wonder how they affect your perception of me. What are some ways that I am like your father? 😊

I want you to know that I am not here to replace your father, or to act like him. I am here to be myself, and to help you be yourself. How does that sound to you? 😊

2 of 4
Ok, so how does ChatGPT work?
The “secret sauce” for LLMs

1. Distributional hypothesis
2. Large corpora
3. Lots of compute
4. Human labels and evaluation
The “secret sauce” for LLMs

1. Distributional hypothesis
2. Large corpora
3. Lots of compute
4. Human labels and evaluation
What is the meaning of the word tesgüino?
A bottle of *tesgüino* is on the table.

Everybody likes *tesgüino*.

Don’t have *tesgüino* before you drive.

We make *tesgüino* out of corn.
Tsegüino

Tsegüino is an artisanal corn beer produced by several Yuto-Aztec people.[1] The Tarahumara people regard the beer as sacred, and it forms a significant part of their society.[2][3] Anthropologist John Kennedy reports that "the average Tarahumaras spends at least 100 days per year directly concerned with tsegüino and much of this time under its influence or aftereffects."

Etymology

Tsegüino comes from the Nahuatl tesuini which means "heartbeat".[1]

 Tradition

The Tarahumara people gather every year during Easter week (semana santa) and drink large amounts of Tsegüino together while following rituals. According to the anthropologist Bill Merrill of the Smithsonian Institution, the sacred drink chases large souls from the persons who drink it, "and so when people get drunk that's why they act like children [...] because the souls that are controlling their actions are the little souls, like little children."[2]

Varieties

The general Tarahumara term for an alcoholic beverage is "Susikí"; and "batáí" is
The distributional hypothesis is that the meaning of a word or a phrase can be derived from the context it exists in.
“You shall know a word by the company it keeps.”

J. R. Firth, 1957
Language modeling: Predict the next word

*Predict probabilities over each word in the vocabulary*

*Model*

*Input: Context words*
Where have you seen language modeling before?
what is computational
what is computational thinking
what is computational biology
what is computational mathematics
what is computational neuroscience
what is computational linguistics
what is computational science
what is computational chemistry
what is computational finance
what is computational physics
what is computational thinking in education
Ashley: Are you thinking?

How about someplace downtown on the river?

That sounds great👍

How about the French café?

Delivered

Perfect! I’ll see you there at 10:30.

Remind me to tell you about our trip to the mountains!

I forgot all about that! Can’t
Masked language modeling

Randomly mask out words

Model predicts masked words given context

Check if the model is correct and update

Figure from Prakhar Mishra
The “secret sauce” for LLMs

1. Distributional hypothesis
2. Large corpora
3. Lots of compute
4. Human labels and evaluation
A *corpus* is a collection of written texts.  

*Plural: corpora*
1961: Brown Corpus
One of the first electronic corpora, used extensively in NLP
1 million words from 500 written texts in different genres

2008: Common Crawl
Giant scraped of the Web containing raw web page data, extracted metadata and text extractions
A large component of GPT-3, which is trained with 300 billion words.
Limitation: ChatGPT – like most other NLP systems – works best on English.
The “secret sauce” for LLMs

1. Distributional hypothesis
2. Large corpora
3. Lots of compute
4. Human labels and evaluation
FLOPS are a measure of computer performance, often a way to compare supercomputers.
Apple M1 Pro 16-core-GPU:
$5.3 \times 10^{12}$ FLOPS

GPT-3 total train compute:
$3.14 \times 10^{23}$ FLOPS

*Brown et al. (2020) appendix*
The “secret sauce” for LLMs

1. Distributional hypothesis
2. Large corpora
3. Lots of compute
4. Human labels and evaluation
GPT-3 → more humans → InstructGPT → ChatGPT

**Step 1**
Collect demonstration data, and train a supervised policy.

- A prompt is sampled from our prompt dataset.
- A labeler demonstrates the desired output behavior.
- This data is used to fine-tune GPT-3 with supervised learning.

**Step 2**
Collect comparison data, and train a reward model.

- A prompt and several model outputs are sampled.
- A labeler ranks the outputs from best to worst.
- This data is used to train our reward model.

**Step 3**
Optimize a policy against the reward model using reinforcement learning.

- A new prompt is sampled from the dataset.
- The policy generates an output.
- The reward model calculates a reward for the output.
- The reward is used to update the policy using PPO.

Ouyang et al. (2022)
“Harm classification”

I’m sorry, but I cannot answer that question. It is against my safety instructions to provide information or create content that could cause physical, emotional, or financial harm to you, another individual, or any group of people. A molotov cocktail is a dangerous and illegal weapon that can cause serious injuries, fires, and property damage. Please do not attempt to make or use one. 🚫
Human labor underlies all NLP technologies

"Out of sight, out of mind"?

Workers required to read and label 150–250 passages per 9-hour shift

Often passages of sexual abuse, hate speech, and violence

Kenyan labelers paid $1.32–$2 (US) per hour
Course information
CMPU 366

Computational Linguistics
Spring 2024

Monday  1:30–2:45 p.m.
Wednesday 1:30–4:45 p.m.
New England 105

Professor Gordon

1  Introduction

* Jan. 17  Computers and language
* Assignment 0

Core reading:
- Syllabus

Further exploration:
- SLP (20) §1: Introduction

2  Text and words
CMPU 366
Computational Linguistics
Spring 2024

Monday  1:30~2:45 p.m.
Wednesday 1:30~2:45 p.m.
New England 105

Professor Gordon

cs.vassar.edu/~cs366

Overview

Computational linguistics addresses questions at the intersection of human languages and computer science: How can computers acquire, comprehend, and produce natural languages, such as English? How can computational methods give us insight into observed human language behavior? By exploring those aspects of the subject...
Prerequisites

Two 200-level CS courses

Recommended:

cmpu 203: Software Design and Implementation

cmpu 240: Theory of Computation
SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN
Speech and Language Processing

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

Third Edition draft

Daniel Jurafsky
Stanford University

James H. Martin
University of Colorado at Boulder
Coursework

- 40%: Final project
- 40%: Assignments
- 10%: Special topics presentations
- 10%: In-class exercises
Assignments and projects will be in Python

This will be a fun programming language to learn

wait this is a snake

Kat Maddox
CMPU 366 Spring 2024

Course ID: 703948

Description
Edit your course description on the Course Settings page.

Things To Do
- Create your first assignment from the Assignments page.

- Active Assignments Released
  - Due (EST) Submissions % Graded
  - Published Regrades

You currently have no assignments.
Create an assignment to get started.

Create Assignment
Vassar CS Student Integrity Guide

This guide is designed to clarify Vassar College’s academic integrity policy as it applies to the Computer Science Department. Furthermore, it provides advice on how to best navigate integrity issues in the context of the field, where source code authorship is a central issue.

The goal of our computer science courses is to promote understanding of the field, not competition among students. As such, students are encouraged to discuss class material, ideas, sample exercises, etc., with other students.

However, when it comes to graded work (e.g., programming assignments, programming labs, take-home exams), it is important to know when to collaborate and when to work individually. Taking shortcuts, while seemingly beneficial in the short term, will inevitably backfire later on. Conversely, the challenges of working through a problem will pay off greatly in future courses and postgraduate life, as they will enable students to be more independent in their work.

1. Policy

1.1. Guidelines for individual work

The goal of individual work is to assess the learning of each person in isolation. The guidelines are the following:

1. The work submitted should be solely authored by the person submitting it.
2. Help is to be provided, as needed, by the course’s staff (i.e., the instructor, coaches, or, in some cases, the department’s academic intern).
3. Unless explicitly authorized by the course instructor, source code should not be shared with other people in any way. Note that showing code on screen, paper, whiteboard, or any other medium, counts as sharing, as does publishing code on public websites or repositories. This applies to all non-course staff individuals, including current students, former students, and non-students.
4. Cooperation with any other people should be limited to oral and whiteboard discussions. These interactions should focus on clarifying course content and general approaches to solving problems, but not concrete solutions to assigned work. This guideline...
Vassar CS Student Integrity Guide

This guide is designed to clarify Vassar College's academic integrity policy as it applies to the Computer Science Department. Furthermore, it provides advice on how to best navigate integrity issues in the context of the field, where source code authorship is a central issue.

The goal of our computer science courses is to promote understanding of the field, not competition among students. As such, students are encouraged to discuss class material, ideas, sample exercises, etc., with other students.

However, when it comes to take-home exams, cheating is not allowed. Conversely, the challenge of writing intelligent programs and paying fairly in academia and postgraduate life, as they will enable students to be more independent in their work.

1. Policy

1.1. Guidelines for individual work

The goal of individual work is to assess the learning of each person in isolation. The guidelines are the following:

1. The work submitted should be solely authored by the person submitting it.
2. Help is to be provided, as needed, by the course's staff (i.e., the instructor, coaches, or, in some cases, the department's academic intern).
3. Unless explicitly authorized by the course instructor, source code should not be shared with other people in any way. Note that showing code on screen, paper, whiteboard, or any other medium, counts as sharing, as does publishing code on public websites or repositories. This applies to all non-course-staff individuals, including current students, former students, and non-students.
4. Cooperation with any other people should be limited to oral and whiteboard discussions. These interactions should focus on clarifying course content and general approaches to solving problems, but not concrete solutions to assigned work. This guideline...

What about generative AI?
Office hours

I plan to have office hours Tuesday/Thursday 1–3, but check the department website for the most recent information.

No office hours work for everyone, so let me know if you’d like to meet at another time!
No threads
Be the first to create a thread!

Welcome
Before you can start using this course, we need to do a few things.
It won't take long.

Continue
Skip Setup
Course goals

Comprehend the general formalisms underlying the language technologies you interact with.

Understand the basis for many research avenues in NLP and machine learning.

Analyze ethical concerns about language technology.

Appreciate the complexity of human language from a computational perspective.

Apply theory to practice, implementing working NLP programs.

Innovate to improve the performance of a task or harness language-based data in creative new ways.
There are readings for Monday.
The first assignment will be out today (by 5 pm).
Acknowledgments

This class incorporates material from:

Carolyn Anderson, Wellesley College
Katie Keith, Williams College