

# 1. Intro & Word vectors

LING-581-Natural Language Processing1

---

Instructor: Hakyung Sung

August 25, 2025

\*Acknowledgment: These course slides are based on materials from CS224N: NLP with Deep Learning @ Stanford University.

# Table of contents

1. Introduction
2. Lesson plan
3. Human language & Word meaning
4. Encoding and embedding
5. Wrap-up

# Introduction

---

- Instructor: Dr. Hakyung Sung

- Instructor: Dr. Hakyung Sung
  - PhD in Linguistics, MS in Computer Science @ University of Oregon

# Course logistics

- Instructor: Dr. Hakyung Sung
  - PhD in Linguistics, MS in Computer Science @ University of Oregon
- Grader: Bea (Bey-uh) Pulido

# Course logistics

- Instructor: Dr. Hakyung Sung
  - PhD in Linguistics, MS in Computer Science @ University of Oregon
- Grader: Bea (Bey-uh) Pulido
  - a second year Experimental Psychology graduate student

# Course logistics

- Instructor: Dr. Hakyung Sung
  - PhD in Linguistics, MS in Computer Science @ University of Oregon
- Grader: Bea (Bey-uh) Pulido
  - a second year Experimental Psychology graduate student
- Time: Tu/Th 12:30PM-1:45PM



# Course logistics

- Instructor: Dr. Hakyung Sung
  - PhD in Linguistics, MS in Computer Science @ University of Oregon
- Grader: Bea (Bey-uh) Pulido
  - a second year Experimental Psychology graduate student
- Time: Tu/Th 12:30PM-1:45PM
- Office: EAS 3173

# Course logistics

- Instructor: Dr. Hakyung Sung
  - PhD in Linguistics, MS in Computer Science @ University of Oregon
- Grader: Bea (Bey-uh) Pulido
  - a second year Experimental Psychology graduate student
- Time: Tu/Th 12:30PM-1:45PM
- Office: EAS 3173
- Office hour: TuTh 3:30-4:30 in-person, or Zoom by appointment

# Course logistics

- Instructor: Dr. Hakyung Sung
  - PhD in Linguistics, MS in Computer Science @ University of Oregon
- Grader: Bea (Bey-uh) Pulido
  - a second year Experimental Psychology graduate student
- Time: Tu/Th 12:30PM-1:45PM
- Office: EAS 3173
- Office hour: TuTh 3:30-4:30 in-person, or Zoom by appointment
- Email: [hks gla@rit.edu](mailto:hks gla@rit.edu)

# Course logistics

- Instructor: Dr. Hakyung Sung
  - PhD in Linguistics, MS in Computer Science @ University of Oregon
- Grader: Bea (Bey-uh) Pulido
  - a second year Experimental Psychology graduate student
- Time: Tu/Th 12:30PM-1:45PM
- Office: EAS 3173
- Office hour: TuTh 3:30-4:30 in-person, or Zoom by appointment
- Email: [hks gla@rit.edu](mailto:hks gla@rit.edu)
- Course website:  
[https://hksung.github.io/Fall25\\_LING581/](https://hksung.github.io/Fall25_LING581/)

# Learning goals

- The foundations of the effective modern methods for deep learning applied to NLP

# Learning goals

- The foundations of the effective modern methods for deep learning applied to NLP
  - Basics first: word vectors, feed-forward networks, recurrent networks, attention

# Learning goals

- The foundations of the effective modern methods for deep learning applied to NLP
  - Basics first: word vectors, feed-forward networks, recurrent networks, attention
  - Then key methods used in NLP: encoder-decoder models, transformers, pre-training, post-training, benchmark and evaluation, NLP applications (to language research), etc.

# Learning goals

- A big picture understanding of human languages and the difficulties in understanding and producing them via computers



# Learning goals

- A big picture understanding of human languages and the difficulties in understanding and producing them via computers
- An understanding of an application to build systems for some of the major problems in NLP: Word meaning, dependency parsing, machine translation, question answering

# Learning goals

- A big picture understanding of human languages and the difficulties in understanding and producing them via computers
- An understanding of an application to build systems for some of the major problems in NLP: Word meaning, dependency parsing, machine translation, question answering
- Hands-on exercises conducted during classes (on Thursday)

# Learning goals

- A big picture understanding of human languages and the difficulties in understanding and producing them via computers
- An understanding of an application to build systems for some of the major problems in NLP: Word meaning, dependency parsing, machine translation, question answering
- Hands-on exercises conducted during classes (on Thursday)
- Opportunities to connect NLP techniques to specific domains of interest using language data (Final project)

# Final grading components

[a × b] a = number; b = points

- Lab exercises [8 × 5]: 40%
- Background research 20%
  - Assignment [1 × 10] 10%
  - Presentation [1 × 10] 10%
- Final project 40%
  - Final project proposal [1 × 10] 10%
  - Final presentation [1 × 15]: 15%
  - Final paper [1 × 15]: 15%

# Final grading components

- Lab exercises [ $8 \times 5$ ]: 40%

Week	Date	Topic	Due (Friday, 11:59 pm)
1	8/26	Introduction, Word vectors	
	8/28	Lab1 – Python basics	Lab exercise 1
2	9/2	Word vectors	
	9/4	Lab2 – Word vectors	Lab exercise 2
3	9/9	Backpropagation, neural network basics	
	9/11	Lab 3 – PyTorch	Lab exercise 3

# Final grading components

- Lab exercises [8 × 5]: 40%

Week	Date	Topic	Due (Friday, 11:59 pm)
1	8/26	Introduction, Word vectors	
	8/28	Lab1 – Python basics	Lab exercise 1
2	9/2	Word vectors	
	9/4	Lab2 – Word vectors	Lab exercise 2
3	9/9	Backpropagation, neural network basics	
	9/11	Lab 3 – PyTorch	Lab exercise 3

- Please bring your laptop on Thursday.

# Final grading components

- Lab exercises [8 × 5]: 40%

Week	Date	Topic	Due (Friday, 11:59 pm)
1	8/26	Introduction, Word vectors	
	8/28	Lab1 – Python basics	Lab exercise 1
2	9/2	Word vectors	
	9/4	Lab2 – Word vectors	Lab exercise 2
3	9/9	Backpropagation, neural network basics	
	9/11	Lab 3 – PyTorch	Lab exercise 3

- Please bring your laptop on Thursday.
- Students are expected to complete their exercises during class and submit their answers before the class ends.

# Final grading components

- Lab exercises [8 × 5]: 40%

Week	Date	Topic	Due ( <b>Friday</b> , 11:59 pm)
1	8/26	Introduction, Word vectors	
	8/28	Lab1 – Python basics	Lab exercise 1
2	9/2	Word vectors	
	9/4	Lab2 – Word vectors	Lab exercise 2
3	9/9	Backpropagation, neural network basics	
	9/11	Lab 3 – PyTorch	Lab exercise 3

- Please bring your **laptop on Thursday**.
- Students are expected to complete their exercises **during class** and **submit their answers before the class ends**.
- If not possible, the official due date is **Friday** of the same week.



# Final grading components

- Background research 20%
  - Assignment [1 × 10] 10%
  - Presentation [1 × 10] 10%
- [https://youtube.com/shorts/Yg7WrDt5I1E?si=12YMKYi\\_OJRj9c6r](https://youtube.com/shorts/Yg7WrDt5I1E?si=12YMKYi_OJRj9c6r)

- To apply NLP technologies to a given domain, we need at least a basic understanding of that domain (and ideally, a more advanced one).

# Final grading components

- To apply NLP technologies to a given domain, we need at least a basic understanding of that domain (and ideally, a more advanced one).
- Think about the areas you are interested in. (e.g., Language learning and NLP applications)

# Final grading components

- To apply NLP technologies to a given domain, we need at least a basic understanding of that domain (and ideally, a more advanced one).
- Think about the areas you are interested in. (e.g., Language learning and NLP applications)
- At the end of this class, I will give you some time to reflect on this; Submit your area of interest (it can be broad), and I will form groups of four.

# Final grading components

- To apply NLP technologies to a given domain, we need at least a basic understanding of that domain (and ideally, a more advanced one).
- Think about the areas you are interested in. (e.g., Language learning and NLP applications)
- At the end of this class, I will give you some time to reflect on this; Submit your area of interest (it can be broad), and I will form groups of four.
- In **Week 7**, you will submit a more specific research topic.

# Final grading components

- To apply NLP technologies to a given domain, we need at least a basic understanding of that domain (and ideally, a more advanced one).
- Think about the areas you are interested in. (e.g., Language learning and NLP applications)
- At the end of this class, I will give you some time to reflect on this; Submit your area of interest (it can be broad), and I will form groups of four.
- In **Week 7**, you will submit a more specific research topic.
- In **Week 13**, you will give a **presentation** about the background research. This presentation will be connected to your **Assignment**.

# Final grading components

- **Final project 40%**
  - Final project proposal [ $1 \times 10$ ]: 10%
  - Final presentation [ $1 \times 15$ ]: 15%
  - Final paper [ $1 \times 15$ ]: 15%

# Final grading components

- **Final project 40%**
  - Final project proposal [1 × 10] 10%
  - Final presentation [1 × 15]: 15%
  - Final paper [1 × 15]: 15%
- Based on the problems and prior solutions you identified in your background research, you will design a final project that applies NLP techniques to address the problem (\*Continue working on the same group)



# Final grading components

- **Final project 40%**
  - Final project proposal [1 × 10] 10%
  - Final presentation [1 × 15]: 15%
  - Final paper [1 × 15]: 15%
- Based on the problems and prior solutions you identified in your background research, you will design a final project that applies NLP techniques to address the problem (\*Continue working on the same group)
- The project proposal is due in **Week 9**.

# Final grading components

- **Final project 40%**
  - Final project proposal [1 × 10] 10%
  - Final presentation [1 × 15]: 15%
  - Final paper [1 × 15]: 15%
- Based on the problems and prior solutions you identified in your background research, you will design a final project that applies NLP techniques to address the problem (\*Continue working on the same group)
- The project proposal is due in **Week 9**.
- The final presentation and paper are due in **Week 16**.

- This process actually follows the typical research flow:

- This process actually follows the typical research flow:
  1. Review previous research

- This process actually follows the typical research flow:
  1. Review previous research
  2. Identify a research gap

- This process actually follows the typical research flow:
  1. Review previous research
  2. Identify a research gap
  3. Apply proposed techniques / conduct analysis

- This process actually follows the typical research flow:
  1. Review previous research
  2. Identify a research gap
  3. Apply proposed techniques / conduct analysis
  4. Present findings

- This process actually follows the typical research flow:
  1. Review previous research
  2. Identify a research gap
  3. Apply proposed techniques / conduct analysis
  4. Present findings
  5. Write up the paper



# Grading policy

- **2-hr grading window:** Any assignment submitted online will automatically have a 2-hour grading window. This will be applied by the system, and no action is required from students.

# Grading policy

- **2-hr grading window:** Any assignment submitted online will automatically have a 2-hour grading window. This will be applied by the system, and no action is required from students.
- **Late penalty:** Late assignments will incur a 10% deduction per day, for up to 5 days (e.g., 1 day late = 10% off). After 5 days, the assignment will receive a grade of zero.

# Grading policy

- **2-hr grading window:** Any assignment submitted online will automatically have a 2-hour grading window. This will be applied by the system, and no action is required from students.
- **Late penalty:** Late assignments will incur a 10% deduction per day, for up to 5 days (e.g., 1 day late = 10% off). After 5 days, the assignment will receive a grade of zero.
- **Extenuating circumstances:** Whenever possible, please request an official document that can prove the circumstances—this allows me to accommodate you fairly while respecting your privacy.

# Grading policy

- **2-hr grading window:** Any assignment submitted online will automatically have a 2-hour grading window. This will be applied by the system, and no action is required from students.
- **Late penalty:** Late assignments will incur a 10% deduction per day, for up to 5 days (e.g., 1 day late = 10% off). After 5 days, the assignment will receive a grade of zero.
- **Extenuating circumstances:** Whenever possible, please request an official document that can prove the circumstances—this allows me to accommodate you fairly while respecting your privacy.
  - If that is not possible, contact me as soon as you can. Extensions are generally not granted retroactively.

# Grading policy

- **2-hr grading window:** Any assignment submitted online will automatically have a 2-hour grading window. This will be applied by the system, and no action is required from students.
- **Late penalty:** Late assignments will incur a 10% deduction per day, for up to 5 days (e.g., 1 day late = 10% off). After 5 days, the assignment will receive a grade of zero.
- **Extenuating circumstances:** Whenever possible, please request an official document that can prove the circumstances—this allows me to accommodate you fairly while respecting your privacy.
  - If that is not possible, contact me as soon as you can. Extensions are generally not granted retroactively.
- **No extensions will be granted for the final paper.**

- For the group works, all members are expected to contribute their time and effort equally.

- For the group works, all members are expected to contribute their time and effort equally.
- Each submission will include a section outlining both individual and group contributions, which will be evaluated separately.

- For the group works, all members are expected to contribute their time and effort equally.
- Each submission will include a section outlining both individual and group contributions, which will be evaluated separately.
- Collaboration with AI tools is permitted, but you are responsible for the quality and integrity of the work produced.



- For the group works, all members are expected to contribute their time and effort equally.
- Each submission will include a section outlining both individual and group contributions, which will be evaluated separately.
- Collaboration with AI tools is permitted, but you are responsible for the quality and integrity of the work produced.
- You must acknowledge and document how AI tools were used in your work (including individual exercises).

Any questions?

## Lesson plan

---

## 1. Course logistics

1. ~~Course logistics~~
2. Human language and word meaning

1. Course logistics
2. Human language and word meaning
3. Encoding and embedding: Encoding

1. ~~Course logistics~~
2. Human language and word meaning
3. Encoding and embedding: Encoding
4. Wrap up

1. ~~Course logistics~~
2. Human language and word meaning
3. Encoding and embedding: Encoding
4. Wrap up



1. Course logistics
2. Human language and word meaning
3. Encoding and embedding: Encoding
4. Wrap up

**Key idea:** Language and writing are remarkable technologies; NLP problems begin with encoding meaning in computers.

## Human language & Word meaning

---

- Enables us to tell stories, ask questions, share knowledge, plan ahead, and even imagine alternate realities.

# Language is *technology*

- Enables us to tell stories, ask questions, share knowledge, plan ahead, and even imagine alternate realities.
- Serves as a tool for coordinating with others, transmitting ideas, and building shared culture.

# Language is *technology*

- Enables us to tell stories, ask questions, share knowledge, plan ahead, and even imagine alternate realities.
- Serves as a tool for coordinating with others, transmitting ideas, and building shared culture.
- Universally present across all human societies (regardless of geography or historical period).

# Language is *technology*

- Enables us to tell stories, ask questions, share knowledge, plan ahead, and even imagine alternate realities.
- Serves as a tool for coordinating with others, transmitting ideas, and building shared culture.
- Universally present across all human societies (regardless of geography or historical period).
- Estimated age: **100,000–200,000 years**, making it one of the oldest and most powerful human inventions.

# Language is *technology*

- Enables us to tell stories, ask questions, share knowledge, plan ahead, and even imagine alternate realities.
- Serves as a tool for coordinating with others, transmitting ideas, and building shared culture.
- Universally present across all human societies (regardless of geography or historical period).
- Estimated age: **100,000–200,000 years**, making it one of the oldest and most powerful human inventions.
- Evidence includes archaeological findings such as *symbolic beads, tools, and burial sites*, which suggest abstract thought and communication.



**Figure 1:** Clay tablet inscribed with the earliest known writing system, cuneiform—recording the receipt of barley and malt (around 3000 BCE, left)—and a close-up of cuneiform text on a mudbrick (around 1200 BCE, right).

Sourced from: <https://en.wikipedia.org/wiki/Cuneiform>



# What is *Writing*?

Writing = another amazing technology!

- Records language, which is otherwise *ephemeral*

# What is *Writing*?

Writing = another amazing technology!

- Records language, which is otherwise *ephemeral*
- Makes language usable across time and space

# What is *Writing*?

Writing = another amazing technology!

- Records language, which is otherwise *ephemeral*
- Makes language usable across time and space
- Enables wide communication—even with strangers!

# What is *Writing*?

Writing = another amazing technology!

- Records language, which is otherwise *ephemeral*
- Makes language usable across time and space
- Enables wide communication—even with strangers!
- Key to history, law, science, culture

# What is *Writing*?

## Writing = another amazing technology!

- Records language, which is otherwise *ephemeral*
- Makes language usable across time and space
- Enables wide communication—even with strangers!
- Key to history, law, science, culture
- Not all people or societies use writing

# What is *Writing*?

## Writing = another amazing technology!

- Records language, which is otherwise *ephemeral*
- Makes language usable across time and space
- Enables wide communication—even with strangers!
- Key to history, law, science, culture
- Not all people or societies use writing
- Estimated age: 5,000–6,000 years

What is NLP?

## What is NLP?

Natural Language Processing (NLP) is the field that enables computers to understand, analyze, and generate human language.



## What is NLP?

Natural Language Processing (NLP) is the field that enables computers to understand, analyze, and generate human language.

- To process language with computers, NLP requires a way to **encode language** → that's where **writing systems** come in!

## What is NLP?

Natural Language Processing (NLP) is the field that enables computers to understand, analyze, and generate human language.

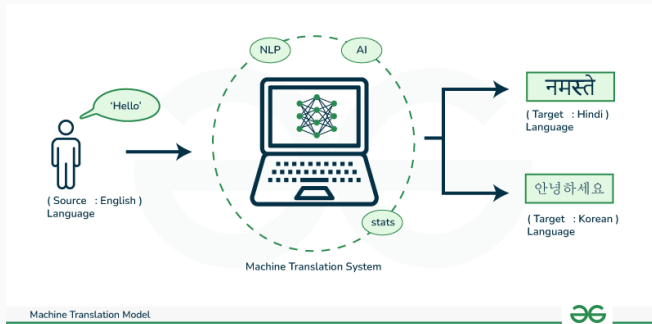
- To process language with computers, NLP requires a way to **encode language** → that's where **writing systems** come in!
- Evolution of writing technologies: **clay** → **papyrus** → **printing press** → **digital text**

## What is NLP?

Natural Language Processing (NLP) is the field that enables computers to understand, analyze, and generate human language.

- To process language with computers, NLP requires a way to **encode language** → that's where **writing systems** come in!
- Evolution of writing technologies: **clay** → **papyrus** → **printing press** → **digital text**
- Digital writing allows for new forms of communication and makes language **machine-readable** → **Leap!**

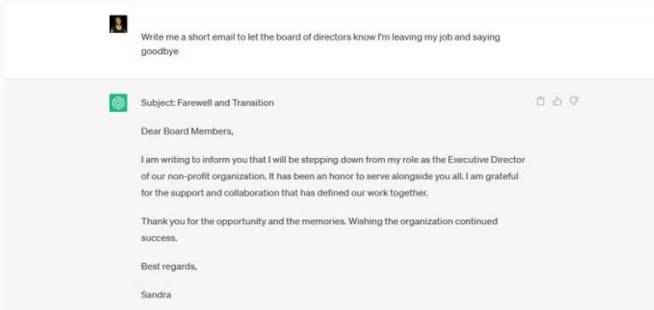
# Machine translations



Sourced from:

<https://www.geeksforgeeks.org/nlp/machine-translation-of-languages-in-artificial-intelligence/>

# Generating texts



Sourced from: <https://www.makeuseof.com/use-chatgpt-write-work-emails/>

# Generating image

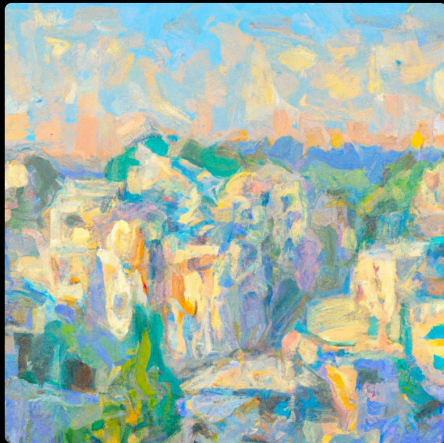
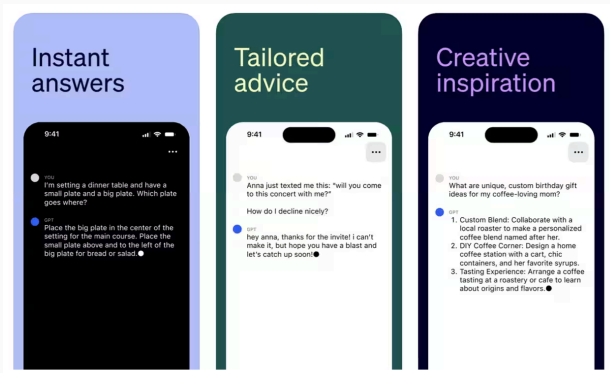


Illustration: Justin Jay Wang × DALL-E

Sourced from: <https://openai.com/about/>



Sourced from:

<https://www.theguardian.com/technology/2024/jan/10/openai-launches-gpt-store-customized-chatbots>

# How do we represent the *meanings*

- Definition: *meaning* (Webster)



# How do we represent the *meanings*

- Definition: *meaning* (Webster)

# How do we represent the *meanings*

- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.

# How do we represent the *meanings*

- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.

# How do we represent the *meanings*

- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.
  - the idea that a person wants to express by using words, signs

# How do we represent the *meanings*

- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.
  - the idea that a person wants to express by using words, signs

# How do we represent the *meanings*

- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.
  - the idea that a person wants to express by using words, signs
  - the idea that is expressed in a work of writing, art, etc.

# How do we represent the *meanings*

- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.
  - the idea that a person wants to express by using words, signs
  - the idea that is expressed in a work of writing, art, etc.

# How do we represent the *meanings*

- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.
  - the idea that a person wants to express by using words, signs
  - the idea that is expressed in a work of writing, art, etc.
- Commonest linguistic way of thinking about meaning:



# How do we represent the *meanings*

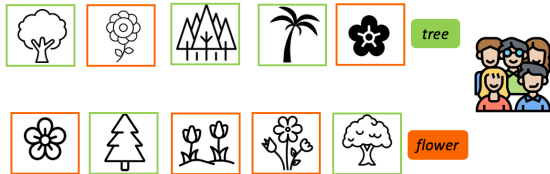
- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.
  - the idea that a person wants to express by using words, signs
  - the idea that is expressed in a work of writing, art, etc.
- Commonest linguistic way of thinking about meaning:
  - **Signifier** (symbol) ↔ **Signified** (idea or thing)

# How do we represent the *meanings*

- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.
  - the idea that a person wants to express by using words, signs
  - the idea that is expressed in a work of writing, art, etc.
- Commonest linguistic way of thinking about meaning:
  - **Signifier** (symbol) ↔ **Signified** (idea or thing)

# How do we represent the *meanings*

- Definition: *meaning* (Webster)
  - the idea that is represented by a word, phrase, etc.
  - the idea that a person wants to express by using words, signs
  - the idea that is expressed in a work of writing, art, etc.
- Commonest linguistic way of thinking about meaning:
  - **Signifier** (symbol) ↔ **Signified** (idea or thing)
  - **Denotational semantics**



# How do we represent the meanings *in computer*

Can computers understand meanings of the words as we do?

# How do we represent the meanings *in computer*

Can computers understand meanings of the words as we do?

No.

# How do we represent the meanings *in computer*

Can computers understand meanings of the words as we do?

No.

Traditional NLP method: Use the sets of synonyms and hypernyms of word by querying some databases (e.g., *WordNet*)

## Problems with the traditional method (like WordNet)

- Missing nuances (e.g., *proficient* is listed as a synonym of *good* (but not always))

## Problems with the traditional method (like WordNet)

- Missing nuances (e.g., *proficient* is listed as a synonym of *good* (but not always))
- Missing new meanings of words (e.g., *rizz*)



## Problems with the traditional method (like WordNet)

- Missing nuances (e.g., *proficient* is listed as a synonym of *good* (but not always))
- Missing new meanings of words (e.g., *rizz*)
  - Word meanings constantly change and adapt based on how people really use the language in the world

## Problems with the traditional method (like WordNet)

- Missing nuances (e.g., *proficient* is listed as a synonym of *good* (but not always))
- Missing new meanings of words (e.g., *rizz*)
  - Word meanings constantly change and adapt based on how people really use the language in the world
- Practically, building/updating a database is expensive and inefficient.

## Problems with the traditional method (like WordNet)

- Missing nuances (e.g., *proficient* is listed as a synonym of *good* (but not always))
- Missing new meanings of words (e.g., *rizz*)
  - Word meanings constantly change and adapt based on how people really use the language in the world
- Practically, building/updating a database is expensive and inefficient.
- Can't compute accurate word similarity

# Encoding and embedding

---

- In traditional NLP, we regard words as discrete symbols

- In traditional NLP, we regard words as discrete symbols
- Words themselves cannot be given as inputs to computers

- In traditional NLP, we regard words as discrete symbols
- Words themselves cannot be given as inputs to computers
- BUT numbers can be given as inputs to computers

- In traditional NLP, we regard words as discrete symbols
- Words themselves cannot be given as inputs to computers
- BUT numbers can be given as inputs to computers



- In traditional NLP, we regard words as discrete symbols
- Words themselves cannot be given as inputs to computers
- BUT numbers can be given as inputs to computers
- Encoding = converting words to numbers

- In traditional NLP, we regard words as discrete symbols
- Words themselves cannot be given as inputs to computers
- BUT numbers can be given as inputs to computers
- Encoding = converting words to numbers
- *Will continue discussions on encoding/embedding and word vectors next Tuesday.*

## Wrap-up

---

- Please check the syllabus; let know if you have any question(s).

- Please check the syllabus; let know if you have any question(s).
- Lab session 1 is planned on Thursday, which we will go over some basic Python functions (If you want to brush up your Python skills, this session will be helpful.)

- Please check the syllabus; let know if you have any question(s).
- Lab session 1 is planned on Thursday, which we will go over some basic Python functions (If you want to brush up your Python skills, this session will be helpful.)
- A mini survey for group projects:  
<https://forms.gle/4dtPDFFhDpccfvBu8>