LG 467 Computers in Linguistics

Semester 1-2021

Sakol Suethanapornkul

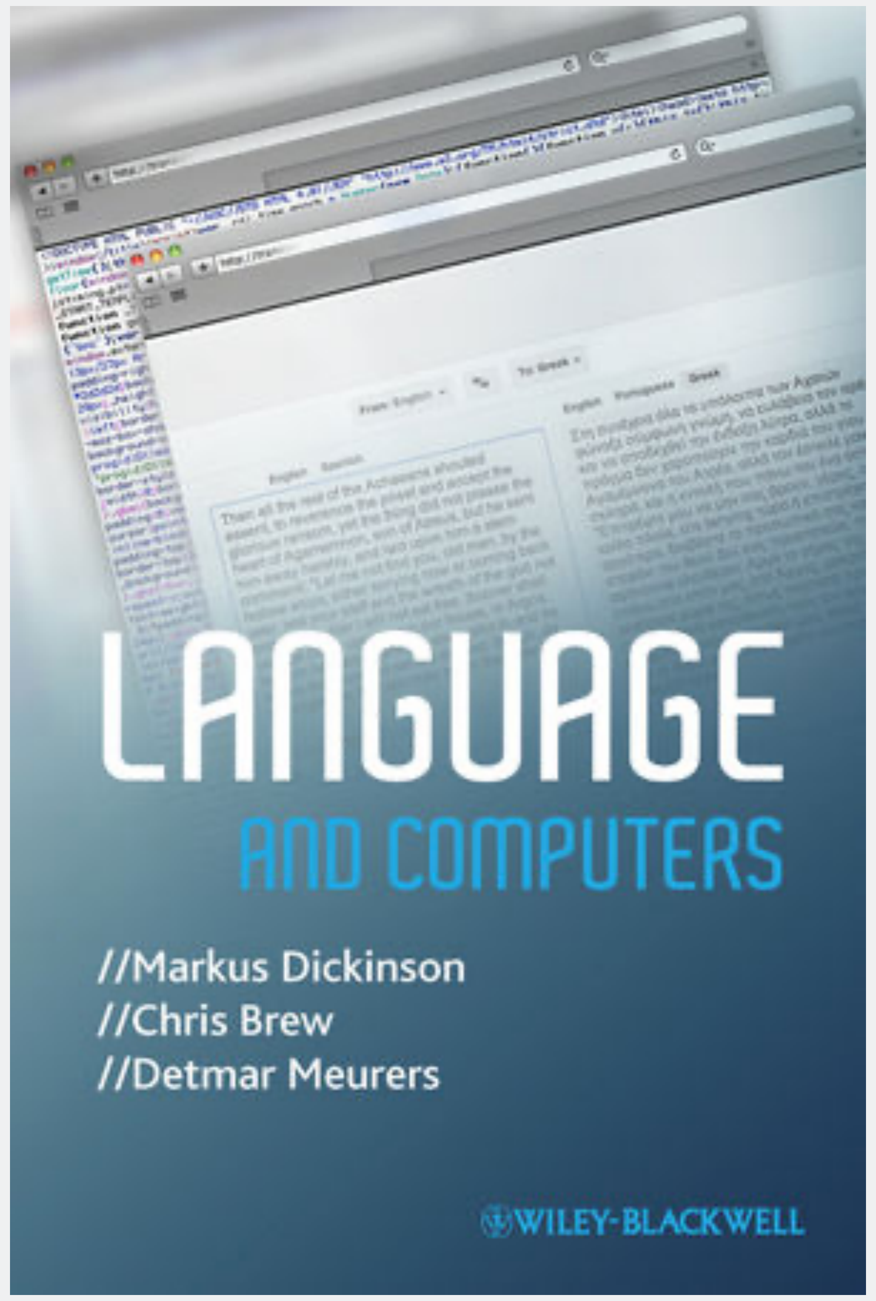
✉

Course Information

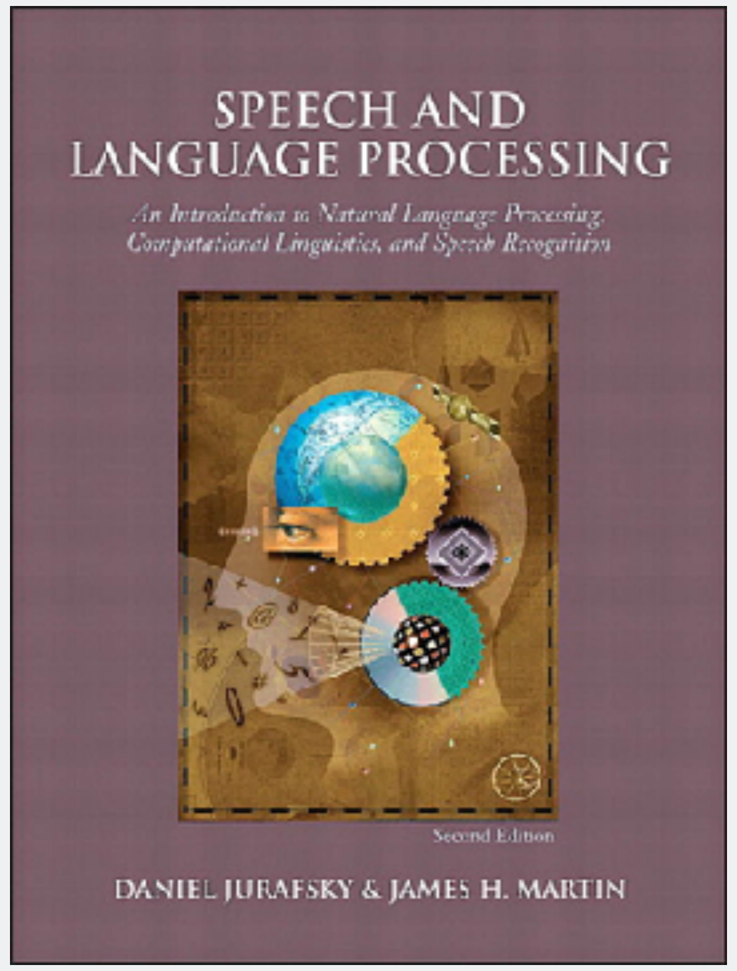
• Instructor:	Sakol Suethanapornkul
• Office hours:	W & TH 1 pm – 4 pm
• Course platform:	Microsoft Teams (Link)
• Course website:	Follow this link to the site!
• Communication:	Teams chat & email (suesakol@staff.tu.ac.th)
• Registration:	Add-drop period from Aug 5 to Aug 12

Before we dive deeper...

Textbooks

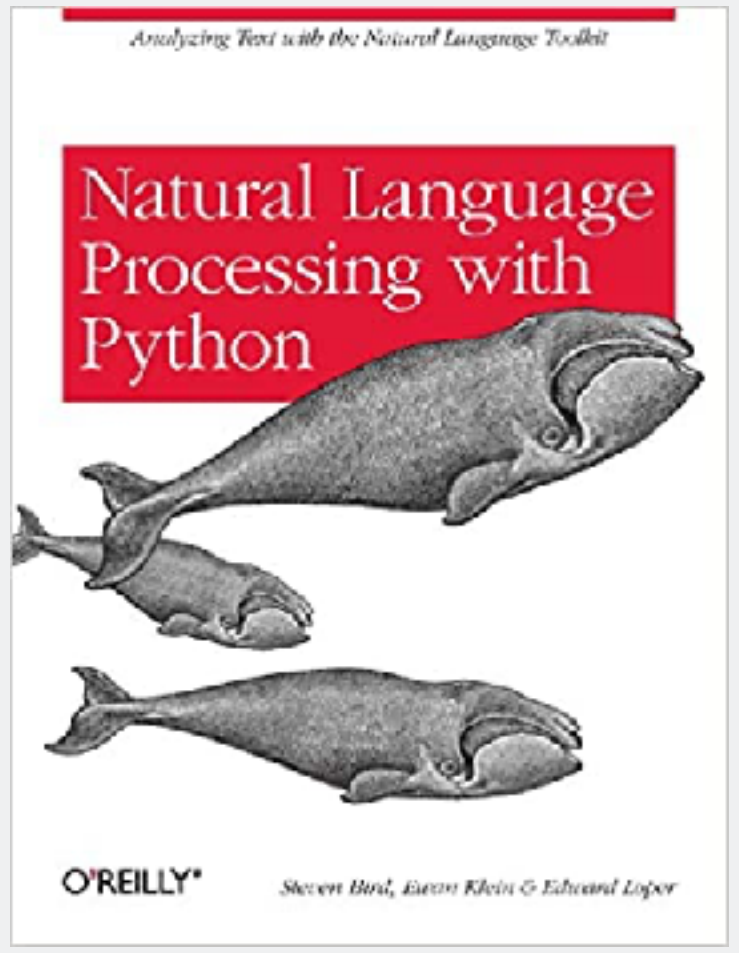


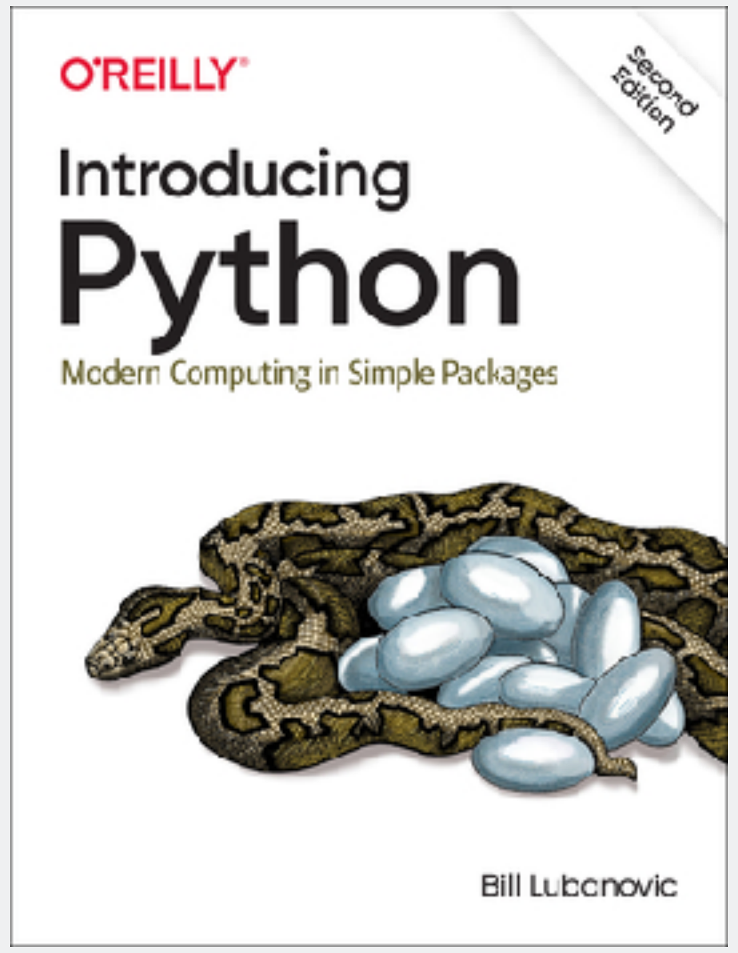
Dickinson et al. (2013)
Language and computers (LC)


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

Jurafsky & Martin. Speech and language processing (3rd ed; online draft)

Bird et al. NLTK with Python (Website)


Analyzing Text with the Natural Language Toolkit
Natural Language
Processing with
Python
O'REILLY
Steven Bird, Ewan Klein & Edward Loper



Lubanovic (2020). Introducing Python (2nd ed)

How to succeed in this course

- Practice!
  - Write codes. Try them out. They won’t break your computer!

- Read and study codes before class
  - Download materials and go through them

- Have a computer with you for in-class exercises
  - You can’t absorb programming by watching me program

Assignments & Grading

- Assignments are of two types:
  - homework assignments
  - programming exercises
- Assignments are given on **an approximately weekly basis**
  - assigned on Wednesday after class; due Saturday night
- Collaboration is encouraged
  - group work
  - individual submission

Assignments & Grading

- If you need help, don’t hesitate to:
  - talk to your peers
  - schedule a Zoom group meeting with me
- This is an unprecedented time.
  - The toll the pandemic is taking on your mental health
  - The uncertainty in your life & your future

Seek help. You’re not alone.

Assignments & Grading

Homework assignments (× 6)	50%
Programming exercises (× 5-6)	40%
Attendance*	5%
Participation	5%

A	85-100%	C	65-69.99
B+	80-84.99	D+	60-64.99
B	75-79.99	D	55-59.99
C+	70-74.99	F	0-54.99

* Don’t miss class except for family or medical emergencies. You’ll find it difficult to catch up.

Prerequisites

- A working laptop (iPad isn’t gonna work for this class....) running
  - Windows 10; or
  - Mac OS 10.15 (Catalina) or Mac OS 11 (Big Sur)

Introduction

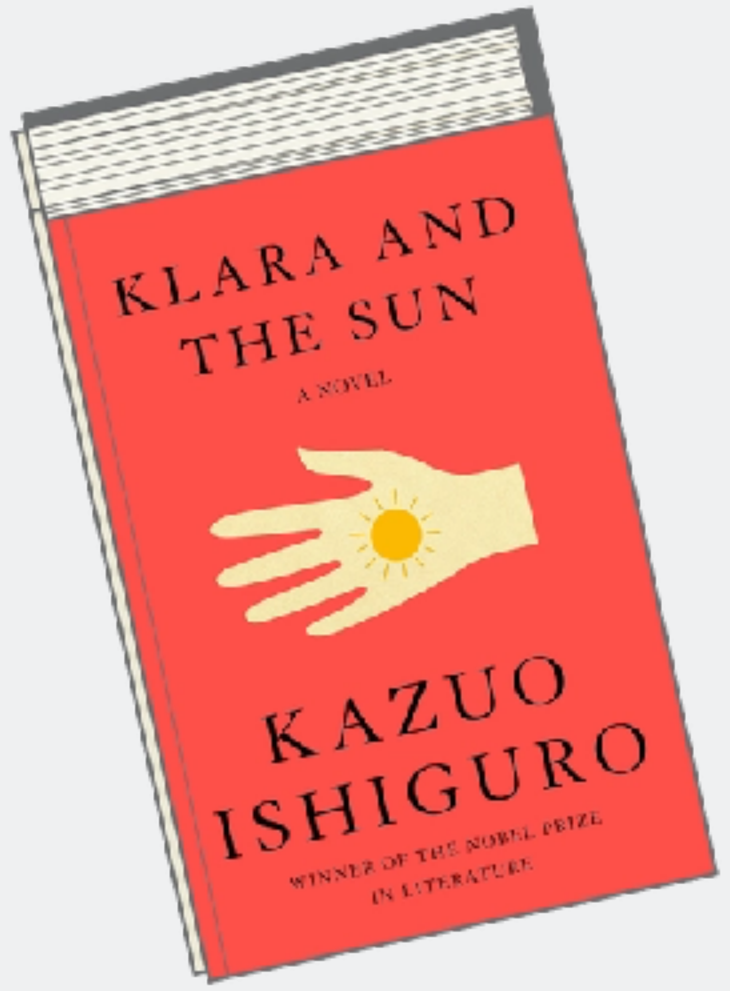
The stuff of imagination....

Goal: Giving computers the ability to process & understand natural language..

A.I. (2001) Stephen Spielberg

JOAQUIN PHOENIX AMY ADAMS ROONEY MARA
OLIVIA WILDE SCARLETT JOHANSSON
her
A SPIKE JONZE LOVE STORY
COMING SOON

Her (2013) Spike Jonze

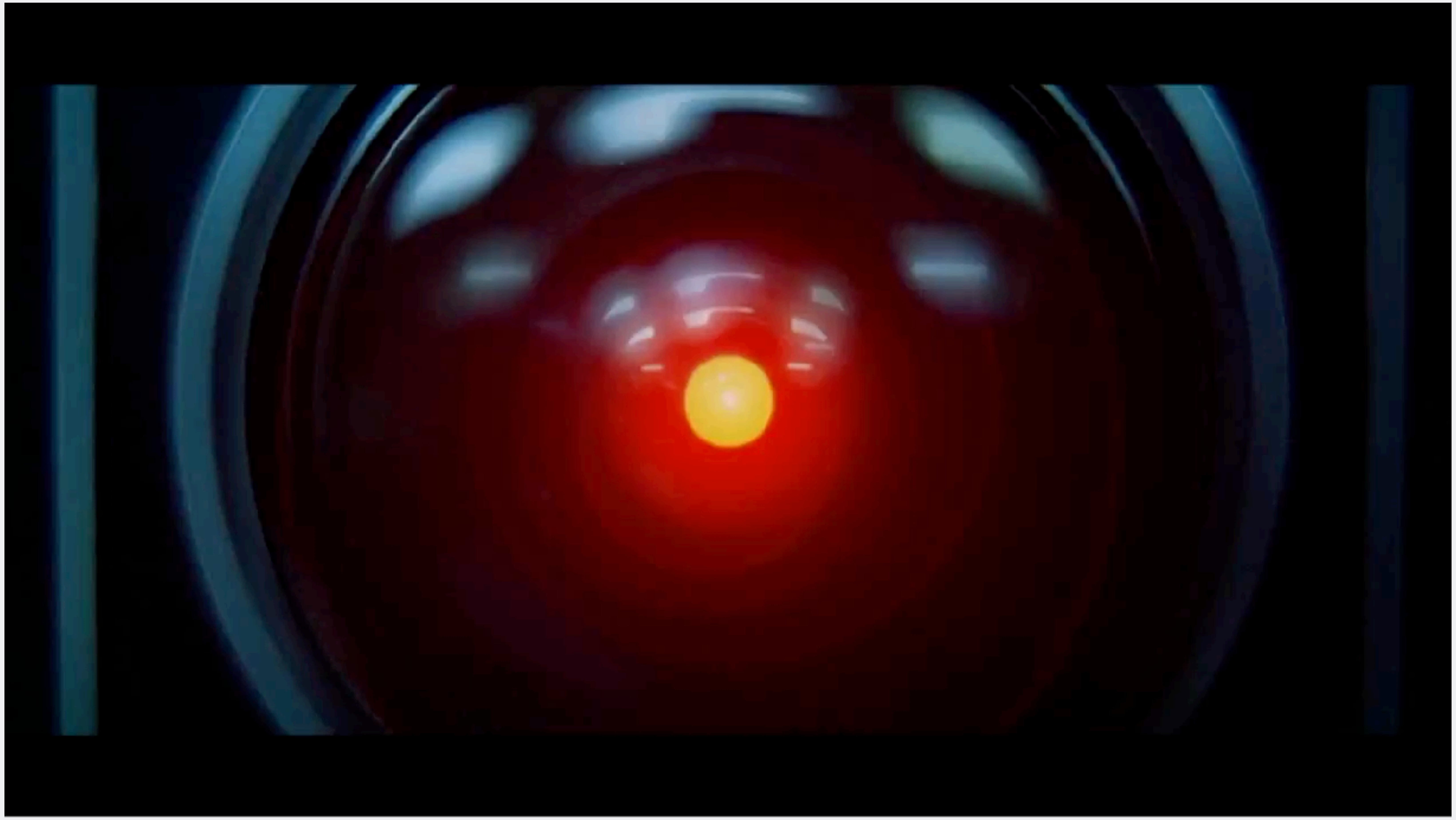


Klara & The Sun (2021)

...so that they can complete useful tasks with language

The stuff of imagination....

HAL 9000, a fictional AI [conversational agent], in *2001 A Space Odyssey*



Credit: Youtube

The stuff of imagination....

David: Open the pod bay door please, HAL.... Hello, HAL. Do you read me?

HAL: Affirmative, Dave. I read you.

David: Open the pod bay door, HAL.

HAL: I’m sorry Dave. I’m afraid I can’t do that.

David: What’s the problem?

HAL: I think you know what the problem is, just as well as I do.

David: What are you talking about?

HAL: The mission is too important for me to allow you to jeopardize it...

David: Where the hell did you get that idea, HAL?

HAL: Dave, although you took very thorough precautions against my hearing you, I could see your lips move...

The stuff of imagination....

David: Open the **pod bay door** please, HAL.... Hello, HAL. Do you read me?

HAL: Affirmative, Dave. I read you.

David: Open the **pod bay door**, HAL.

HAL: **I’m sorry** Dave. **I’m afraid** I can’t do that.

David: What’s **the problem**?

HAL: I think you know what the problem is, just as well as I do.

David: **What** are you **talking about**?

HAL: **The mission** is too important for me to allow you to jeopardize it...

David: Where the hell did you get **that idea**, HAL?

HAL: Dave, although you took very thorough precautions against my hearing you, I could see your lips move...

HAL has: **domain knowledge**, **discourse knowledge** (on top of linguistic knowledge)

The vision: Ultimate UI

User: *Hey Siri*, when did **Bill Gates** found **Microsoft**?

Siri: He and Paul Allen found the company on April 4, 1975.

User: *Who's* **the other guy**?

Siri: Paul Allen was a researcher, computer programmer, and philanthropist. He passed in 2018.

User: *What was* **his cause of death**?

Siri: He died of septic shock related to the cancer on October 15, 2018, at the age of 65.

Siri needs: **domain knowledge**, **discourse knowledge** and **world knowledge**
(not to mention linguistic knowledge!)

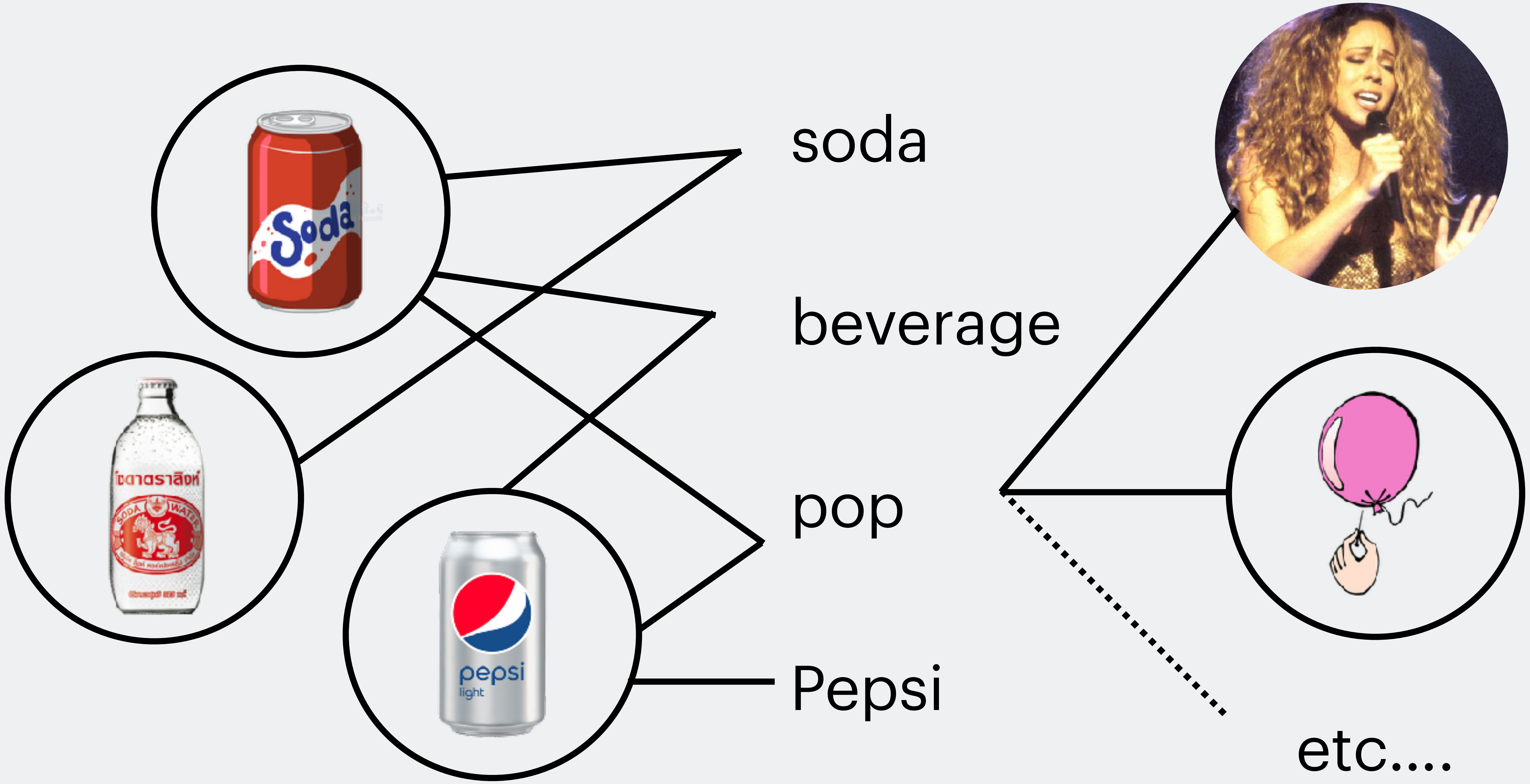
Adapted from: Bill MacCartney's Sym Sys 100 lecture at Stanford

Why is it so hard?

- Because language is:
  - highly **ambiguous** at all levels
  - complex, with recursive structures and coreference
  - subtle, exploiting context to convey meaning
  - and etc.

Adapted from: Bill MacCartney's Sym Sys 100 lecture at Stanford

Ambiguity: 1+ meanings



soda

beverage

pop

Pepsi

etc....

Adapted from: Bill MacCartney's Sym Sys 100 lecture at Stanford

Ambiguity: 1+ interpretations

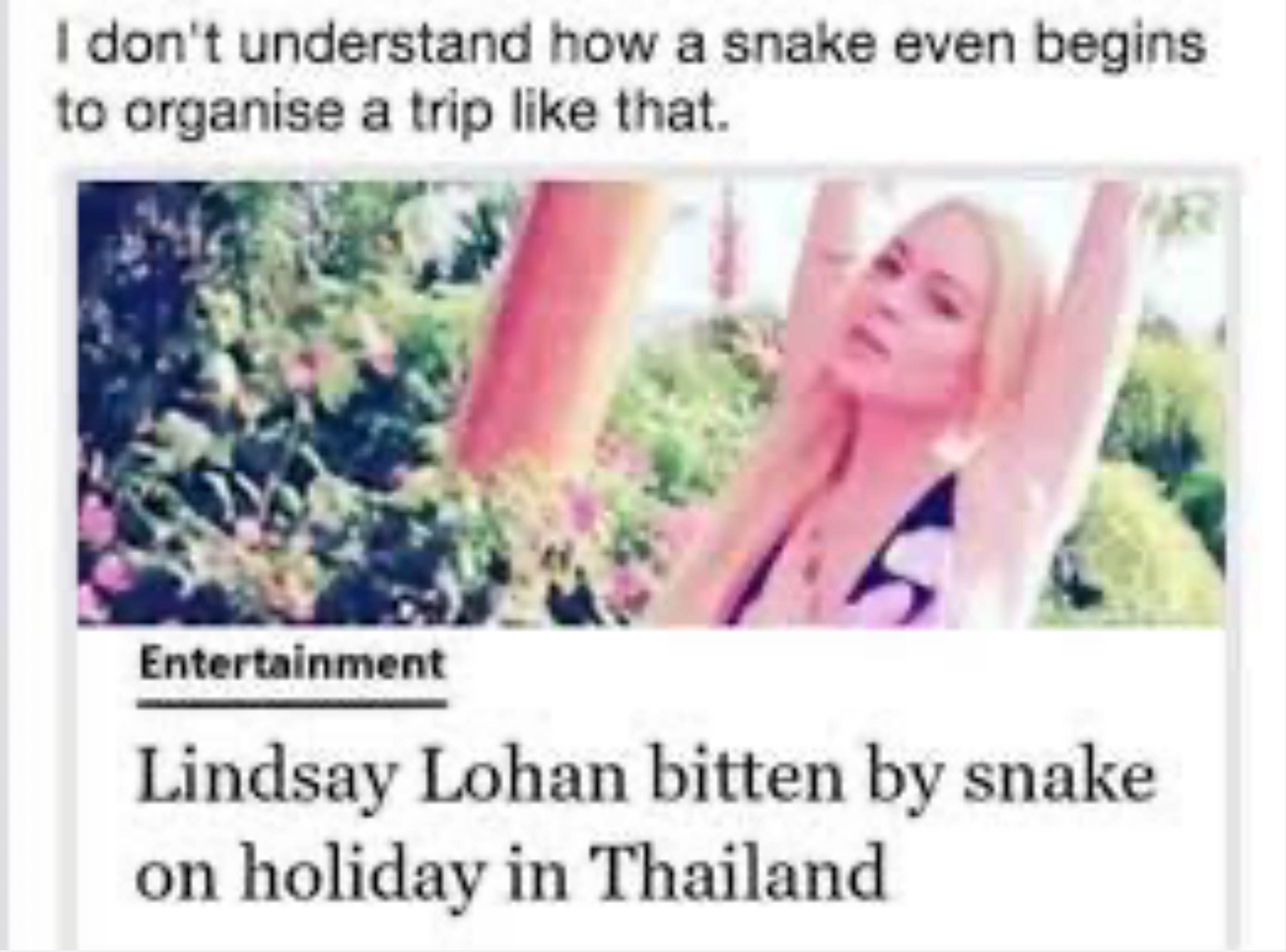
(Evening Standard) Lindsay Lohan bitten by snake on holiday in Thailand

(BBC) Queen mother tried to help abuse girl

(BBC) Knife crime: St John Ambulance to teach teens to help stab victims

(Guardian) Supreme Court plans an attack on independent judiciary, says Labour

(Epoch Times) Chinese citizen files new lawsuit against authorities seeking COVID-19 damages



I don't understand how a snake even begins to organise a trip like that.

Entertainment

Lindsay Lohan bitten by snake on holiday in Thailand

If you wanna waste some time: https://languagelog.ldc.upenn.edu/nll/?cat=118

Complexity and subtlety

(New York Times) **Britney Spears** Quietly Pushed for Years to End Her Conservatorship
Confidential court records… reveal that **the singer** has urged changes to the arrangement

He bumped against the sauce boat containing the raspberry syrup, spilling **it**

The city council refused the women a parade permit because **they feared** violence

The city council refused the women a parade permit because **they advocated** violence

Unless stated, examples are from Stanford Encyclopedia of Philosophy

The world needs linguists!

The roles of linguists:

- design and manage linguistic annotation projects
- provide consultation for quality enhancement (machine translation systems, search results, etc.)
- build NLP systems, train models
- and on and on....

Working knowledge of computational linguistics is definitely a plus!

Adapted from: Na-Rae Han's LING 1330/2330 lecture

The world needs linguists!

Computational linguistics:

…is the scientific and engineering discipline concerned with understanding language from a computational perspective, and building tools/programs that process and produce language.

Source: Stanford Encyclopedia of Philosophy

The world needs linguists!

- *Theoretical computational linguistics* focuses, for example, on:
  - formulating syntactic and semantic frameworks for analysis
  - developing computational models of language processing & learning

- *Applied computational linguistics* seeks to:
  - apply theories/frameworks to build practical applications: spell checkers, MT, dialogue systems/agents, automatic scoring (ETS), etc.
  - natural language processing (NLP); natural language understanding (NLU), language technology (LT)

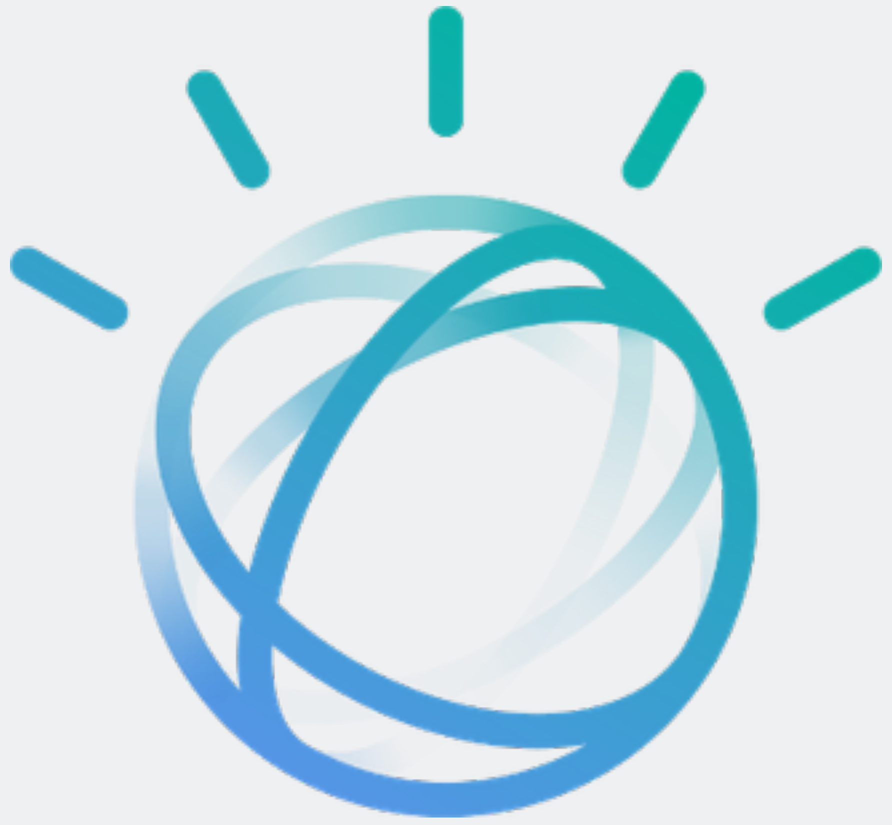
Source: Stanford Encyclopedia of Philosophy

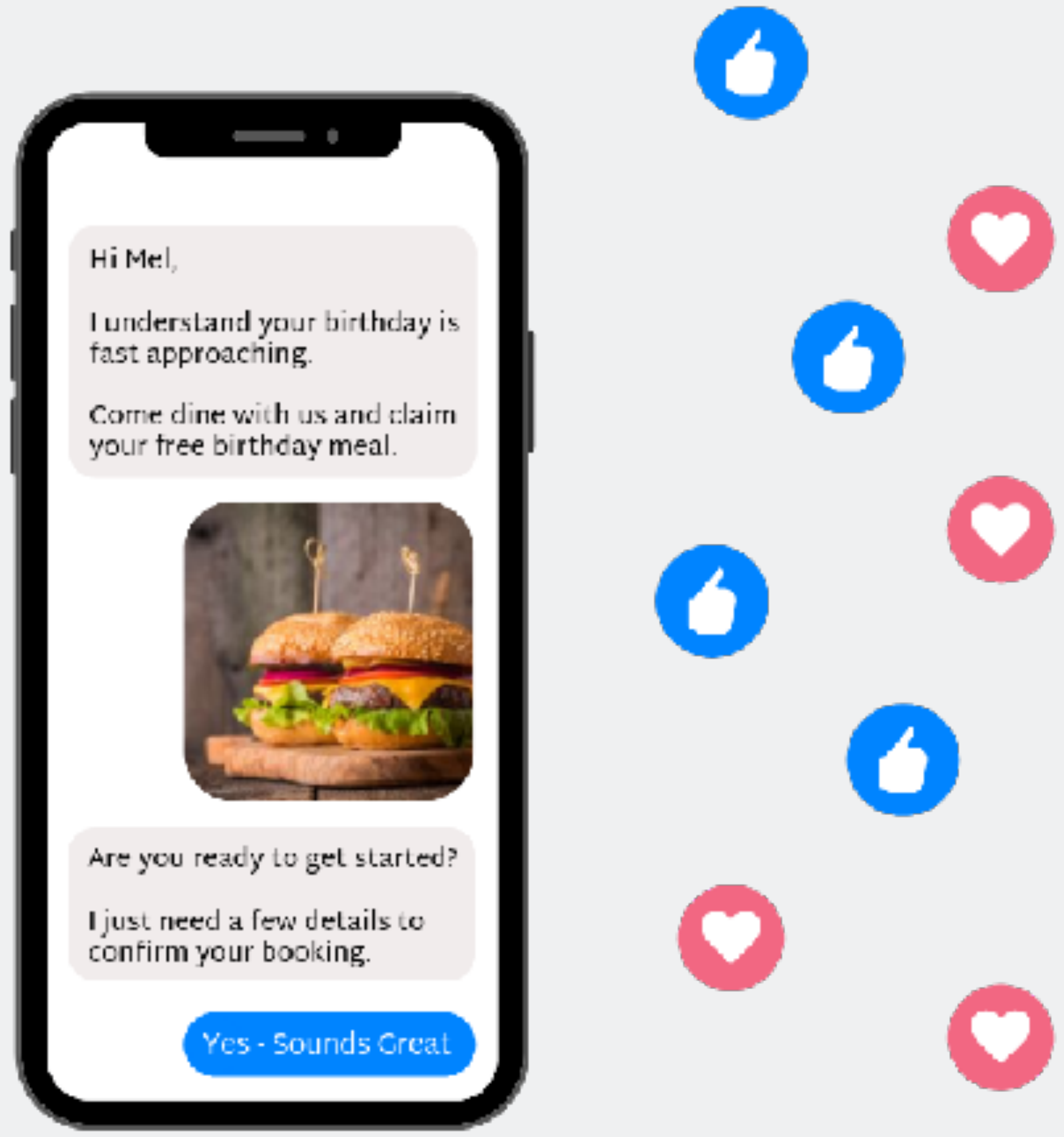
Our focus in this semester

- *Theoretical computational linguistics*:
  - some theories (finite-state automata, context-free grammar, etc)

- *Applied computational linguistics*:
  - text processing
  - NLP applications as discussed in *Language and Computers* (tokenization, morphological analyzer, POS tagging, etc.)

Language-related applications





Hi Mel,

I understand your birthday is fast approaching.

Come dine with us and claim your free birthday meal.

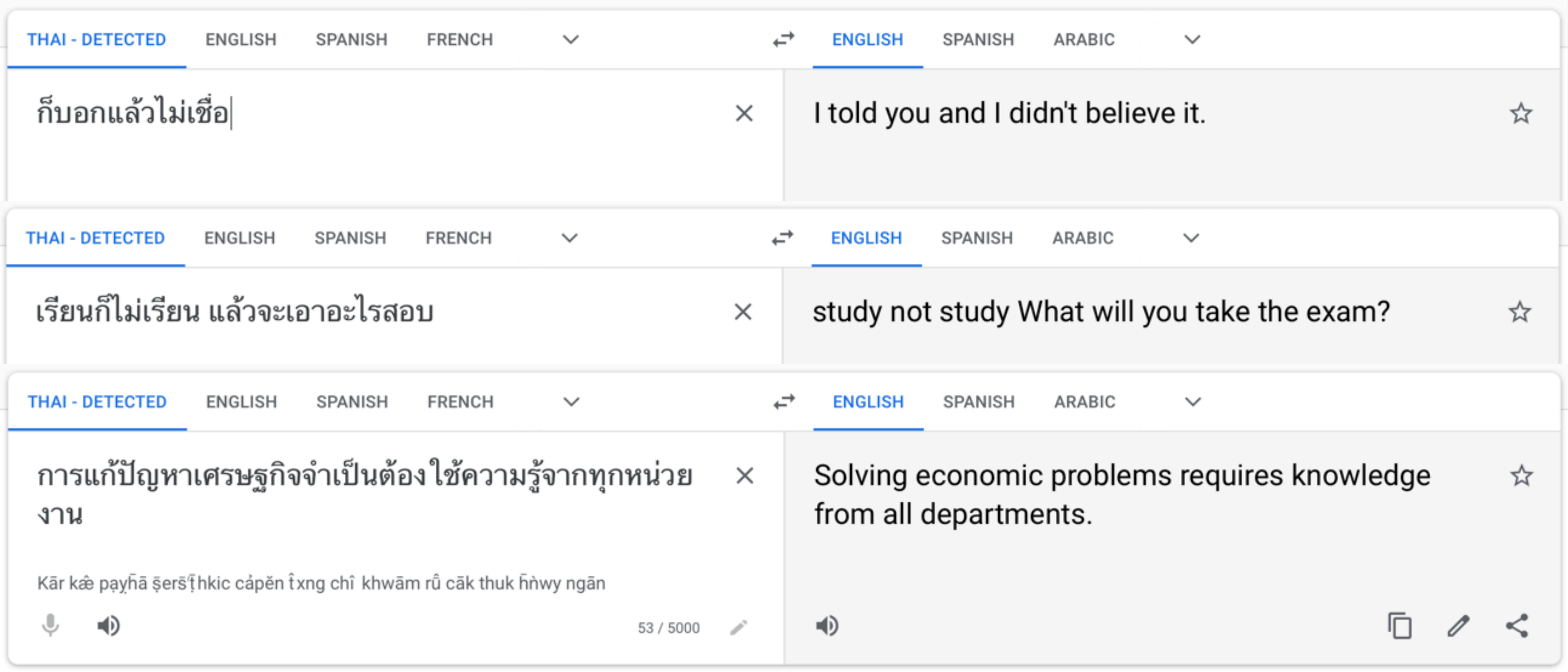
Are you ready to get started?

I just need a few details to confirm your booking.

Language-related applications

amazon alexa

Google Assistant



THAI - DETECTED ENGLISH SPANISH FRENCH ⇄ ENGLISH SPANISH ARABIC

ก็บอกแล้วไม่เชื่อ

I told you and I didn't believe it.

THAI - DETECTED ENGLISH SPANISH FRENCH ⇄ ENGLISH SPANISH ARABIC

เรียนก็ไม่เรียน แล้วจะเอาอะไรสอบ

study not study What will you take the exam?

THAI - DETECTED ENGLISH SPANISH FRENCH ⇄ ENGLISH SPANISH ARABIC

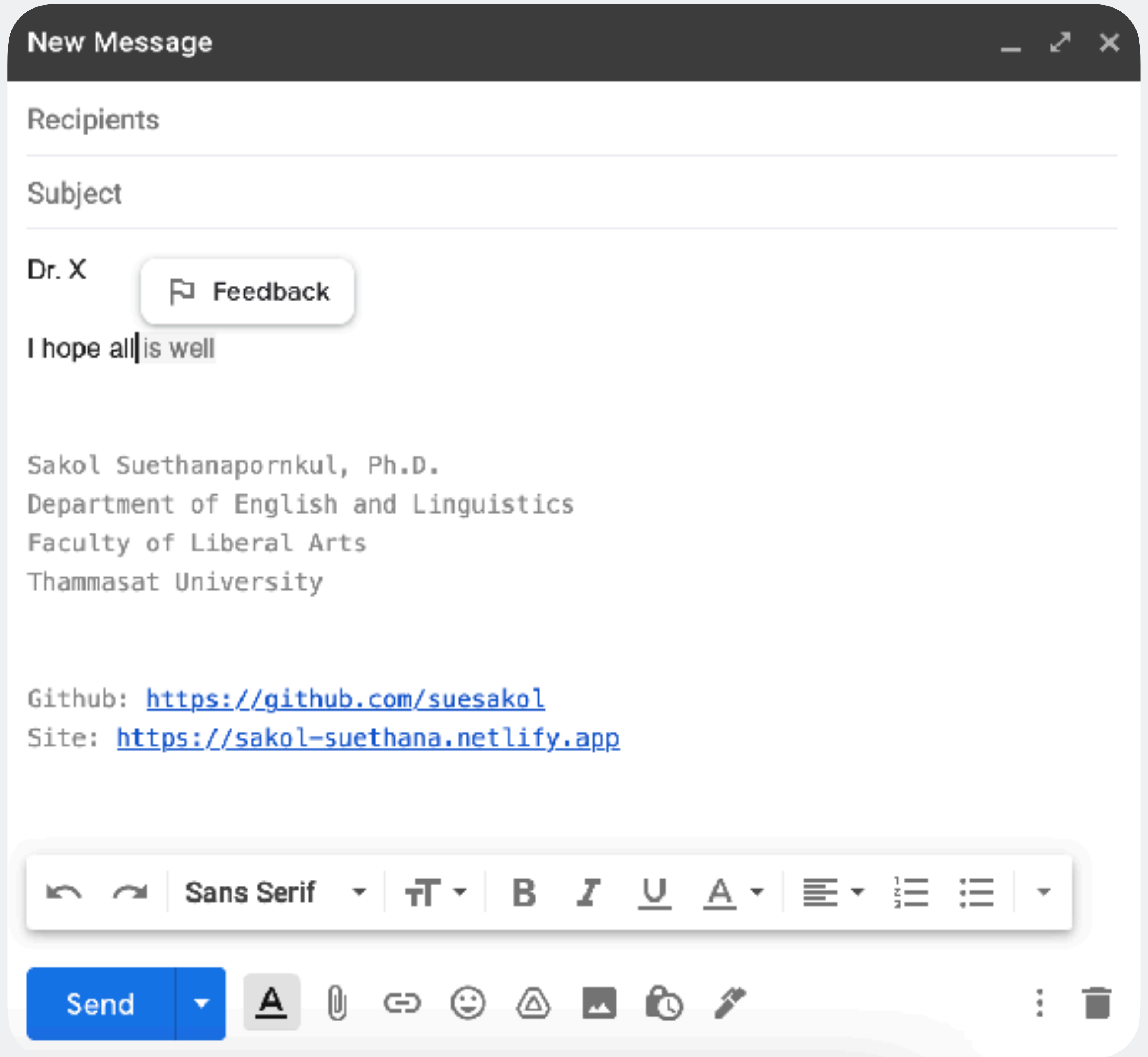
การแก้ปัญหาเศรษฐกิจจำเป็นต้องใช้ความรู้จากทุกหน่วยงาน

Kār kæ̂ pạỵh̄ā ṣ̄ers̄ʹṭ̄hkic cảpĕn t̂xng chı̂ khwām rū̂ cāk thuk h̄ǹwy ngān

53 / 5000

Solving economic problems requires knowledge from all departments.

Language-related applications



Recipients

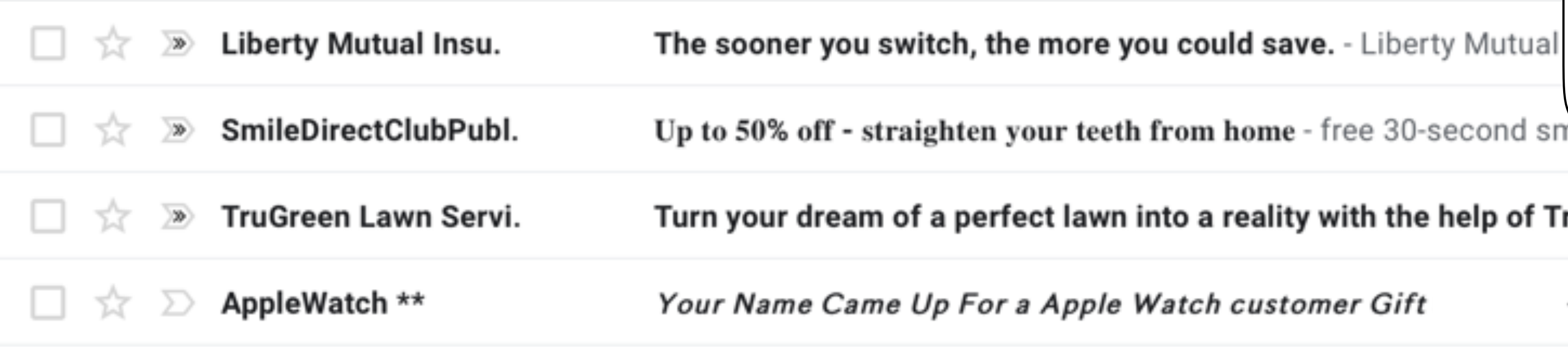
Subject

Dr. X

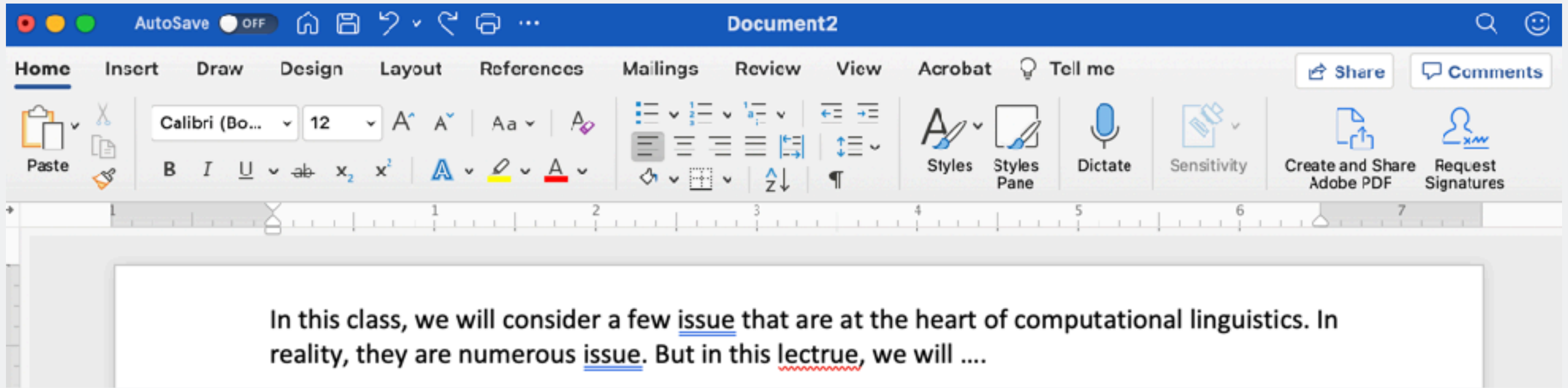
I hope all is well

Sakol Suethanapornkul, Ph.D.
Department of English and Linguistics
Faculty of Liberal Arts
Thammasat University

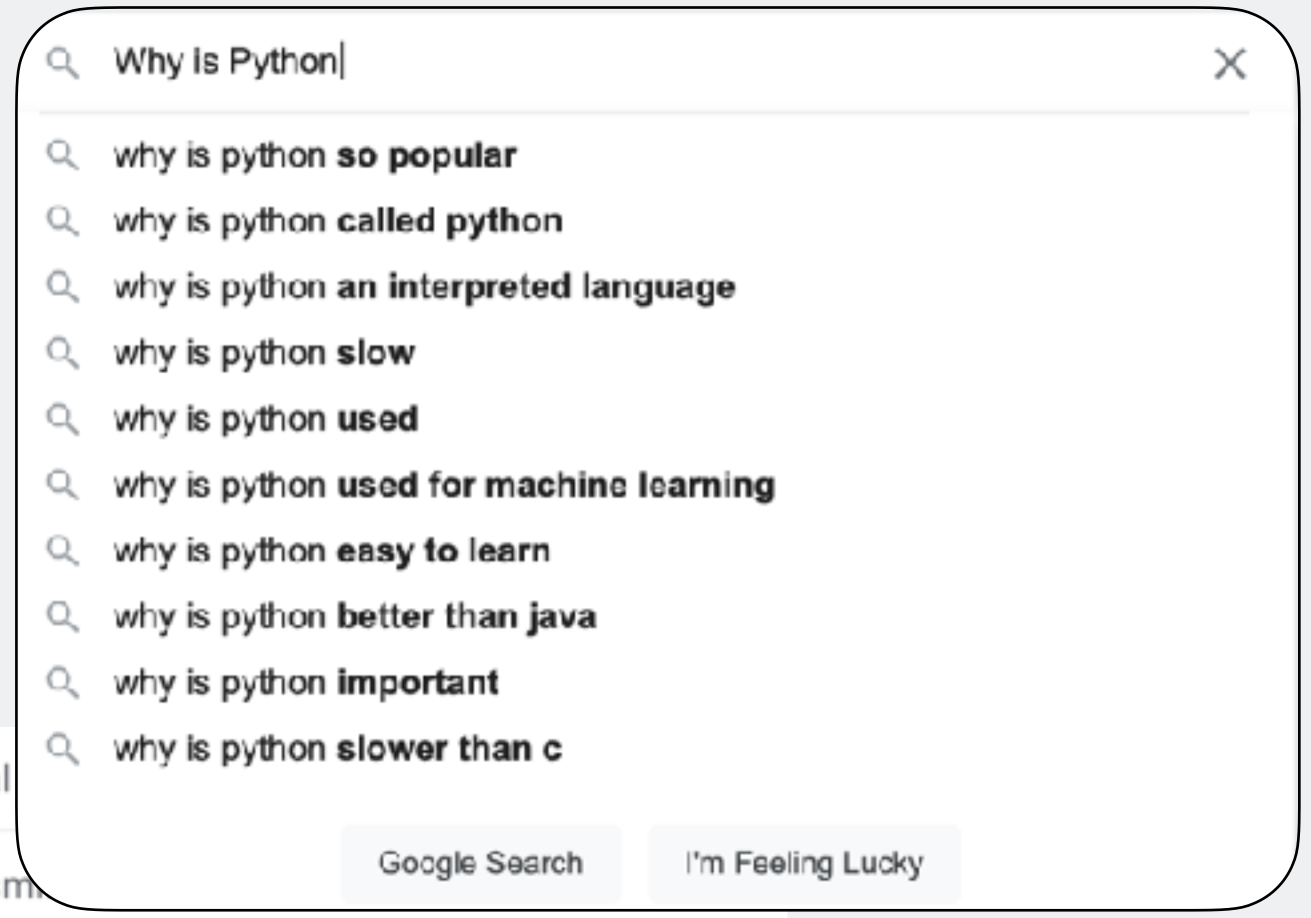
Github: https://github.com/suesakol
Site: https://sakol-suethana.netlify.app



Liberty Mutual Insu.	The sooner you switch, the more you could save.	- Liberty Mutual
SmileDirectClubPubl.	Up to 50% off - straighten your teeth from home	- free 30-second sm
TruGreen Lawn Servi.	Turn your dream of a perfect lawn into a reality with the help of TruGreen	- ALL EXPERT LAWN CARE 50%OFF◇
AppleWatch **	*Your Name Came Up For a Apple Watch customer Gift*	- ANSWER & WIN! A Brand New Apple Watch

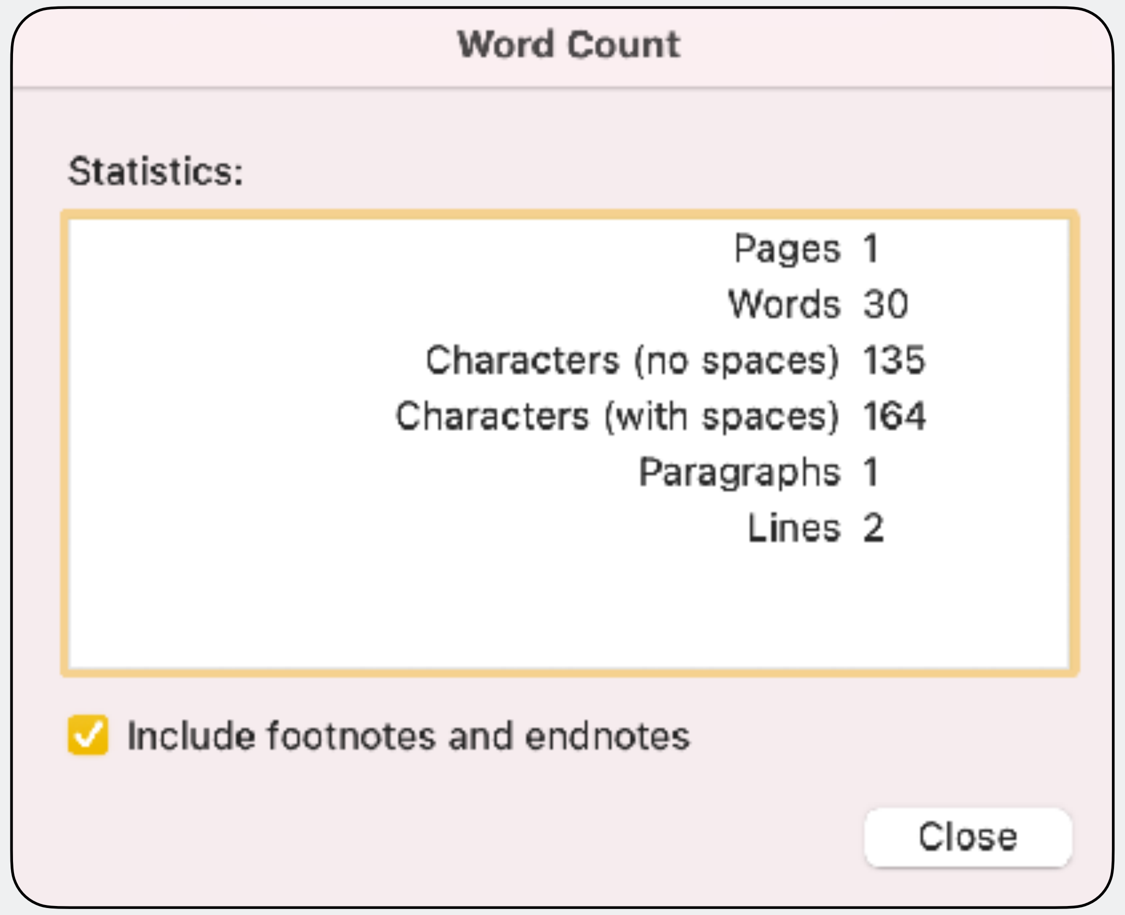


In this class, we will consider a few issue that are at the heart of computational linguistics. In reality, they are numerous issue. But in this lectrue, we will ....



Why is Python

- why is python **so popular**
- why is python **called python**
- why is python **an interpreted language**
- why is python **slow**
- why is python **used**
- why is python **used for machine learning**
- why is python **easy to learn**
- why is python **better than java**
- why is python **important**
- why is python **slower than c**



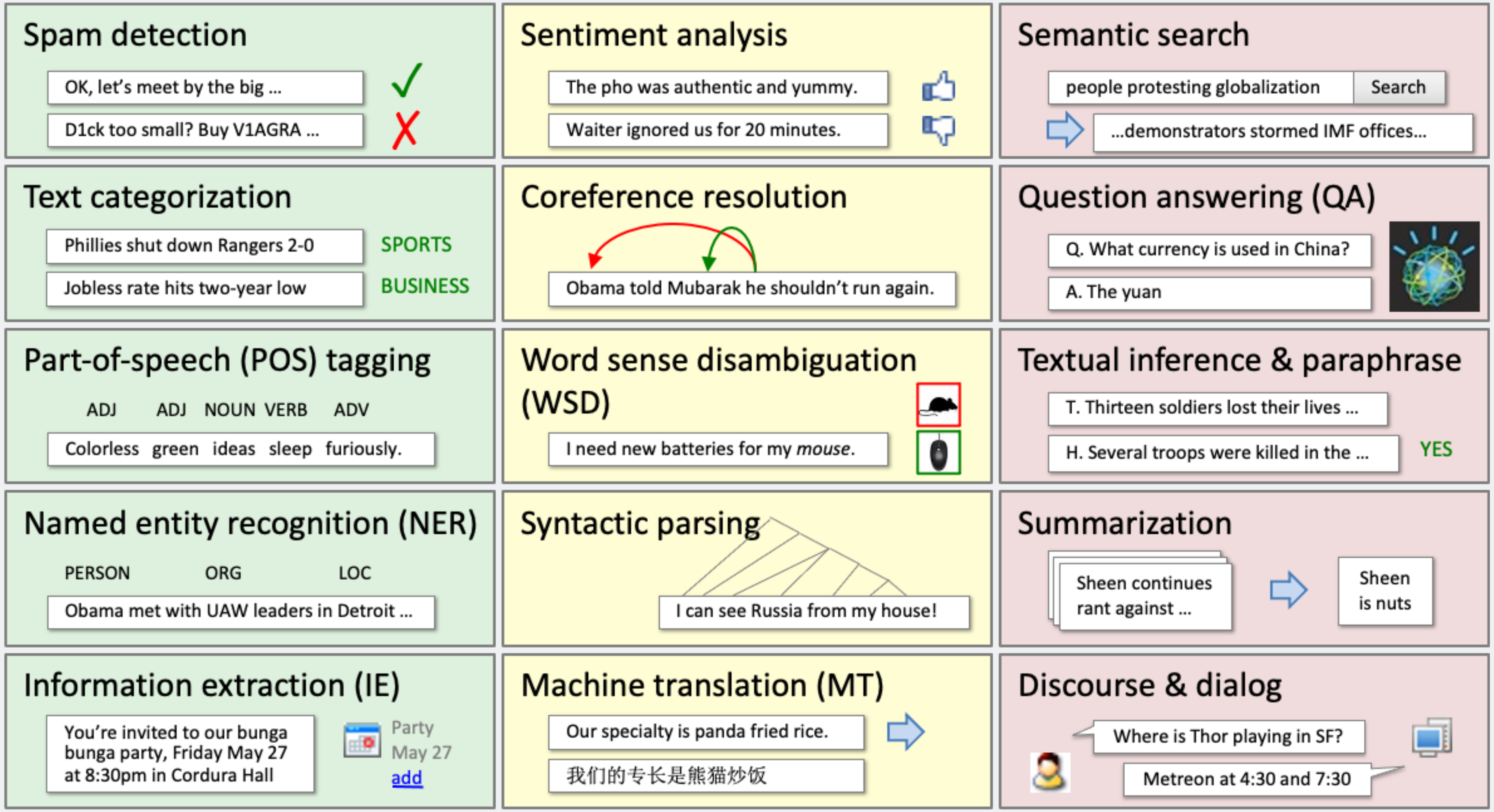
Word Count

Statistics:

Pages	1
Words	30
Characters (no spaces)	135
Characters (with spaces)	164
Paragraphs	1
Lines	2

☑ Include footnotes and endnotes

Language-related applications



mostly solved	making good progress	still really hard
Spam detection OK, let's meet by the big ... ✓ D1ck too small? Buy V1AGRA ... ✗	Sentiment analysis The pho was authentic and yummy. 👍 Waiter ignored us for 20 minutes. 👎	Semantic search people protesting globalization Search ⇨ ...demonstrators stormed IMF offices...
Text categorization Phillies shut down Rangers 2-0 SPORTS Jobless rate hits two-year low BUSINESS	Coreference resolution Obama told Mubarak he shouldn't run again.	Question answering (QA) Q. What currency is used in China? A. The yuan
Part-of-speech (POS) tagging ADJ ADJ NOUN VERB ADV Colorless green ideas sleep furiously.	Word sense disambiguation (WSD) I need new batteries for my *mouse*.	Textual inference & paraphrase T. Thirteen soldiers lost their lives ... H. Several troops were killed in the ... YES
Named entity recognition (NER) PERSON ORG LOC Obama met with UAW leaders in Detroit ...	Syntactic parsing I can see Russia from my house!	Summarization Sheen continues rant against ... ⇨ Sheen is nuts
Information extraction (IE) You're invited to our bunga bunga party, Friday May 27 at 8:30pm in Cordura Hall Party May 27 add	Machine translation (MT) Our specialty is panda fried rice. ⇨ 我们的专长是熊猫炒饭	Discourse & dialog Where is Thor playing in SF? Metreon at 4:30 and 7:30

Adapted from: Bill MacCartney's Sym Sys 100 lecture at Stanford

Our plan next week!

- Language and Computer, Chapter 1
  - Sections 1.1 and 1.3
- To make computers work with language, we need to start by:
  - Making it encode language, but how?

Homework announcement (HW1):

- Key terms in CL/NLP
- Up on the site; due Saturday night (→ Bonus assignment ☺️)