



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014



Introduction to  
the course

Why is natural  
language  
understanding  
so difficult?



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

## Problems, Problems, Problems

### Understanding Natural Language

Contemporary methods for automated natural language processing depend, to a large extent, on the use to which each application of the processing is to be put. Consider the following excerpt from Erle Stanley Gardner's "The Case of the Demure Defendant":



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

“Cross-examine,” Hamilton Burger snapped at Perry Mason.

Mason said, “Mr. Dayton, when you described your occupation you gave it as that of a police expert technician. Is that correct?”

“yes sir.”

“What is an expert technician?”

“Well, I have studied extensively on certain fields of science that are frequently called upon in the science of criminology.”

“That is what you meant by an expert technician?”

“Yes sir.”

“Now what is a police expert technician?”

“Well that means that ... well, it all means the same thing.”

“What means the same thing?”

“An expert technician.”

“An expert technician is the same as a police expert technician?”

“Well I am in the employ of the police department.”

“Oh the police employ you as an expert witness, do they?”

Instructor: Nick Cercone - 3050 CSEB - [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

*“Yes sir, ... I mean no, sir. I am an expert investigator, not an expert witness.”*

*“You are testifying now as an expert witness are you not?”*

*“Yes sir.”*

*“Then what did you mean by saying you were an expert technician but not an expert witness?”*

*“I am employed as a technician but not as a witness.”*

*“You draw a monthly salary?”*

*“Yes.”*

*And you are being paid for your time while you are on the stand as an expert witness?”*

*“Well, I'm paid for being a technician”*

*“Then you won't accept any pay for being a witness?”*

*“I can't divide my salary.”*

*“So you are being paid?”*

*“Of course - as part of my employment.”*



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

*“And are you now employed by the police?”*

*“Yes.”*

*“And are you an expert witness?”*

*“Yes.”*

*“Then you are now being employed as an expert witness.”*

*“I guess so. Have it your own way.”*

*“When you described yourself as a police expert technician that means your testimony is always called by the police. Isn't that so?”*

*“No, sir.”*

*“Who else calls you?”*

*“Well, I ... I could be called by either party.”*

*“How many times have you been on the witness stand?”*

*“Oh, I don't know. I couldn't begin to tell you.”*

*“Dozens of times?”*

*“Yes.”*

*“Hundreds of times?”*



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

*“Probably.”*

*“Have you ever been called by the defense as a defense witness?”*

*“I have not been directly subpoenaed by the defense. No, sir.”*

*“So that you have always testified for the police, for the prosecution?”*

*“Yes, sir. That's my business.”*

*“That was what I was trying to bring out,” Mason said. ...*

*“Yes.”*

*“Then you are now being employed as an expert witness.”*

*“I guess so. Have it your own way.”*

*“When you described yourself as a police expert technician that means your testimony is always called by the police. Isn't that so?”*

*“No, sir.”*

*“Who else calls you?”*

*“Well, I ... I could be called by either party.”*



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

*“How many times have you been on the witness stand?”*

*“Oh, I don't know. I couldn't begin to tell you.”*

*“Dozens of times?”*

*“Yes.”*

*“Hundreds of times?”*

*“Probably.”*

*“Have you ever been called by the defense as a defense witness?”*

*“I have not been directly subpoenaed by the defense. No, sir.”*

*“So that you have always testified for the police, for the prosecution?”*

*“Yes, sir. That's my business.”*

*“That was what I was trying to bring out,” Mason said. ...*



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

Mr. Dayton needs to understand the subtleties of noun phrases such as “*police expert technician*”, to answer Mr. Mason’s questions. Understanding such phrases are troublesome to automate since “*police*”, “*expert*” and “*technician*” are all nouns.

Generalizing semantic considerations for such constructions have proven evasive. The compositional approach to NLU favoured by logic grammarians becomes combinatorially explosive. Many researchers represent noun-noun constructions as single lexical entries, constraining the computation required to disambiguate them.





CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

When the domain of discourse is well specified and the number of such phrases is small, this approach works adequately. But is it practical? Consider “western region outage log” employed by telecommunications service personnel. Would the designer of their system resort to separate lexical entries for “eastern region outage log”, “southern region outage log”, “northern region outage log”, “northeastern region outage log”, ..., “western district outage log”, ..., “western prefecture outage log”, ..., “western region service log”, ..., “western region record”, ...?



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

Imagine further, the processing required by Perry Mason. Not only must the subtleties of language understanding realized by Mr. Dayton be mastered but also the reasoning capabilities of Mr. Mason and extraction of relevant and salient conversational features be identified in order to generate the appropriate next question. Actually, Mr. Mason's task is much simpler than Mr. Dayton's - to generate an utterance which conveys a presumably preexisting thought. Mr. Dayton's task as listener is to decide what Mr. Mason must have been thinking in order to motivate his utterance in the particular context in which he uttered it.



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

## Some NLP Applications

- machine translation
- speech analysis and generation systems
- spell checking and grammatical correction
- conversational agents
- document generation
- text classification, summarization, mining
- information retrieval and information extraction
- question answering
- support applications, such as: stemming, POS tagging, semantic tagging, and partial parsing



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

## Short History of NLP

1947–54 pioneers and foundational insights

1954–66 decade of optimism, two camps: symbolic and stochastic

1966 ALPAC report in US (negative MT report)

1970-90's emergence various systems & approaches:

- stochastic paradigm
- logic-based
- NLU
- discourse modeling

1990–2000 stochastic NLP, Web, unification-based grammars

2000– "The rise of Machine Learning"



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

## Levels of NLP

1. phonetics: physical sounds
2. phonology: sound system (phonemes) of a spoken language
3. morphology: word structure
4. syntax: inter-word structure up to sentence structure
5. semantics: meaning up to the sentence level
6. pragmatics: “speaker’s meaning” —extended from the literal sentence meaning
7. discourse: units larger than an utterance (e.g., inter-sentence meaning, references)



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

## Why is NLP hard?

Aside: Turing test, evidence from neural sciences

1. highly ambiguous
2. vague (the principle of minimal effort)
3. universal (domain independent)



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

## Ambiguities at all levels of NLP

- phonological (phonetical) ambiguity
- lexical ambiguity
- syntactic ambiguity
- semantic ambiguity
- referential ambiguity
- pragmatic ambiguity



## Ambiguity Examples at Various Levels of NLP

- lexical ambiguity, e.g. “hot” may mean: warm, spicy, good looking, or stolen
- syntactic ambiguity, e.g. “Time flies like an arrow.”

Two structures and two meanings.

- semantic ambiguity

Examples:

1. What does “coast road” mean? Is it a road that leads to a coast, or a road that follows the coast?
2. “carriage return” — Is it a return of a carriage, or an ASCII character?
3. “kick the bucket”, and other idioms





CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

– referential ambiguity— a kind of semantic ambiguity

Example: ‘It,’ or ‘he’ in a text – to what and who does it refer?

– pragmatic ambiguity

12am— is it noon or midnight?

Do you have a quarter? — Do you have or quarter, or can you give me a quarter?

Can you pass the salt? — Can you, or would you pass the salt?



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

## Some More Problems with Natural Language

English, for example, is indeed a *crazy language*.

- A hot cup of coffee → A cup of hot coffee
- A near miss → in reality, a near miss is a collision
- I could care less → I could not care less
- Put on your shoes and socks → this is exceedingly difficult since we, in reality, put on our socks and shoes.



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

English is also *weird*, wearing tonal grooves in our heads, ignoring the rigid rules for forming proper expressions. Consider the oxymorons:

- Fresh frozen jumbo shrimp
- Old news
- Freezer burn
- Voice mail
- Final draft
- Act naturally
- Conspicuously absent



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

These oxymorons give rise to material for comedians and new web sites on the Internet. Nonetheless we readily interpret them without extending our complexity of language analysis. In fact we take liberties with language all of the time by editorializing oxymorons, for example:

- Military intelligence
- Peace offensive
- Postal service
- Airline food
- Microsoft works
- Good grief – we find oxymorons lurking everywhere.



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

It is common in the best of the Yiddish tradition to answer a question with a question, often the very same question with different emphasis and intonation.

For example, in response to the question *Did you buy flowers for your mother on her birthday?*, the response would be quite different if a different word were emphasized in the answer. Thus the answer *Did I buy flowers for your mother on her birthday?* is quite different from *Did I buy flowers for your mother on her birthday?* which is different from *Did I buy flowers for your mother on her birthday?* and so on and so on.



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

Imagine the reasoning required to interpret the following and correctly ask the question that Mr. Rosenberg asks of Murray Goldwag at the end of this story.

One Friday afternoon, Murray Goldwag leaves work in Brooklyn to catch the bus to Poughkeepsie upstate to spend the weekend with his finance Lennie Rosenfeld. During the trip an elderly gentleman, Mr. Rosenberg, returning to his home in Poughkeepsie from a visit with his brother in Brooklyn, strikes up a conversation with Murray. “So, you’re up to the Pok’ for the weekend”? “Yes sir”, replied Murray. “By yourself?” “Just visiting friends”, replied Murray. Mr. Rosenberg then reflects for a while. He is visiting friends,



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

leaving early enough to arrive in time for a late dinner and to get himself a motel room. He must be visiting a girl friend since no one would go out of their way so much for another boy friend. Now who could he be visiting? He is a handsome lad. Could it be the Goldberg twins? - No, Meryl is away at University out of state and Fran is out of town visiting relatives. Could it be Maxine Kriebel? - No, she has a boyfriend. What about Melinda Eaman? Probably not, she is very rich and would not be going out with any young man who had to travel by bus in order to see her. Well, what about Sarah Lavine? That could be the one, Sarah has been on cloud nine recently



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

and acting mysterious, visiting all the shops and making preparations for a big day soon. She is not graduating from College, nor celebrating a promotion or new job. That's it, thought Mr. Rosenberg, Murray must be coming to Poughkeepsie to see Sarah and make plans for their upcoming wedding. Mr. Rosenberg turns to Murray and says "Congratulations, son, on your upcoming wedding to Sarah Lavine." "wha--what", said Murray. "How did you know, we haven't even told her parents yet." "My boy", said Mr. Rosenberg, "its obvious."





## An Application: NL-DB Interface

Consider the sentence:

*I want the marks in CS110.* [1]

- Its interpretation would seem straightforward.
- The sentence structure is simple – any parser
- It is not difficult to map “marks” onto the corresponding database attribute, and to require that only marks in the course CS110 be accessed.
- However, many assumptions underly this interpretation which may, in fact, not correspond to what the user intended.



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

- The first problem which must be resolved is how to even recognize [1] as a question.
- Its surface form is declarative, not interrogative.
- Must recognize underlying *request* speech act
- Imagine “I want to change all marks of 49% to 50%.”, “I want to add something to what I just told you.”, etc.
- Other phraseologies are more problematic
- “Can you give me the marks in CS110?” - a yes/no answer is inappropriate



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

- Recognizing underlying speech act(s) is just one item which must be inferred about the user.
- What does the user want with the marks - needed to determine what exactly to present to the user.
- If the user is the registrar, who wants to mail the marks out to the students, it would be a good idea to include addresses with names.
- If the user is the university's statistician, it may be sufficient to give the distribution of the marks.
- Certain users may not be even be allowed access to the marks (for example, another CS 110 student).



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

- Other unspecified aspects of [1] must be clarified if an appropriate response is to be given.
- Does the user want the marks for all sections of CS 110 or just one?
- If just one section, is it possible to infer which section without asking the user
- Does the user want the marks for last year's CS 110 course? Next year's? The course currently underway?
- If the course currently in progress, should the system hold the request in abeyance and respond at the end of the term when the course is done?
- How about holding the request if the marks are for next year's course?



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

- How about holding the request if the marks are for next year's course?
- Once the course and the section are determined, the system must determine whether all marks are required, or just final exam marks, or the final grade for each student.
- The system must also decide what to do if the request is inappropriate. For example, CS 110 may not have been offered, or there may not be any course named CS 110, or the marks may not have been assigned as yet.

*To be continued next class*



CSE6339 3.0 Introduction to Computational Linguistics  
Instructor: Nick Cercone – 3050 LAS – [nick@cse.yorku.ca](mailto:nick@cse.yorku.ca)  
Mondays, Wednesdays 10:00-11:20 – North Ross 836A  
Winter Semester, 2014

## Concluding Remarks

### The Road to Wisdom

*The road to wisdom? –  
Well, it's plain and simple to express:  
Err and err and err again  
but less and less and less.*