

# COGS 200: Introduction to Cognitive Systems

Bob Woodham

Department of Computer Science  
University of British Columbia

Lecture Notes 2014/2015 Term 1

14W1: September–December, 2014

# Readings Summary

Today's assigned "reading" was to watch Peter Norvig's September 23, 2010, talk, "The unreasonable effectiveness of data," UBC Computer Science Distinguished Lecture Series. ([video](#))

- Peter Norvig introduces us to the statistical machine learning approach to artificial intelligence
- Example problems discussed in the talk:
  - ▶ scene completion (in image processing)
  - ▶ scene carving (in image processing)
  - ▶ speech recognition
  - ▶ orbital mechanics
  - ▶ flu trends
  - ▶ word sense disambiguation
  - ▶ word segmentation
  - ▶ spelling correction
  - ▶ **language translation**

# Today's Learning Goals

- ➊ To explore additional examples of challenges in language translation, for both human and machine
- ➋ To understand that computer science is about language design
  - ▶ programming languages are precise about both syntax and semantics
  - ▶ syntax determines whether a sequence of statements is valid in a language
  - ▶ given a valid sequence of statements in a language, semantics determines what the statements do (i.e., determine what is computed)
  - ▶ implementation of a programming language typically involves multiple language translations
- ➌ To explore one approach to machine translation of natural language that resolves semantic ambiguity by appeal to large corpora of empirical data

## Example 1: Back Translation

The following is a classic, but apocryphal, translation from English into Russian and back into English:

The spirit is willing but the flesh is weak

## Example 1: Back Translation

The following is a classic, but apocryphal, translation from English into Russian and back into English:

The spirit is willing but the flesh is weak

came back as

The vodka is fine but the meat is rotten

## Example 2: Bill Gates and Steve Jobs

Scene: Bill Gates and Steve Jobs are having lunch together. The conversation is a bit strained

Bill Gates (glancing at his watch): “I have to leave now. I have an appointment at the bank to talk about a loan”

Steve Jobs (incredulous): “Why do you need a loan?”

Bill Gates (smiling): “I don’t. The bank does”

## Example 3: Spoken versus Written English

## Example 3: Spoken versus Written English

Fill in the blank:

There are three \_\_\_\_\_ in the English language: to, too and two



## Example 4: Crown and Anchor Sign



## Example 4: Crown and Anchor Sign



I like the sign. But, please leave more space between crown and and and and and anchor

## Example 5: Simple Arithmetic

Calculate

## Example 5: Simple Arithmetic

Calculate

$$3 + 5 \times 2 =$$

## Example 5: Simple Arithmetic

Calculate

$$3 + 5 \times 2 = 13$$

## Example 5: Simple Arithmetic

Calculate

$$3 + 5 \times 2 = 13$$

This result depends, in part, on the shared convention that multiplication has precedence over addition

That is, convention dictates that we interpret the left hand side of the equation as  $3 + (5 \times 2)$ , not as  $(3 + 5) \times 2$

## Example 5: Simple Arithmetic

Calculate

$$3 + 5 \times 2 = 13$$

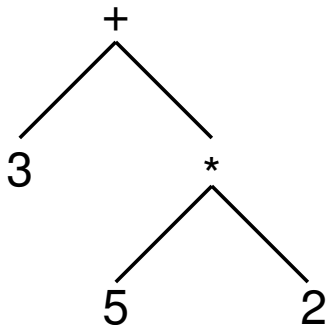
This result depends, in part, on the shared convention that multiplication has precedence over addition

That is, convention dictates that we interpret the left hand side of the equation as  $3 + (5 \times 2)$ , not as  $(3 + 5) \times 2$

For those contemplating or taking CPSC 110, in Scheme the expression would be written as `(+ 3 (* 5 2))`

## Example 5 (cont'd): “Abstract” Syntax Tree

The intended “meaning” of the arithmetic expression can be made explicit as a tree



regardless of the syntax (and associated conventions) used to represent the original expression



## Example 6: GCC – the GNU Compiler Collection

Source Language “Front Ends”	Abstract Syntax Tree (AST)	Target Machine “Back Ends”
C C++ Objective C Objective C++ Fortran Java Ada Go ⋮	common internal representation	45 processor families (in standard release)  25 additional processors (from 3rd parties)  ⋮

[http://en.wikipedia.org/wiki/GNU\\_Compiler\\_Collection](http://en.wikipedia.org/wiki/GNU_Compiler_Collection)

## Example 7: Translation Fail – An Arm Gesture

... the interpreter translated their [Liberals Bill Graham and Marlene Jennings] description for the gesture – “bras d’honneur” – into “Italian salute”

“I want an apology on behalf of all Canadians of Italian heritage and I want it withdrawn,” said Tory House Leader Bob Nicholson. “That is an insult”

Graham and Jennings demanded an apology for the mistaken suggestion they had made a racist slur

No apology was forthcoming

“I have a right to listen to it in translation. . .” Nicholson retorted

Speaker of the House Peter Milliken could do only one thing – throw up his arms and move onto to other business

House of Commons, June 14, 2006

## Example 8: (A More Serious) Translation Fail



[http://en.wikipedia.org/wiki/We\\_will\\_bury\\_you](http://en.wikipedia.org/wiki/We_will_bury_you)

## Example 8 (cont'd):

Original Russian:

**Мы вас похороним**

(transliterated as **Mi vas pokhoronim**)

Alternate English Translations:

We will be present when you are buried

We will be present at your funeral

We will outlast you

We will outlive you

## Example 9:

Time flies like an arrow

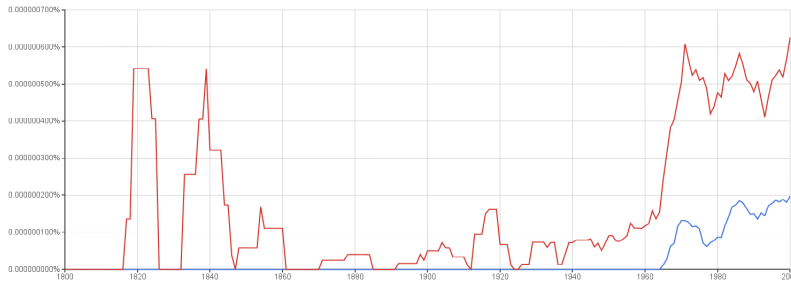
## Example 9:

Time flies like an arrow

But,

Fruit flies like a banana

## Example 9: Google Books Ngram



time flies like an arrow  
fruit flies like a banana

## Example 10: "THE CAT"

THE CAT



## Example 10 (cont'd): “THE CAT”

An **on-line dictionary of American English** (file /usr/share/dict/american in OpenSuSE 12.3) contains a total of 305,089 “distinct” entries

Sequence	Count	Proportion
ca	16,152	$5.29 \times 10^{-2}$
ch	15,311	$5.02 \times 10^{-2}$
at	30,039	$9.85 \times 10^{-2}$
ht	1,992	$6.53 \times 10^{-3}$
ta	15,389	$5.04 \times 10^{-2}$
th	11,342	$3.72 \times 10^{-2}$
ae	2,394	$7.85 \times 10^{-3}$
he	14,890	$4.88 \times 10^{-2}$
cat	2,724	$8.91 \times 10^{-3}$
cht	238	$7.80 \times 10^{-4}$
tae	79	$2.59 \times 10^{-4}$
the	3,424	$1.12 \times 10^{-2}$

## Example 10 (cont'd): “THE CAT”

Here are the 79 “**tae**” words. . .

Actaeon, antae, Antaeon, Antaeus, aortae, arborvitae, arborvitae's, Archegoniatae, Archegoniatae's, Aristaeus, arytaenoid's, aspiratae, ballistae, chaetae, Clytaemnestra, Clytaemnestra's, Compositae's, Conjugatae, Conjugatae's, Crataegus, Crataegus's, cristae, crustae, emeritae, etaerio, etaerios, glutaeal, hetaera, hetaerae, hetaerai, hetaeras, hetaerismic, hetaerisms, hetaerist, hetaerists, Heterocontae, Heterocontae's, Isokontae, Isokontae's, Labiatae, Labiatae's, locustae, lyttae, metaethical, metaethics, Nabataea, Nabataean, Nabataea's, notaeum, notaeums, placentae, Plataea, Plataean, Plataeans, Plataea's, Ratitae, Ratitae's, setae, spirochaetaemia, Stael, Stael's, taedium, Taegu, Taegu's, Taejon, tael, tael's, taeniate, Therapeutae, Therapeutae's, Tsvetaeva, Tsvetaeva's, Tyrtaean, Tyrtaean's, vistaed, vitae, vittae, voltaelectric, voltaelectricity

## Example 10 (cont'd): “THE CAT”

A decade (2000–2009) of **Bob's (spam filtered) email** contained a total of 51,106,155 “words”

Sequence	Count	Proportion
ca	1,722,653	$3.37 \times 10^{-2}$
ch	1,237,612	$2.42 \times 10^{-2}$
at	2,592,333	$5.07 \times 10^{-2}$
ht	552,908	$1.08 \times 10^{-2}$
ta	997,182	$1.95 \times 10^{-2}$
th	4,104,575	$8.03 \times 10^{-2}$
ae	73,151	$1.43 \times 10^{-3}$
he	3,088,885	$6.04 \times 10^{-2}$
cat	192,728	$3.77 \times 10^{-3}$
cht	4,373	$8.56 \times 10^{-5}$
tae	1,372	$2.69 \times 10^{-5}$
the	2,307,626	$4.52 \times 10^{-2}$

## Example 10 (cont'd): “THE CAT”

Comparison between actual (Bob's email) usage and dictionary

Sequence	Proportion Email	Proportion Dictionary	Ratio Email-to-Dictionary
ca	$3.37 \times 10^{-2}$	$5.29 \times 10^{-2}$	0.637
ch	$2.42 \times 10^{-2}$	$5.02 \times 10^{-2}$	0.482
at	$5.07 \times 10^{-2}$	$9.85 \times 10^{-2}$	0.515
ht	$1.08 \times 10^{-2}$	$6.53 \times 10^{-3}$	1.654
ta	$1.95 \times 10^{-2}$	$5.04 \times 10^{-2}$	0.387
th	$8.03 \times 10^{-2}$	$3.72 \times 10^{-2}$	2.159
ae	$1.43 \times 10^{-3}$	$7.85 \times 10^{-3}$	0.182
he	$6.04 \times 10^{-2}$	$4.88 \times 10^{-2}$	1.238
cat	$3.77 \times 10^{-3}$	$8.91 \times 10^{-3}$	0.423
cht	$8.56 \times 10^{-5}$	$7.80 \times 10^{-4}$	0.110
tae	$2.69 \times 10^{-5}$	$2.59 \times 10^{-4}$	0.104
the	$4.52 \times 10^{-2}$	$1.12 \times 10^{-2}$	4.036

# Google: All Our N-gram are Belong to You

File sizes: approximately 24GB compressed (gzip'ed) text files

Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

Credit: [Alex Franz and Thorsten Brants, Google Machine Translation Team](#)

# Norvig: Natural Language Corpus Data

Applications considered:

- Word Segmentation
- Secret Codes
- Spelling Correction

Other tasks (briefly described):

- Language Identification
- Spam Detection and Other Classification Tasks
- Author Identification (Stylometry)
- Document Unshredding and DNA Sequencing
- Machine Translation

## Example 10 (one last time): “THE CAT”

Data for this table derived from the **Google Web Trillion Word Corpus**

There are  $26 \times 26 = 676$  distinct two letter combinations and  
 $26 \times 26 \times 26 = 17,576$  distinct three letter combinations of the  
characters A–Z

Sequence	Count	Rank
ca	35,964,886,206	51
ch	34,170,408,619	54
at	80,609,883,139	11
ht	10,529,516,272	172
ta	42,344,542,093	42
th	133,210,262,170	2
ae	12,677,383,584	384
he	106,498,528,786	5
cat	6,716,493,366	133
cht	875,755,609	1,571
tae	55,488,859	5,242
the	82,103,550,112	1

## Chris' Example Revisited

The monkeys ate the bananas because they were ripe

The monkeys ate the bananas because they were hungry



# Google Translate API

***Important:** Google Translate API v2 is now available as a paid service. The courtesy limit for existing Translate API v2 projects created prior to August 24, 2011, will be reduced to zero on December 1, 2011. In addition, the number of requests your application can make per day will be limited. Google Translate API v1 will be shut off completely on the same date (December 1, 2011); it was officially deprecated on May 26, 2011. These changes are being made due to the substantial economic burden caused by extensive abuse. For website translations, we encourage you to use the Google Website Translator gadget*

<http://code.google.com/apis/language/translate/overview.html>

# Google Translate API

*Google is dropping an automatic-translation tool, because overuse by spam-bloggers is flooding the internet with sloppily translated text, which in turn is making computerized translation even sloppier*

*... I think it is very interesting in its implications – about language, “big data,” Google’s strategies, and the never-ending recalibration of goods vs bads, “signal to noise,” on the internet*

“An ‘Economic Burden’ Google Can No Longer Bear?”

James Fallows, The Atlantic, June 12, 2011

[http://www.theatlantic.com/technology/archive/2011/06/  
an-economic-burden-google-can-no-longer-bear/240283/](http://www.theatlantic.com/technology/archive/2011/06/an-economic-burden-google-can-no-longer-bear/240283/)