# HG2052
## Language, Technology and the Internet

# Introduction, Organization: Overview, Main Issues

Francis Bond

**Division of Linguistics and Multilingual Studies**

http://www3.ntu.edu.sg/home/fcbond/

bond@ieee.org

Lecture 1

# Introduction

➢ How technology affects our use of language

➢ How language is used on the internet

    ➢ Some fun things we can now do, that we couldn't before

➢ Collaboration and shared authoring

➢ Meta-languages

➢ The Semantic Web

# Goals

➤ Gain an understanding of how technology affects language use

➤ Develop familiarity with markup and meta information in texts

➤ Get a feel for what research is all about, especially relating to web mining and online frequency counting

**Upon successful completion, students will:**

➤ have an understanding of how technology shapes language use

➤ be able to test linguistic hypotheses against web data.

# Administrivia

**Coordinator** Francis Bond

**Email** bond@ieee.org; **Subject** [HG2052]

**\*** Seminar: combined lecture/tutorial

The timetable and slides are on the web: http://compling.hss.ntu.edu.sg/courses/hg2052/

# Assessment

➤ *Continuous Assessment* (100%)

  ➤ Assignment One (30%);
    * Describe one new modality of communication and compare it to speech and text
  ➤ Assignment Two (30%) Group Work
    * Create or enhance a wikipedia page about linguistics
  ➤ Assignment Three (30%)
    * TBA
  ➤ Classroom participation (10%)

# Recommended Texts

➤ *Wikipedia*

➤ Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press

➤ Crystal, D. (2001). *Language and the Internet*. Cambridge University Press

➤ Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly. (`www.nltk.org/book`)

➤ Sproat, R. (2010). *Language, Technology, and Society*. Oxford University Press

➤ Dickinson, M., Brew, C., and Meurers, D. (2013). *Language and Computers*. Wiley-Blackwell

# Complimentary Courses

➤ **HG2051 Language and the Computer** — solving NLP problems with Python: introduces both programming and linguistics

➤ **HG3051 Corpus Linguistics** — empirical research on language use

# Guidelines for Written Work in LMS

➢ All assignments must follow the *Guidelines to Submitting Written Work for the Division of Linguistics and Multilingual Studies*

  ➢ You can get it from: http://linguistics.hss.ntu.edu.sg/CurrentStudents/Pages/Resources.aspx
  ➢ Useful advice on citation, transcription, formatting
  ➢ Proper citation is important
    — failure to cite is plagiarism — **fail subject**
    See the NTU code of academic integrity
    http://academicintegrity.ntu.edu.sg/

➢ I also strongly recommend my own style guide: *(Computational) Linguistic Style Guidelines: a guide for the flummoxed*

  ➢ You can get it from: http://www3.ntu.edu.sg/home/fcbond/data/ling-style.pdf

# House Rules

➤ No late work without prior permission

➤ Eating in class OK, so long as I can browse

➤ Talking encouraged (but only to the whole class)

➤ Sleeping tolerated, but your own bed recommended

➤ Start at 5 minutes past the half hour

# Acknowledgments and Disclaimer

These slides contain material from Steven Bird, and various other people from the web.

**Disclaimer:**

➢ I am experimenting with using fewer slides and talking more

   ➢ You will have trouble if you don't come to class

➢ I have vetted and will watch all Wikipedia pages I cite

   ➢ i.e., I have vetted them once, and will monitor changes.

➢ I am trying to find good on-line readings

# Themes

➤ Language and Technology

    ➤ Writing and Speech Technology

➤ Language and the Internet

    ➤ Email; Chat; Virtual Worlds; WWW; IM; Blogs; Facebook; Wikis; Twitter

➤ The Web as Corpus

➤ The Web beyond Language

    ➤ Semantic Web and Networks

# Language and Technology

➤ The two things that separate humans from animals (Sproat, 2010, §1.1)

   ➤ Language
      ∗ large vocabulary (10,000+)
      ∗ complicated syntax (no upper length; recursion; embedding)
   ➤ Technology
      ∗ Widespread tool use
      ∗ Widespread tool manufacture

➤ Speech — the start of language

➤ Writing — the first great intersection

# Language and the Internet

➢ New forms of communication

  ➢ Neither speech nor text
  ➢ Massively interactive

➢ Extremely rapid change

➢ A first hand narrative (I was online before the internet :-)

  ➢ but I am probably behind you all now

# New forms

➤ Email (from PC, phone, other)

➤ Chat; Usenet

➤ Virtual Worlds

➤ WWW

➤ Blogs (overlap)

➤ Facebook, LinkedIn

➤ Wikis

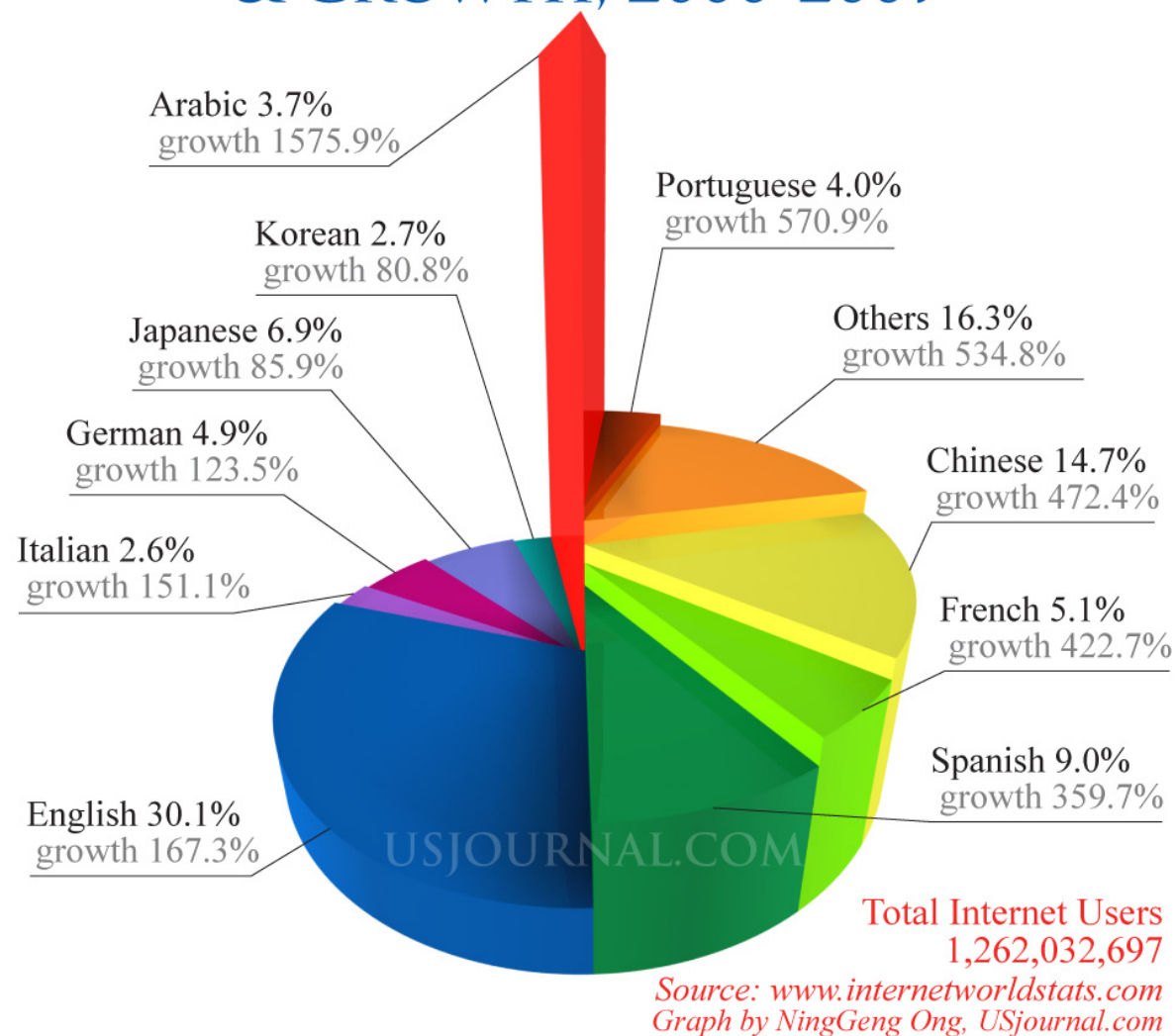➤ Twitter

# Hidden data on the web

➤ Web query logs

  ➤ http://www.google.com/trends
  ➤ http://www.google.org/flutrends/about/how.html

➤ Clicks and time browsing

# The Internet and Language Diversity



INTERNET USAGE BY LANGUAGE 2007 & GROWTH, 2000-2007

Arabic 3.7%
growth 1575.9%

Korean 2.7%
growth 80.8%

Japanese 6.9%
growth 85.9%

German 4.9%
growth 123.5%

Italian 2.6%
growth 151.1%

English 30.1%
growth 167.3%

Portuguese 4.0%
growth 570.9%

Others 16.3%
growth 534.8%

Chinese 14.7%
growth 472.4%

French 5.1%
growth 422.7%

Spanish 9.0%
growth 359.7%

Total Internet Users
1,262,032,697
Source: www.internetworldstats.com
Graph by NingGeng Ong, USjournal.com

USJOURNAL.COM

# Gradually Changing



Top Ten Languages in the Internet
in Millions of users - December 2017

| Language | Millions of Users |
|---|---|
| English | 1052 |
| Chinese | 804 |
| Spanish | 337 |
| Arabic | 219 |
| Portuguese | 169 |
| Indonesia | 168 |
| French | 134 |
| Japan | 118 |
| Russia | 109 |
| German | 92 |
| All the rest | 950 |

**Millions of Users**
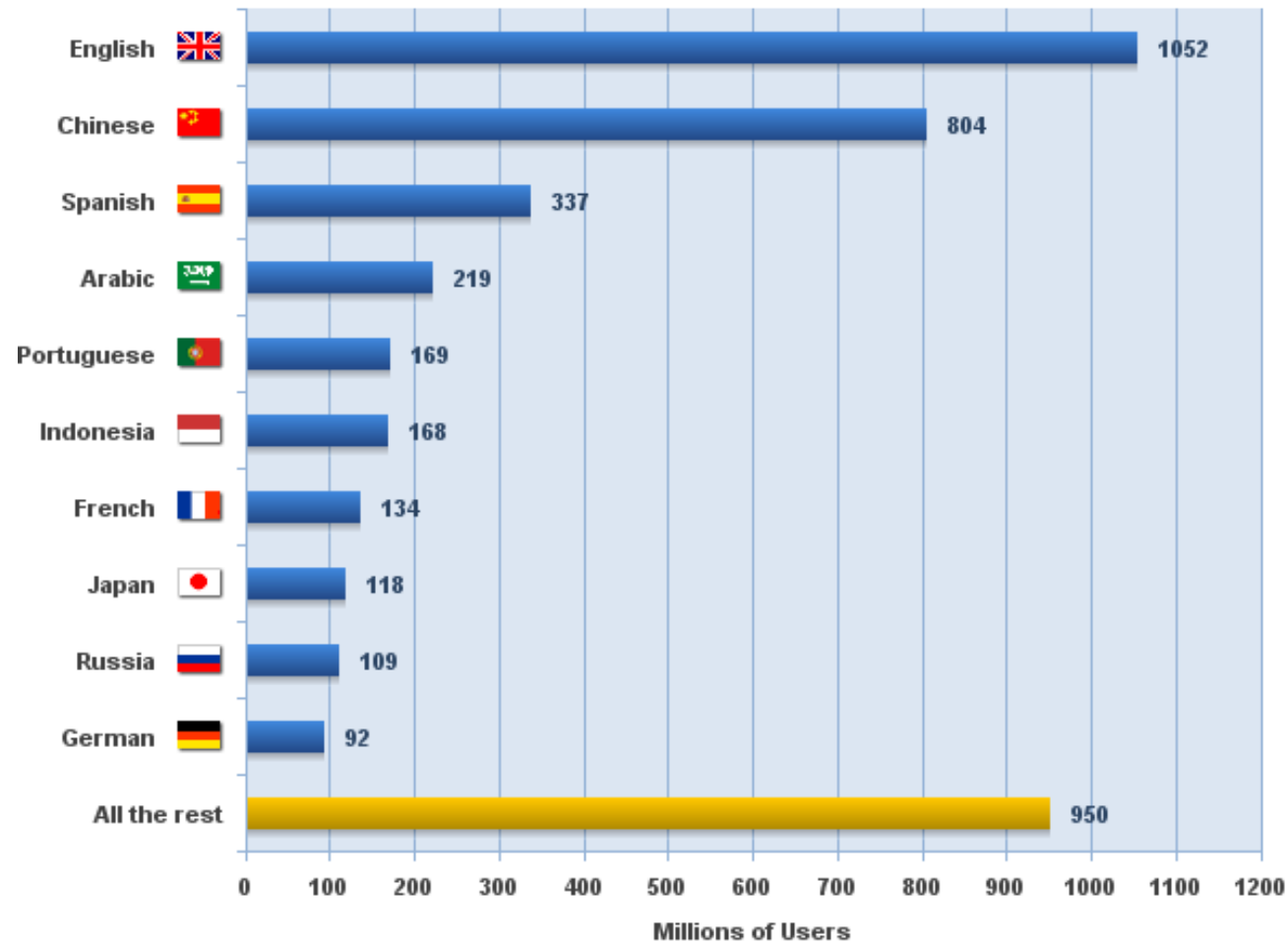
Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated total Internet users are 4,156,932,140 in December 31, 2017
Copyright © 2018, Miniwatts Marketing Group

"On the Internet, nobody knows you're a dog."

Page 61 of July 5, 1993 issue of The New Yorker, (Vol.69 (LXIX) no. 20)

http://www.unc.edu/depts/jomc/academics/dri/idog.html

# Some Technical Terms (the squiggly bits)

# A gentle introduction to Information theory

➢ Language has many uses, only one of which is to convey information
  *A bit of Fry and Laurie Season 1 Episode 2*
  — but surely transferring information is important

➢ How can we measure information?

  ➢ Information as Bits
    Shannon, C.E. (1948), "A Mathematical Theory of Communication", Bell System Technical Journal, 27, pp. 379–423 & 623–656, July & October, 1948. `http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf`

  ➢ Minimum Description Length
    Andrey Kolmogorov (1968), "Three approaches to the quantitative definition of information" in International Journal of Computer Mathematics.

# Consider an abstract case

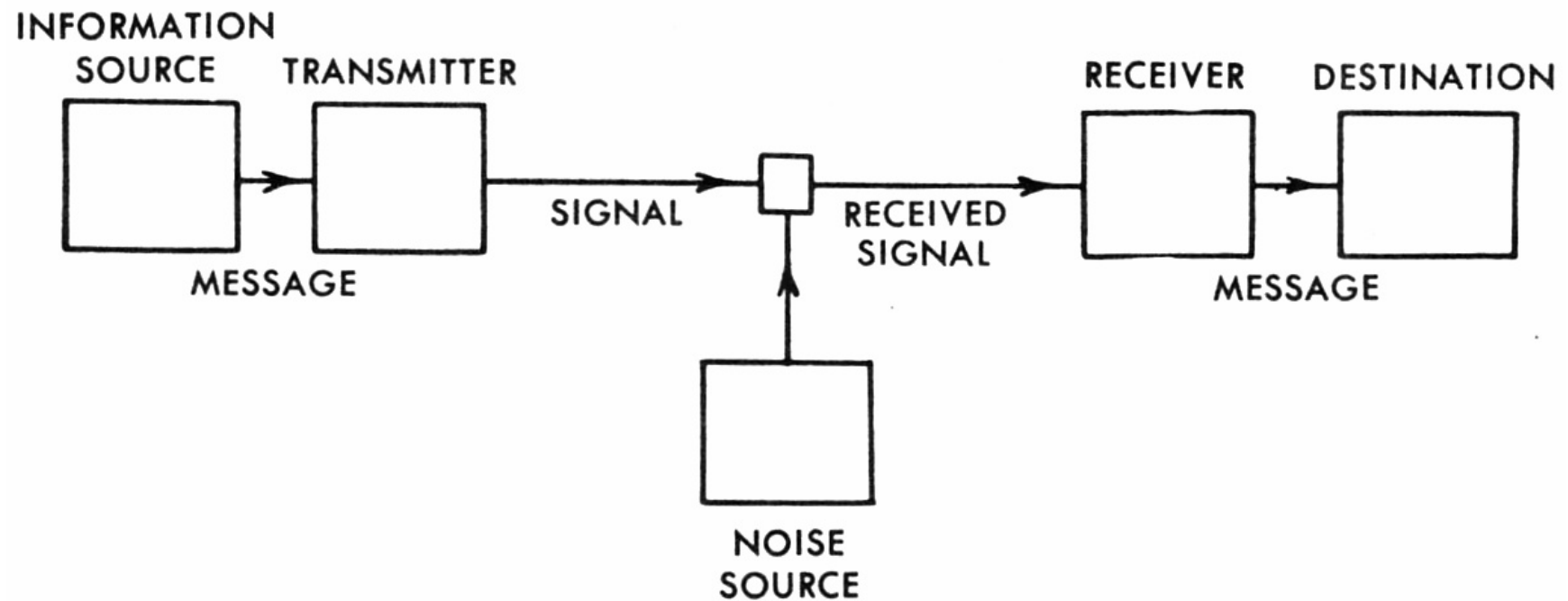34      *The Mathematical Theory of Communication*



Fig. 1. — Schematic diagram of a general communication system.

# Information as bits

➢ Suppose we have 8 equally likely facts (A, B, C, ...H).
  How many yes-no questions does it take to pinpoint the fact?

  ➢ This is the information in bits
    you need $n$ bits of information

➢ Technically the Entropy

$$H(\mathrm{p}) = -\sum_{x \in X} \mathrm{p}(x) \log_2 \mathrm{p}(x)$$

  ➢ We won't go there in HG2052

# What if some facts are more common?

➢ Simplified Polynesian:
p $\left(\frac{1}{8}\right)$; k $\left(\frac{1}{8}\right)$; i $\left(\frac{1}{8}\right)$; u $\left(\frac{1}{8}\right)$; t $\left(\frac{1}{4}\right)$; a $\left(\frac{1}{4}\right)$

➢ We can do this in $2\frac{1}{2}$ bits

   ➢ Is it (a or t) or ( p, k, i, u), …

➢ We can define a $2\frac{1}{2}$ bit code

   ➢ p (100); k (101); i (110); u (111); t (00); a (01)

➢ Which codes are longer — frequent or infrequent letters?

   ➢ What does this imply for language?

   (Manning and Schütze, 1999, §2.2)

# What if there is mutual information?

➤ What is the next letter?

➤ What is the next letter following /./?

    t
    a
    q

➤ What is the next letter following /../?

    th
    as
    qu

➤ A language model and more context improves our guess

# Different Models for English

Consider only 26 lowercase letters and a space, and a language model based on probability (Hidden Markov Model). How many guesses do we need on average to guess the next letter?

➢ Zeroth order (random) $= \log_2 27 = 4.76$

➢ First order (frequency) $= 4.03$  (pick $e$)

➢ Second order (one previous letter) $= 2.8$

➢ Human $= 1.34$

Surrounding context helps interpretation

# What if there is noise?

➤ Imagine you want to send a signal, but randomly a bit gets flipped (noise)

  ➤ Original message /pa/ 100 01
  ➤ Received message /ka/ 101 01

➤ If we make the message longer, we can guard against this

  ➤ Original message /pa/ 100 100 100 01 01 01
  ➤ Received message /ka/ 101 100 100 01 01 01
    We add redundancy to the signal
  ➤ There are much better encodings than this (Hamming codes)

➤ Hmn lngge s vr rdndnt

# The cloze test

Modern linguistics has been             to provide theories and             for the specification of             that express this mapping             a declarative and transparent             .  Computational linguistics has contributed             platforms and tools for             development.

A few large             grammars have been             that exhibit sufficient             and coverage for             application tasks.  However,             encouraging developments were             hampered by a             of methods for             analysis that fulfill             minimal requirements in             , robustness, and specificity.

This simply          that all          working with          grammars have
    too slow          too brittle          real applications.          , they have
    been able          manage the          ambiguity in          language, i.e.
    could not          among large          of linguistically          analyses.

# How did we go?

Modern linguistics has been able to provide theories and formalisms for the specification of grammars that express this mapping in a declarative and transparent way. Computational linguistics has contributed elaborate platforms and tools for grammar development.

A few large scale grammars have been designed that exhibit sufficient accuracy and coverage for real application tasks. However, these encouraging developments were seriously hampered by a lack of methods for language analysis that fulfill the minimal requirements in efficiency, robustness, and specificity.

This simply means that all systems working with these grammars have been too slow and too brittle for real applications. Furthermore, they have not been able to manage the vast ambiguity in natural language, i.e. they could not select among large numbers of linguistically correct analyses.

http://www.delph-in.net/index.php?page=1

# Minimum Description Length

➢ Which message has more information?

  ➢ ababababababababababab
  ➢ lakdsfiuy,mwskfsfdends

➢ We can write the first as "ab 11 times" half as many letters

➢ The Minimum description length is the shortest description in some fixed description language
note: MDL includes the algorithm and data

➢ The difference between the actual length and the MDL is the redundancy

➢ English appears to have more redundancy that Chinese:
size(English/Chinese): Text — 1.49; Compressed — 1.14
http://158.130.17.5/~myl/languagelog/archives/002379.html

# DIKW

**Data** is unprocessed facts and figures without any added interpretation or analysis. "The price of crude oil is $80 per barrel."

**Information** is data that has been interpreted so that it has meaning for the user. "The price of crude oil has risen from $70 to $80 per barrel" gives meaning to the data and so is said to be information to someone who tracks oil prices.

**Knowledge** is a combination of information, experience and insight that may benefit the individual or the organisation. "When crude oil prices go up by $10 per barrel, it's likely that petrol prices will rise by 2p per litre" is knowledge.

**Wisdom** "knowing the right things to do"

# What about you?

➤ name (I won't remember it, sorry)

➤ what you hope to get out of the class

➤ languages you speak

➤ something else?

# HomeWork: Who uses what? (will you tell?)

➤ Telnet, ftp, ssh

➤ WWW

➤ Wikis

➤ Blogs (overlap)

➤ Email (from PC, phone, other)

➤ Virtual Worlds

➤ Facebook, LinkedIn

➤ Twitter, Tumbler, ...

➤ LT: MT, dictation, other

# HomeWork and Readings

➢ Keep a media usage diary for one day (Friday) and add it to the shared spreadsheet

➢ Read the following:

   ➢ What can search terms tell us?
      https://www.nature.com/articles/nature07634 (use the proxy)
   ➢ Which is more efficient: Chinese or English:
      http://itre.cis.upenn.edu/~myl/languagelog/archives/002379.html
   ➢ Rants about technology through the ages:
      Vaughan Bell (2010) Don't touch that Dial! *Slate*
      http://www.slate.com/id/2244198/ accessed 2010-09-03.