LTI
Language, Technology and the Internet

Introduction, Organization: Overview, Main Issues

Francis Bond
Department of Asian Studies
Palacký University
https://fcbond.github.io/
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: https://bond-lab.github.io/Language-Technology-and-the-Internet/

Assessment

- Assignment One (20%);
  - Describe one new modality of communication and compare it to speech and text
- Assignment Two (40%) Group Work
  - Create or enhance a wikipedia page about linguistics
- Assignment Three (30%)
  - New! Something to do with LLMs
    was phish, fake, harvest or analyze
- Classroom participation (10%)

Recommended Texts

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

Complimentary Courses

- **LAC Language and the Computer** — solving NLP problems with Python: introduces both programming and linguistics
- **??? Corpus Linguistics** — empirical research on language use

Guidelines for Written Work

- All assignments must follow the local guidelines
  - Be careful with citation, transcription, formatting
  - Proper citation is important
    — failure to cite is plagiarism — **fail subject**
- I also strongly recommend my own style guide:
  (Computational) Linguistic Style Guidelines: a guide for the flummoxed

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 on time, maybe finish early (depends on the week)
- I will try to be in my room for consultation, but safest to email me and confirm — I do not check teams often

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
- Language Models as tools
  - AI and generative language models

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

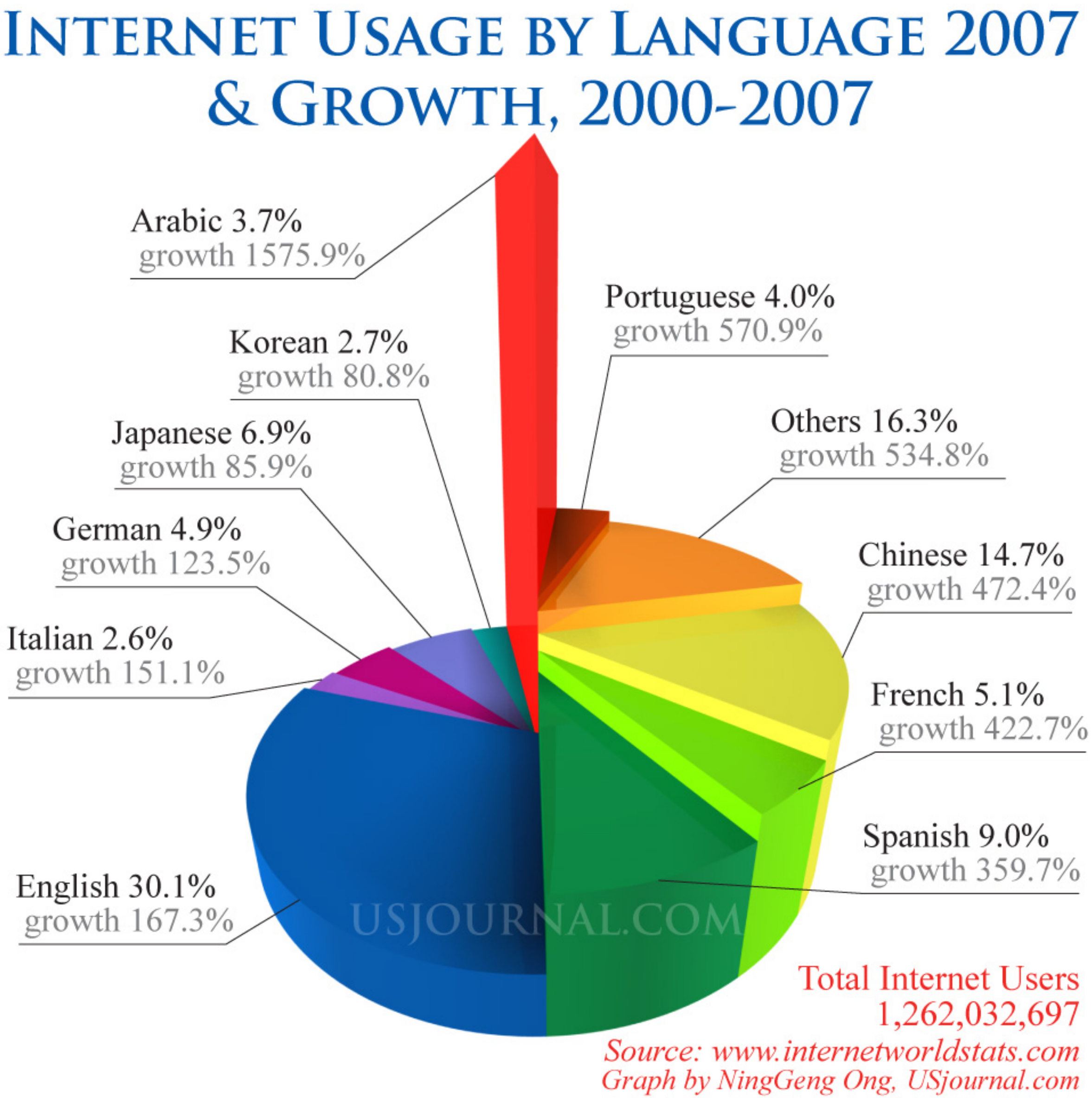
New forms

- Email (from PC, phone, other)
- Chat; Usenet
- Virtual Worlds
- WWW
- Blogs (overlap)
- Facebook, LinkedIn
- Wikis
- Twitter

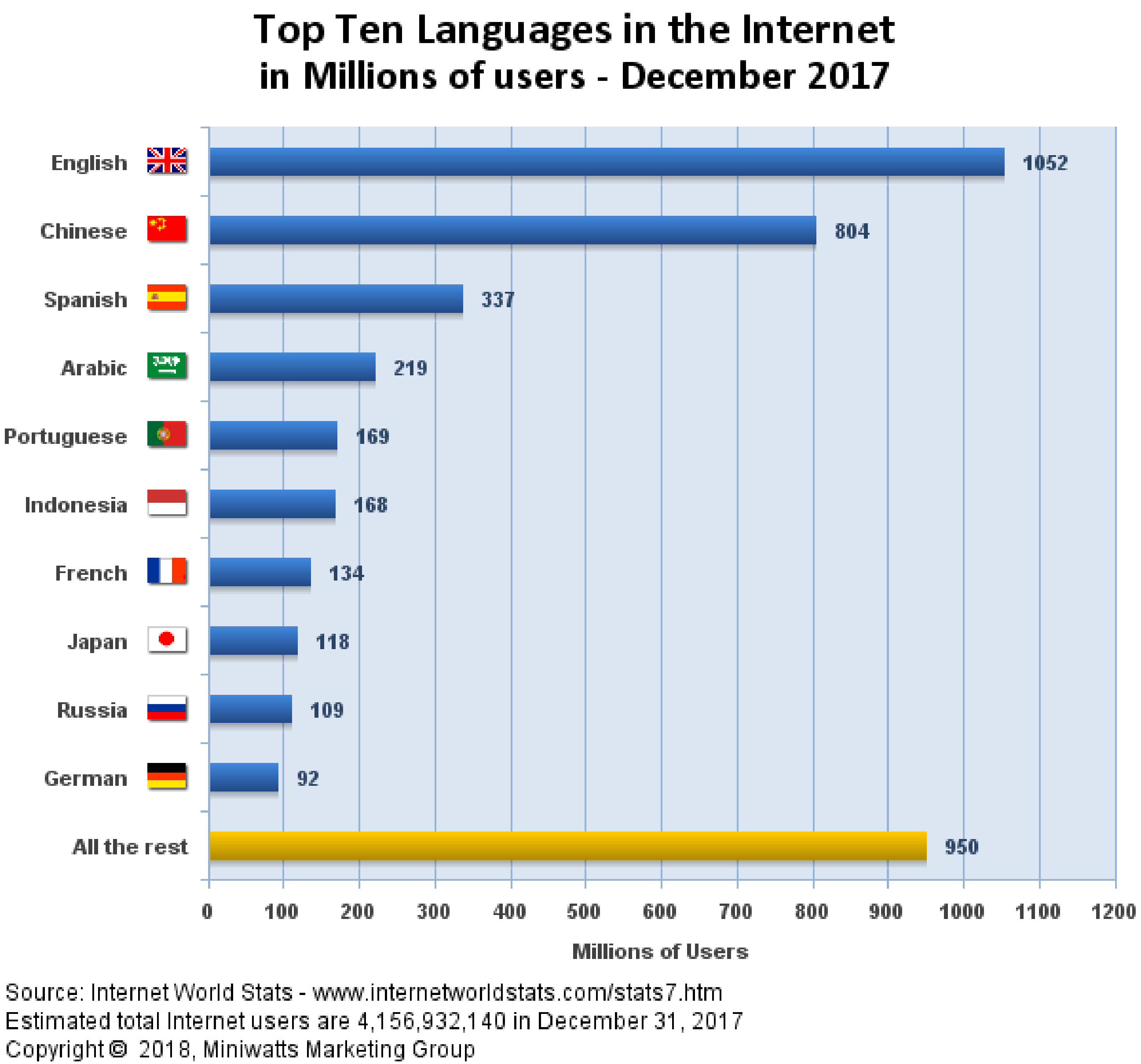
Hidden data on the web

- Web query logs
  - http://www.google.com/trends
  - https://en.wikipedia.org/wiki/Google_Flu_Trends
  - Google Flu Trends Shows Good Data > Big Data Mar 26, 2014 By Kaiser Fung
- Clicks and time browsing

The Internet and Language Diversity



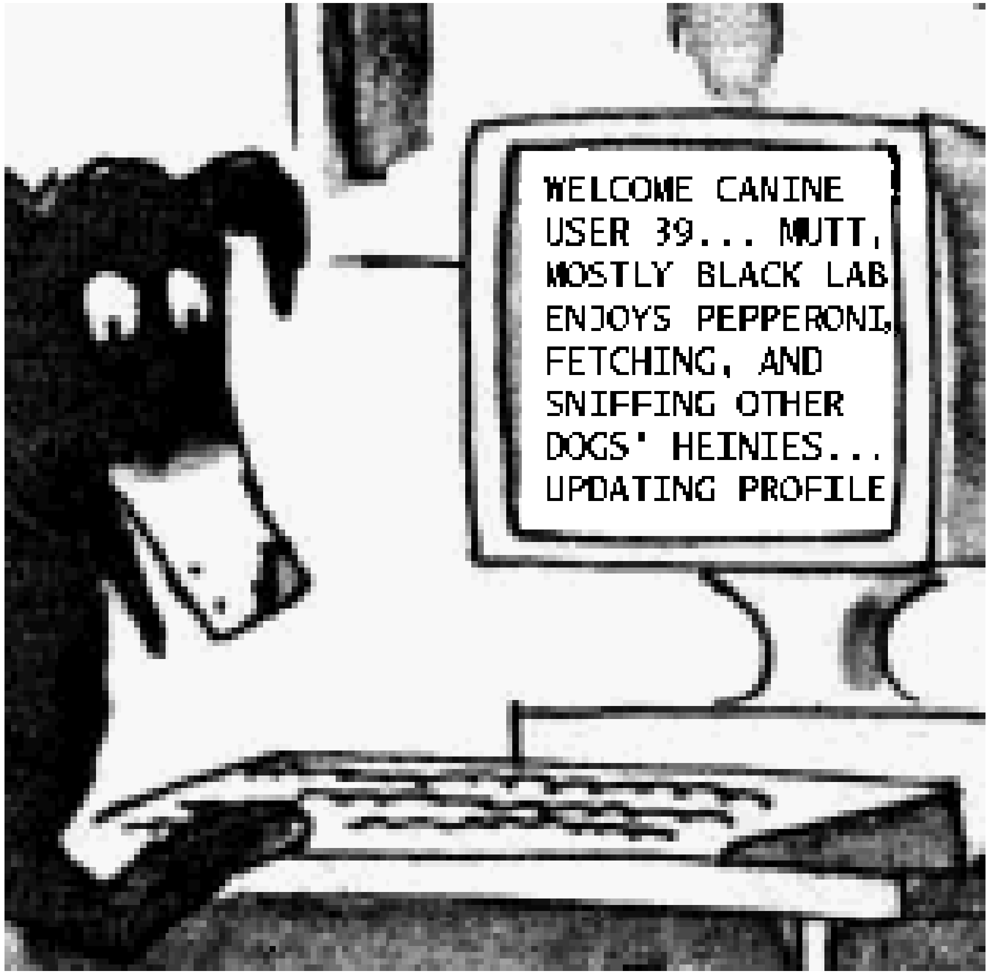
Gradually Changing



On the Internet, nobody knows you are a dog

"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 — Concerning Language
  — 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*

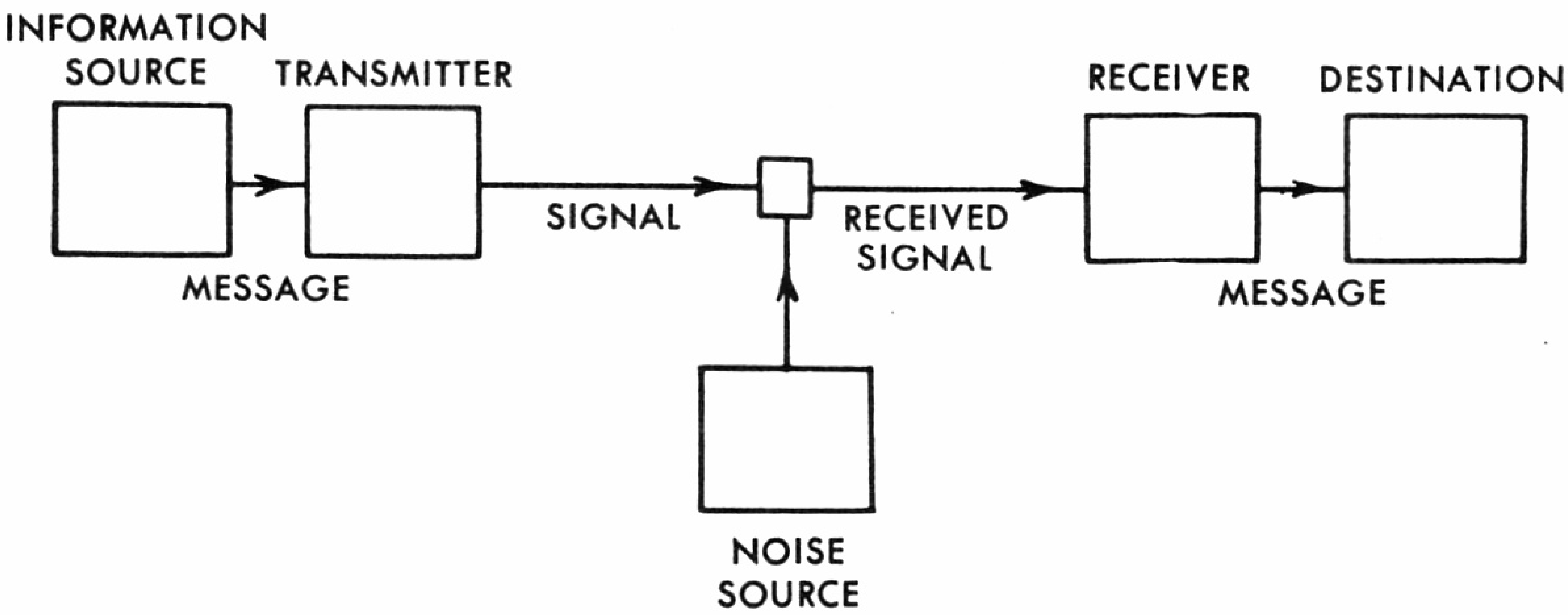


Fig. 1. — Schematic diagram of a general communication system.

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 this class

What if some facts are more common?

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

- 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?
  - abababababababababab
  - lakdsfiuy,mwskfsfdends
- We can write the first as "ab 11 times" half as many letters
- The Minimum description length (Kolgoromov complexity) is the shortest possible description in some fixed description language
  - MDL includes the algorithm and data
  - it approximates the entropy for long strings
- 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": we should burn less oil to lessen the climate catastrophe

What about you?

- name (I won't remember it, sorry), pronouns (I will do my best)
- 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)
- Chat (Whatsapp, Signal, Telegram)
- 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?
    Ginsberg, J., Mohebbi, M., Patel, R. et al. *Detecting influenza epidemics using search engine query data*. Nature 457, 1012–1014 (2009) https://doi.org/10.1038/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.