The Case for Empiricism (with and without statistics)

Kenneth Church

IBM

Kenneth.Ward.Church@gmail.com

Empirical ≠ Statisical

- These days, *empirical* and *statistical*
  - Are used somewhat interchangeably
  - But it wasn't always this way
  - (And probably, for good reason)
- In *A Pendulum Swung Too Far* (Church, 2011),
  - I argued that grad schools should make room for
  - Both *Empiricism* and *Rationalism*
- We don't know what will be hot tomorrow
  - But it won't be what's hot today
- We should prepare the next generation
  - For all possible futures (or at least all probable futures)
- This paper argues for a diverse interpretation of *Empiricism*
  - That makes room for everything
  - from Humanities to Engineering (and then some)

Pendulum Swung Too Far (Church, 2011)

- When we revived empiricism in the 1990s,
  - we chose to reject the position of our teachers for pragmatic reasons.
  - Data had become available like never before.
- What could we do with it?
  - We argued that it is better to do something simple than nothing at all.
  - Let's go pick some low hanging fruit.
- While trigrams cannot capture everything,
  - they often work better than alternatives.
  - It is better to capture the agreement facts that we can capture easily,
    - than to try for more and end up with less.
- That argument made a lot of sense in the 1990s,
  - especially given unrealistic expectations
  - that had been raised during the previous boom.
- But today's students might be faced with a very different set of challenges in the not-too-distant future.
  - What should they do when most of the low hanging fruit
  - has been picked over?

Linguistic Representations

- Fillmore
  - Sound & Meaning >> Spelling
- Jelinek
  - *Every time I fire a fire a linguist,*
  - *performance goes up*

Some of my Best Friends are Linguists

(LREC 2004)

Frederick Jelinek
Johns Hopkins University

THANKS TO: E. Brill, L. Burzio, W. Byrne, C. Cieri, J. Eisner, R. Frank, L. Guthrie, S. Khudanpur, G. Leech, M. Liberman, M. Marcus, M. Palmer, P. Smolensky, and D. Yarowsky

On firing linguists…

Introduction to the Special Issue on Computational Linguistics Using Large Corpora

Kenneth W. Church*
AT&T Bell Laboratories

Robert L. Mercer†
IBM T.J. Watson Research Center

- Finally, they removed the dictionary lookup HMM,
  - taking for the pronunciation of each word its spelling.
  - Thus, a word like *t-h-r-o-u-g-h* was assumed to have a pronunciation like *tuh huh ruh oh uu guh huh*.
- After training, the system learned that
  - with words like *l-a-t-e* the front end often missed the *e*.
  - Similarly, it learned that *g*'s and *h*'s were often silent.
  - This crippled system was still able to recognize
    - 43% of 100 test sentences correctly as compared with
    - 35% for the original Raleigh system.

On firing linguists… (2 of 2)

- These results firmly established the importance of a ***coherent, probabilistic*** approach to speech recognition and the importance of data for estimating the parameters of a probabilistic model.
  - One by one, pieces of the system that had been assiduously assembled by speech experts yielded to probabilistic modeling.
  - Even the elaborate set of hand-tuned rules for segmenting the frequency bank outputs into phoneme-sized segments would be replaced with training (Bakis 1976; Bahl et al. 1978).
- By the summer of 1977, performance had reached **95%** correct by sentence and 99.4% correct by word,
  - a considerable improvement over the same system with hand-tuned segmentation rules (**73%** by sentence and 95% by word).
- Progress in speech recognition at Yorktown and almost everywhere else as well has continued along the lines drawn in these early experiments.
  - As computers increased in power, ever greater tracts of the ***heuristic wasteland*** opened up for colonization by probabilistic models.
  - As greater quantities of recorded data became available,
    - these areas were tamed by automatic training techniques.

Sound & Meaning >> Spelling

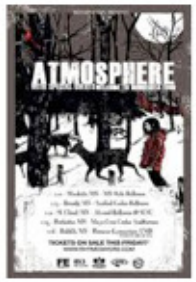
LTA-2012: Charles J Fillmore

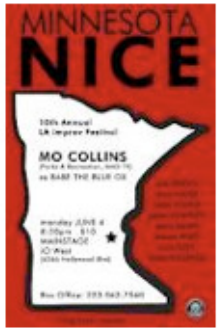
- Technology
  - Video/Skype
  - Credits:
    - Lily Wong Fillmore
- Highlights
  - Case for Case
    - 7k citations in Google Scholar
  - Framenet
    - 2 papers with 1k citations each
- "Minnesota Nice"
  - Nice things to say about everyone: Chomsky/Schank
  - Self-deprecating humor
    - (but don't you believe it)

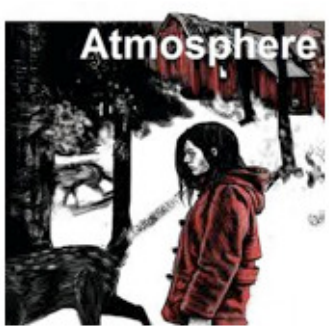
Migration from the cold: Minnesota → Berkeley

“Minnesota Nice”

(Stereotypes aren’t nice, but…)







The “Minnesota Nice” Version

Of the story of Chuck’s migration
from Minnesota to Berkeley

Self-deprecating humor (but don’t you believe it)

The Significance of *Case for Case*: C4C

- For many of us in my generation,
  - C4C was the introduction to a world
  - beyond Rationalism and Chomsky
- This was especially the case for me,
  - since I was studying at MIT,
  - where we learned many things
  - (but not Empiricism).

Case for Case (C4C): Practical Apps

- Information Extraction (MUC)
- Semantic Role Labeling

- Key Question: Who did what to whom?
  - Not: What is the NP and the VP of S?

Commercial Information Extraction

Google

Noam Chomsky

Kenneth Church + Share

Search

About 11,700,000 results (0.22 seconds)

Web
Images
Maps
Videos
News
Shopping
Books
More

Show search tools

Noam Chomsky - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Noam_Chomsky** - Cached
Avram **Noam Chomsky** is an American linguist, philosopher, cognitive scientist, historian, and activist. He is an Institute Professor and Professor (Emeritus) in the ...

Noam Chomsky's political views
Noam Chomsky is a widely known intellectual, political activist, and ...

Bibliography
Noam Chomsky bibliography. From Wikipedia, the free ...

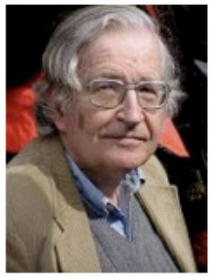
More results from wikipedia.org »

chomsky.info : The **Noam Chomsky** Website
www.**chomsky**.info/ - Cached
Official **Noam Chomsky** website. Resource site including books, articles, interviews, talks, debates, plus writings about Chomsky.
↳ Articles - Audio & Video - Books - Interviews

Images for **noam chomsky** - Report images

MIT Department of Linguistics: People: Faculty: **Noam Chomsky**
web.mit.edu/linguistics/people/faculty/**chomsky**/index.html - Cached

Noam Chomsky



en.wikipedia.org

Avram Noam Chomsky is an American linguist, philosopher, cognitive scientist, historian, and activist. He is an Institute Professor and Professor in the Department of Linguistics & Philosophy at MIT, where he has worked for over 50 years. Wikipedia

Born: December 7, 1928 (age 83), East Oak Lane

Spouse: Carol Chomsky (m. 1949–2008)

Education: University of Pennsylvania (1955), University of Pennsylvania (1951), More

Appears in: Manufacturing Dissent, Noam Chomsky: Rebel Without a Pause, More

Children: Aviva Chomsky, Harry Chomsky, Diane Chomsky

Books

Do Read “Case for Case”

- Great arg but also
  - Demonstrates strong command of
    - Classic literature as well as
    - Linguistic facts
- Our field:
  - Too “silo”-ed
  - Too few citations to
    - Classic literature, other fields and other types of facts
- We could use more “Minnesota Nice”

Historical Motivation: A Case for Case From Morphology → MUC

- Context Free Grammar is attractive for
  - Langs with more word order and less morphology (English)
- But Case Grammar is attractive for
  - Langs with more morphology and less word order
  - Examples: Latin, Greek & Japanese
- Latin (over-simplified):
  - Subject: Nominative case
  - Object: Accusative case
  - Indirect Object: Dative case
  - Other args: Ablative case

Japanese I/Vocabulary/Case Markers

< Japanese I | Vocabulary

These are to be placed after a wo

- wa（は） - topic marker
- ga（が） - subject
- (w)o（を） - direct object
- mo（も） - "also" (substitutes
- no（の） - possessive (revers
- na（な） - marks an adjective
- de（で） - "by means of", "in"
- ni（に） - indirect object, "in"/
- to（と） - "and", object of "sa
- ya（や） - "and" for a list
- (h)e（へ） - destination "to"
- ka（か） - question mark (poli

Using these semantic features valency patterns of the basic predicates necessary in the task domain are defined. As an example, the predicate **'okuru'** ('send' in English) is given the following valency patterns:

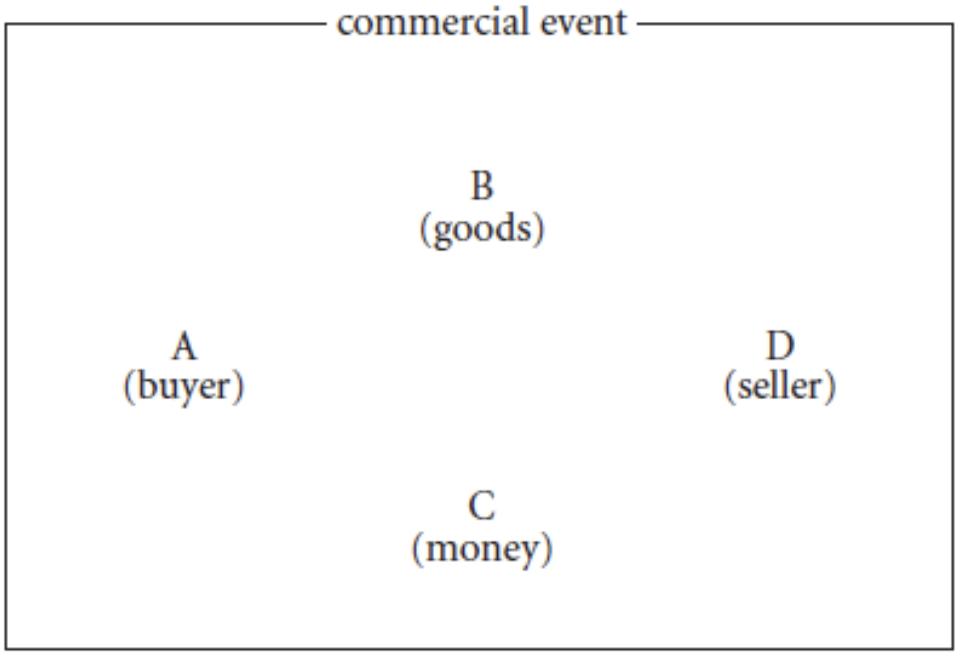
N[con/-tra]***'wo'*** + V,
N[con/-tra]***'wa'*** + N[loc]***'ni'*** + V,
N[con/-tra]***'wa'*** + N[hum]***'ni'*** + V,
N[con/-tra]***'wa'*** + N[tim/pro]***'madeni'*** + V,
N[tim/pro]***'madeni'*** + N[con/-tra]***'wo'*** + V,
N[hum]***'ni'*** + N[con/-tra]***'wo'*** + V,
N[hum]***'ga'*** + N[con/-tra]***'wo'*** + V,
N[hum]***'ga'*** + N[con/-tra]***'wo'*** + N[hum/-pro]***'ni'*** + V,
N[hum/-pro]***'wa'*** + N[con/-tra]***'wo'*** + N[hum]***'ni'*** + V,

Nouns: instrument	***Jitensha de*** *ikimashō.* 自転車で行きましょう。	Let's go **by bicycle.**
Nouns: location	***Koko de*** *yasumitai.* ここで休みたい。	I want to rest **here.**

C4C: Capturing Generalizations over Related Predicates & Arguments

VERB	BUYER	GOODS	SELLER	MONEY	PLACE
buy	subject	object	from	for	at
sell	to				
cost	indirect object	subject		object	at
spend	subject	on		object	at

Harry bought the puppy from Mr. Smith for $60.
Harry bought the puppy with the $60 that his mot
Mr. Smith sold the puppy.
Mr. Smith sold the puppy to Harry.
Mr. Smith sold the puppy for $60.



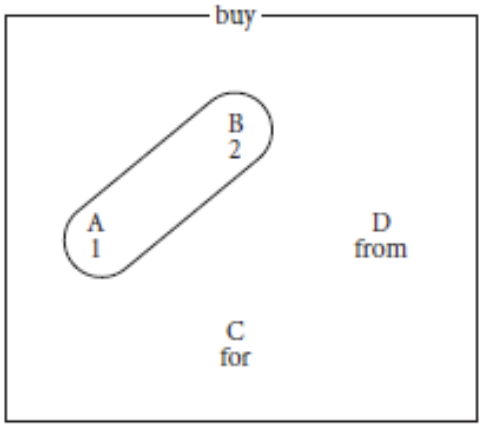


Fig. 3

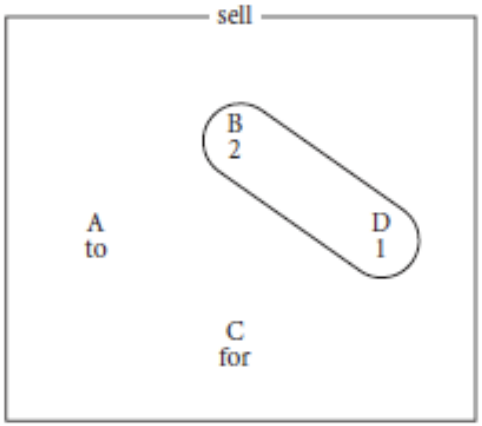


Fig. 4

to him.)

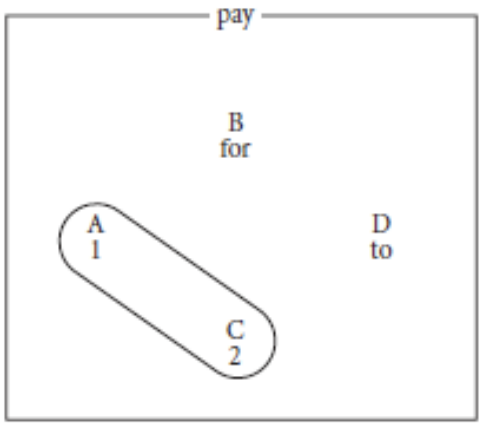


Fig. 5

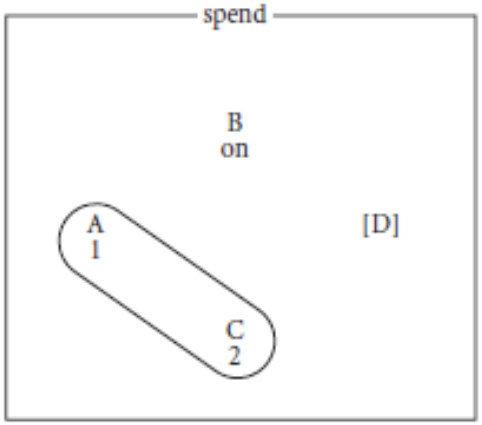


Fig. 6

C4C: Deep Cases → Surface Order/Morphology/Preps

"The case for case" Fillmore, 1968

- An alternative was spelled out in a rambling paper called "the case for case".
- It proposed a universal list of semantic role types ("cases"), configurations of which could characterize the st

Case Grammar → Frames / Lexicography

Valency → Scripts (Roger Schank) / Lexicography (Sue Atkins)

- Valency: Predicates have args (optional & required)
  - Example: "give" requires 3 arguments:
    - Agent (A), Object (O), and Beneficiary (B)
    - Jones (A) gave money (O) to the school (B)
  - Latin Morphology: Nominative, Accusative & Dative
- Frames
  - Commercial Transaction Frame: Buy/Sell/Pay/Spend
  - *Save* <good thing> *from* <bad situation>
  - *Risk* <valued object> *for* <situation>|<purpose>|<beneficiary>|<motivation>
- Collocations & Typical predicate argument relations:
  - *Save whales from extinction* (not vice versa)
  - *Ready to risk everything for what he believes*
- Representation Challenges: What matters for practical apps/NLU?
  - Stats on POS? Word order? Frames (typical predicate-args/collocations)?

Examples >> Definitions: Erode (George Miller)

Example: Save whales from extinction

Generalization: Save <good thing> from <bad thing>

- Exercise: Use "erode" in a sentence:
  - *My family erodes a lot.*
- ***to eat*** into or ***away***; destroy by slow consumption or disintegration
  - *Battery acid had eroded the engine.*
  - *Inflation erodes the value of our money.*
- Miller's Conclusion:
  - Dictionary examples are more helpful than definitions
- Implications for representations:
  - Stats on examples:
    - Easier to estimate/learn/apply than def/generalizations
  - Note: web search is currently more effective with
    - Examples (product number) than
    - Descriptions (cheap camera, camera under $200)

Definition

Examples

Corpus-Based Traditions: Empiricism Without Statistics

- As mentioned above,
  - There is a direct connection between Fillmore
  - And Corpus-Based Lexicographers (Sue Atkins)
- Corpus-based work has a long tradition in
  - lexicography,
  - linguistics,
  - psychology and
  - computer science
- Much of this tradition is documented in ICAME
- ICAME was co-founded by Francis
  - Brown Corpus: Francis and Kučera

Brown Corpus:

Influential across a wide range of fields

- Brown Corpus is cited by 10+ papers with 2k+ citations in 5+ fields:
  - Information Retrieval
    - Baeza-Yates and Ribeiro-Neto (1999)
  - Lexicography
    - Miller (1995)
  - Sociolinguistics
    - Biber (1991)
  - Psychology
    - MacWhinney (2000)
  - Computational Linguistics
    - Marcus et al (1993)
    - Jurafsky and Martin (2000)
    - Church and Hanks (1990)
    - Resnik (1995)
- All of this work is empirical,
  - though much of it is not all that statistical.

Empiricism in Humanities & Engineering

- The Brown Corpus and corpus-based methods have been particularly influential in the Humanities,
  - but less so in other fields such as Machine Learning and Statistics.
- I remember giving talks at top engineering universities and being surprised,
  - when reporting experiments based on the Brown Corpus,
  - that it was still necessary in the late 1990s to explain
    - what the Brown Corpus was,
    - as well as the research direction that it represented.
- While many of these top universities were beginning to warm up to statistical methods and machine learning,
  - there has always been less awareness of empiricism and
  - less sympathy for the research direction.

Little Room for Contrarians

- It is ironic how much the field has changed
  - (and how little it has changed).
- Back in the early 1990s,
  - it was difficult to publish papers that digressed
  - from the strict rationalist tradition
    - that dominated the field at the time.
- We created EMNLP/WVLC
  - to make room for empirical work (with and without statistics)
- These days,
  - it is difficult to publish a paper that digresses from today's fads (stats)
  - just as it used to be difficult
    - to publish papers that digressed from the fads of the day (rationalism)

Names of our meetings no longer make much sense

- There is less discussion than there used to be
  - Of the E-word in EMNLP, and
  - The the C-word in WVLC

Bitter Sweet Moment

- Kučera and Francis, Invited Talk, WVLC-1995
- Location: MIT
  - Long history of hostility to Empiricism
- Received a Standing Ovation
  - Mostly for their contribution to the field
  - But also because they both stood up for the hour
    - Even though they were well past retirement
    - (and standing wasn't easy)

Computational Linguistics → Engineering (away from Humanities)

- Unfortunately, while there was widespread appreciation for Kučera and Francis,
  - it was difficult for them to appreciate
    - what we were doing.
  - Henry tried to read my paper and others in WVLC-1995,
    - but they didn't make much sense to him.
- We had turned away from Humanities
    - (and C4C and FrameNet)
  - toward where we are today
    - (more Statistical than Empirical).

Challenge for Next Generation:

General Linguistics → Computational Linguistics

- Do methods from corpus-based lexicography scale up?
- Are they too manually intensive?
- If so, could we use machine learning methods
  - to speed up manual methods?
- Just as statistical parsers learn phrase structure rules (S → NP VP)
  - Can we learn valency?
  - Collocations?
  - Typical predicate argument relations?

When can we expect to learn frames?

- Corpus-size requirements:
  - freq(content words) ≈ parts per million
- 1970s Corpora: 1 M words (Brown Corpus)
  - Large enough to make a list of common content words
- 1990s: 100 M words (British National Corpus)
  - Large enough to see associations of common predicates with function words
    - "save" + "from"
  - Useful for parsing phrasal verbs: **V NP P** (Hindle & Rooth, 1993)
    - Most parsers are trained on Brown Corpus
    - (too small for phrasal verbs, let alone conjunction)
- Coming soon: $1M^2$ words (Google?)
  - Large enough to see associations of pairs of content words (collocations)
    - "give" + $$
    - "save" + "whale"
    - "save" + "extinction"
    - "risk" <valued object> for <purpose>
  - Useful for parsing every-way ambiguous Catalan Constructions (Church, 1980)
    - Conjunction, NN modification, PP attachment

Page Hits Estimates by MSN and Google

(August 2005)

Larger corpora → Larger counts → More signal

More Freq

Query	Hits (MSN)	Hits (Google)
A	2,452,759,266	3,160,000,000
The	2,304,929,841	3,360,000,000
Kalevala	159,937	214,000
Griseofulvin	105,326	149,000
Saccade	38,202	147,000

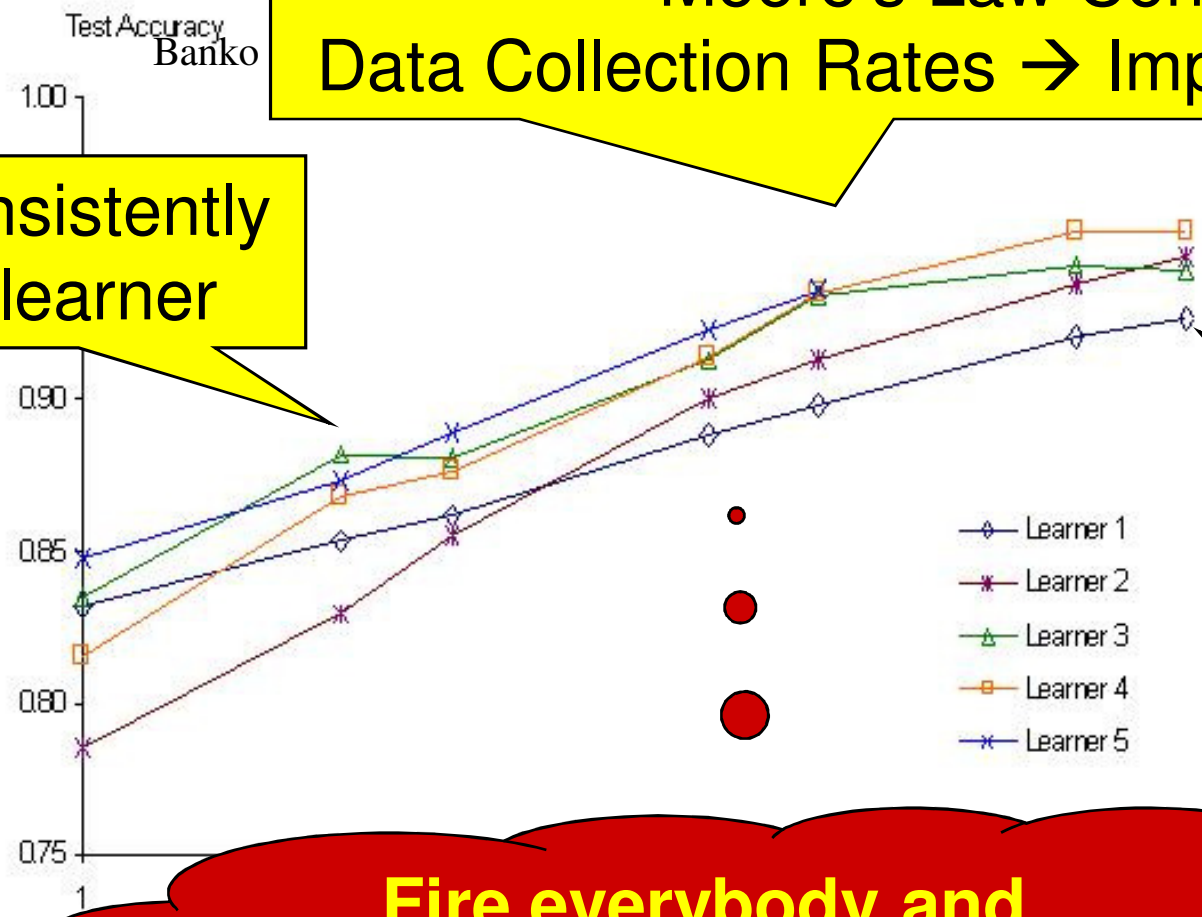
Less Freq

of (English) documents $D \approx 10^{10}$.
Lots of hits even for very *rare* words.

"It never pays to think until you've run out of data" – Eric Brill

Moore's Law Constant:
Data Collection Rates → Improvement Rates

No consistently best learner



More data is better data!

Fire everybody and spend the money on data

Quoted out of context

I(x;y)	fxy	fx	fy	x	y
10.47	7	7809	28	strong	northerly
9.76	23	7809	151	strong	showings
9.30	7	7809	63	strong	believer
9.22	14	7809	133	strong	second-place
9.17	6	7809	59	strong	runup
9.04	10	7809	108	strong	currents
8.85	62	7809	762	strong	supporter
8.84	8	7809	99	strong	proponent
8.68	15	7809	208	strong	thunderstorm
8.45	7	7809	114	strong	odor
8.66	7	1984	388	powerful	legacy
8.58	7	1984	410	powerful	tool
8.35	8	1984	548	powerful	storms
8.32	31	1984	2169	powerful	minority
8.14	9	1984	714	powerful	neighbor
7.98	9	1984	794	powerful	Tamil
7.93	8	1984	734	powerful	symbol
7.74	32	1984	3336	powerful	figure
7.54	10	1984	1204	powerful	weapon
7.47	24	1984	3029	powerful	post

Church and Hanks (1990)

Strong: 427M (Google)

Powerful: 353M (Google)

Google (2005) ≈ 1M x Church & Hanks (1990)

Counts increase 1000x per decade

Rising Tide of Data Lifts All Boats

If you have a lot of data, then you don't need a lot of methodology

- 1985: "*There is no data like more data*"
  - Fighting words uttered by radical fringe elements
  - (Mercer at Arden House)
- 1993 Workshop on Very Large Corpora
  - Perfect timing: Just before the web
  - Couldn't help but succeed
  - Fate
- 1995: The Web changes everything
- All you need is data (magic sauce)
  - No linguistics
  - No artificial intelligence (representation)
  - No machine learning
  - No statistics
  - No error analysis

It's tough to make predictions, especially about the future.

The Center For Language and Speech Processing
at the Johns Hopkins University

Seminars are on hiatus. Check back soon for updates.

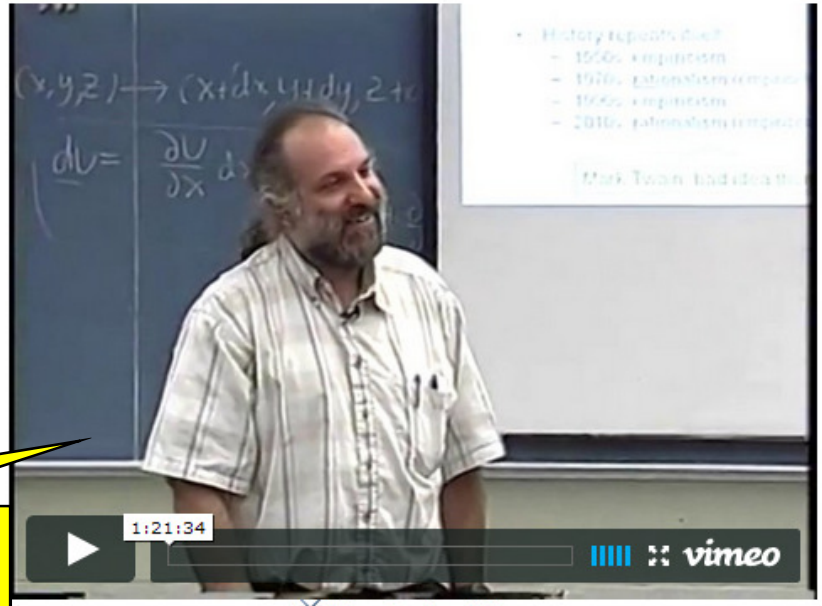
ABOUT
PEOPLE
PUBLICATIONS
SEMINARS
Upcoming Seminars
Past Seminars
2014 2013 2012 2011 2010 2009 2008 2007 2006 2005 2004
2003 2002 2001 2000 1999 1998 1997 1996 1995 1993 1990
WORKSHOPS
EVENT CALENDAR

Home » Seminars » Archive » 2003 » October 21, 2003

Speech and Language Processing: Where have we been and where are we going

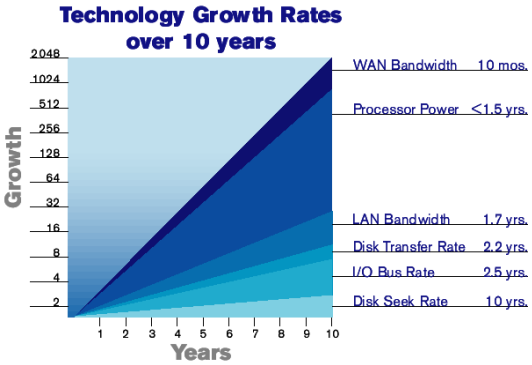
Ken Church, AT&T Research Labs

October 21, 2003

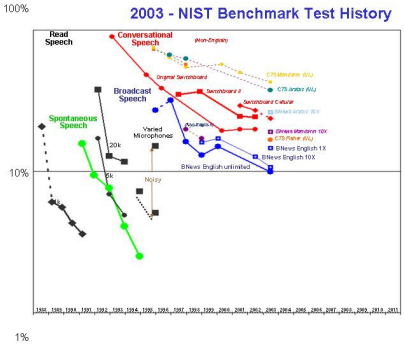


View Seminar Video

Don't record predictions



The Disk Space Conjecture



- Improvements in Speech, Language (& more)
  - are indexed to improvements in disk capacities
  - because falling disk prices → larger corpora → more training data
- 2003 Prediction:
  - Disks improve 1000x per decade → Counts increase 1000x per decade
  - 1TB: $1k (2003) → $1 (2013)
  - ***Missed by 30x (a TB is currently ~ $30 >> $1)***

Toshiba - Canvio 3TB External USB 3.0 Hard Drive - Black

Model: HDWC130XK3J1 **SKU:** 6914041

USB 3.0 interface, backward compatible with USB 2.0; NTI Backup Now EZ software; data transfer rates up to 5 Gbps; 32MB cache buffer

Customer Reviews: 4.2 of 5 (241 reviews)

Check Shipping & Availability ›

Compare

Sale: $104.99

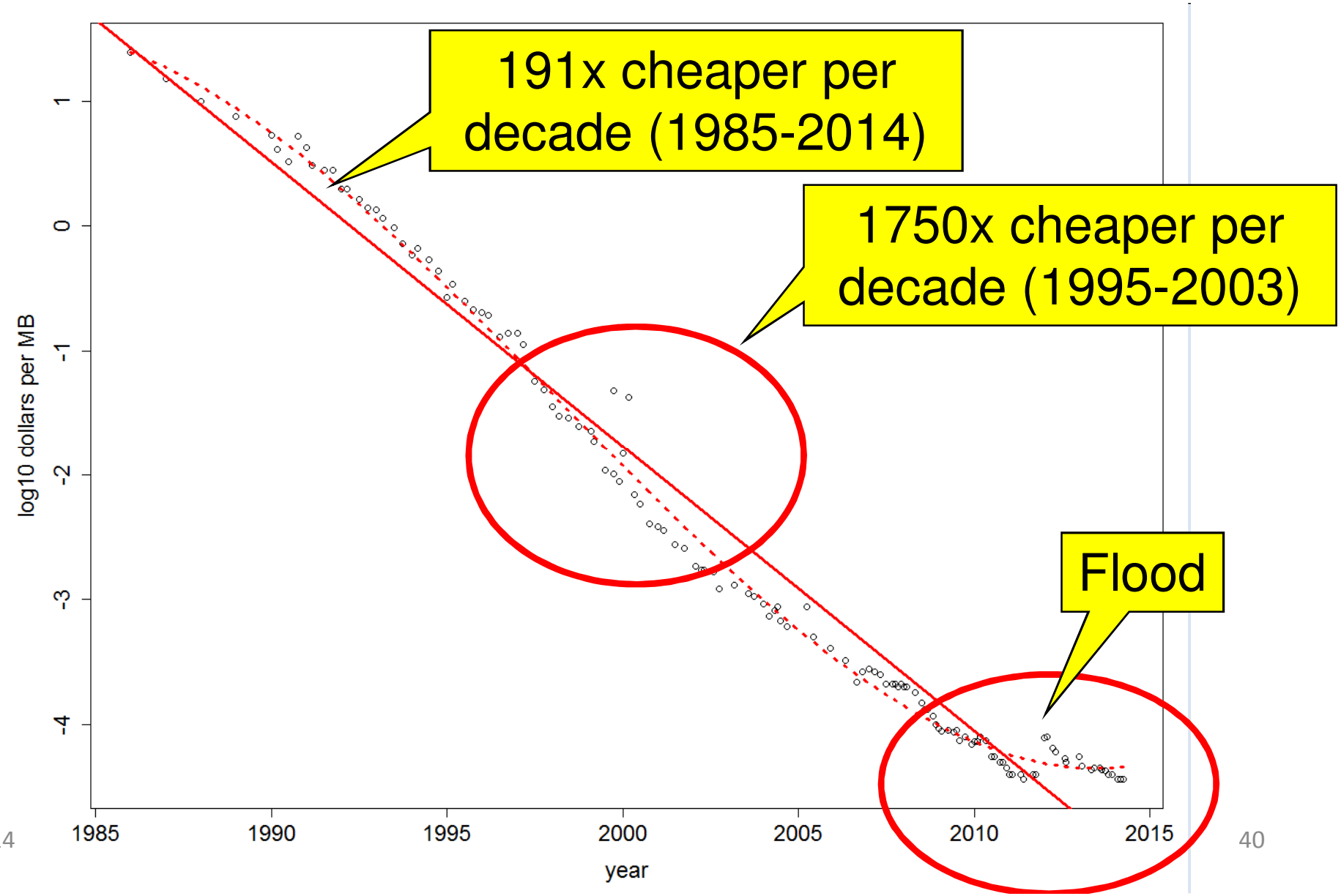
Reg. Price: $119.99
You Save: $15.00

- Free Shipping
- Get up to 6% Back in Rewards: See How

Add to Cart

Disk Prices Over 30 Years

http://www.jcmit.com/diskprice.htm



Speech and Language Processing: Where have we been and where are we going?

That's my story (and I'm sticking to it)

Consistent Progress over Decades

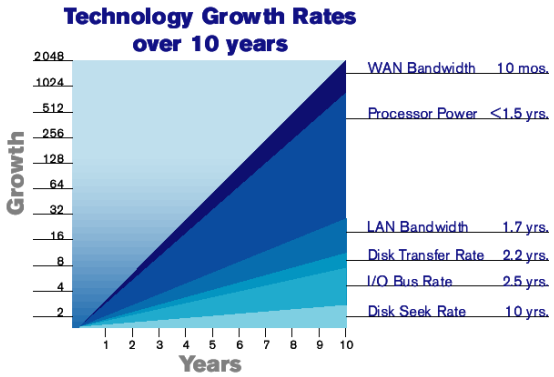
Kenneth Ward Church

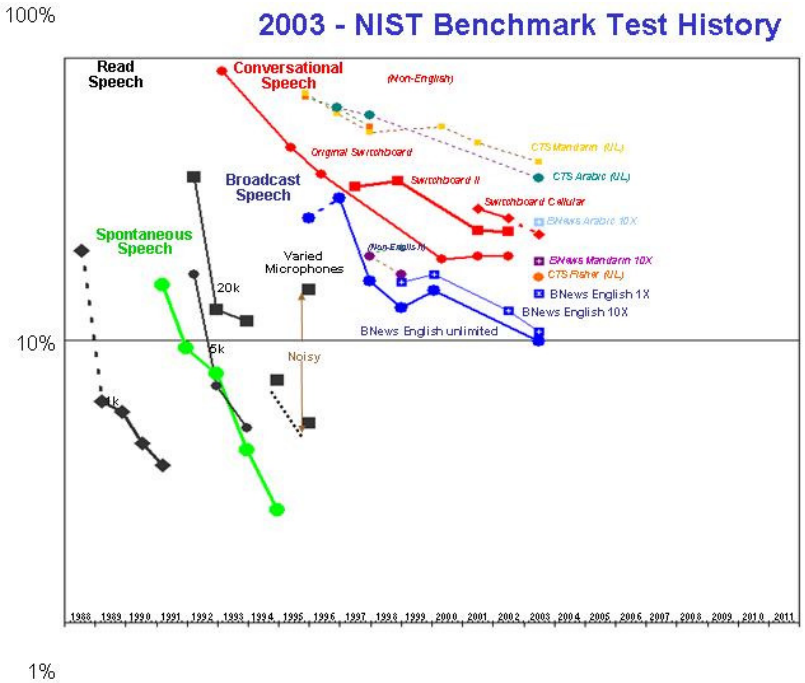
AT&T Labs-Research

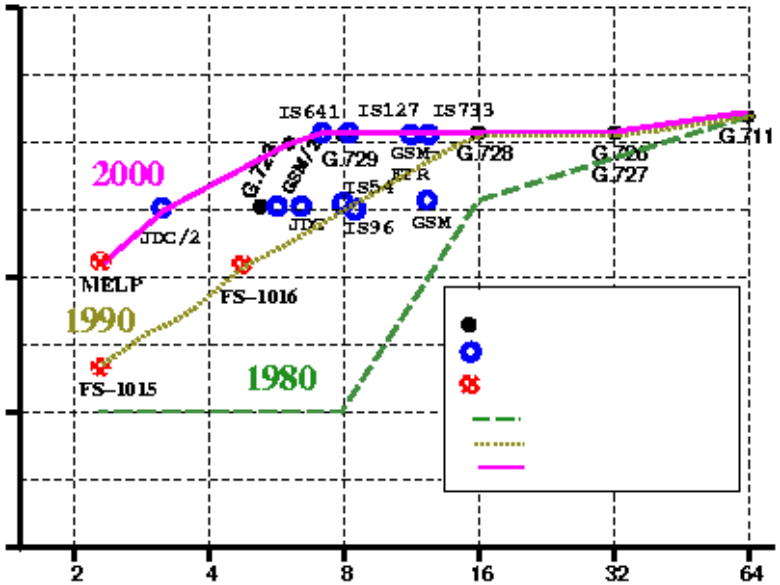
chu

www.research.att.com/~kwc

No Breakthroughs







Conclusions

- Fads come and fads go,
  - but seminal papers such as “Case for Case” (C4C)
  - are here to stay.
- As mentioned above,
  - we should train the next generation with the technical engineering skills to take advantage of the opportunities,
- but more importantly,
  - we should encourage the next generation
  - to read seminal papers in a broad range of disciplines
- so they know about lots of interesting linguistic patterns
  - that will, hopefully, show up
  - in the output of their machine learning systems.