Learning
From Machine Learning Errors

Mark Liberman
University of Pennsylvania

Mathematical modeling of human errors has long played an important role in theories of human perception; and human interpretation of machine errors has always played a central role in guiding algorithmic improvements.

This talk will feature three simple applications of human-machine interaction in predicting, identifying and responding to errors:

- eliminating untrustworthy machine output from research datasets;
- improving productivity and quality in semi-automatic annotation by better management of the human/machine division of labor;
- and adjusting machine output towards human-annotation norms.

None of these applications are new, but recent changes create new opportunities for developing and deploying them.

Three narratives:

1. Eliminating untrustworthy machine output
   from research datasets

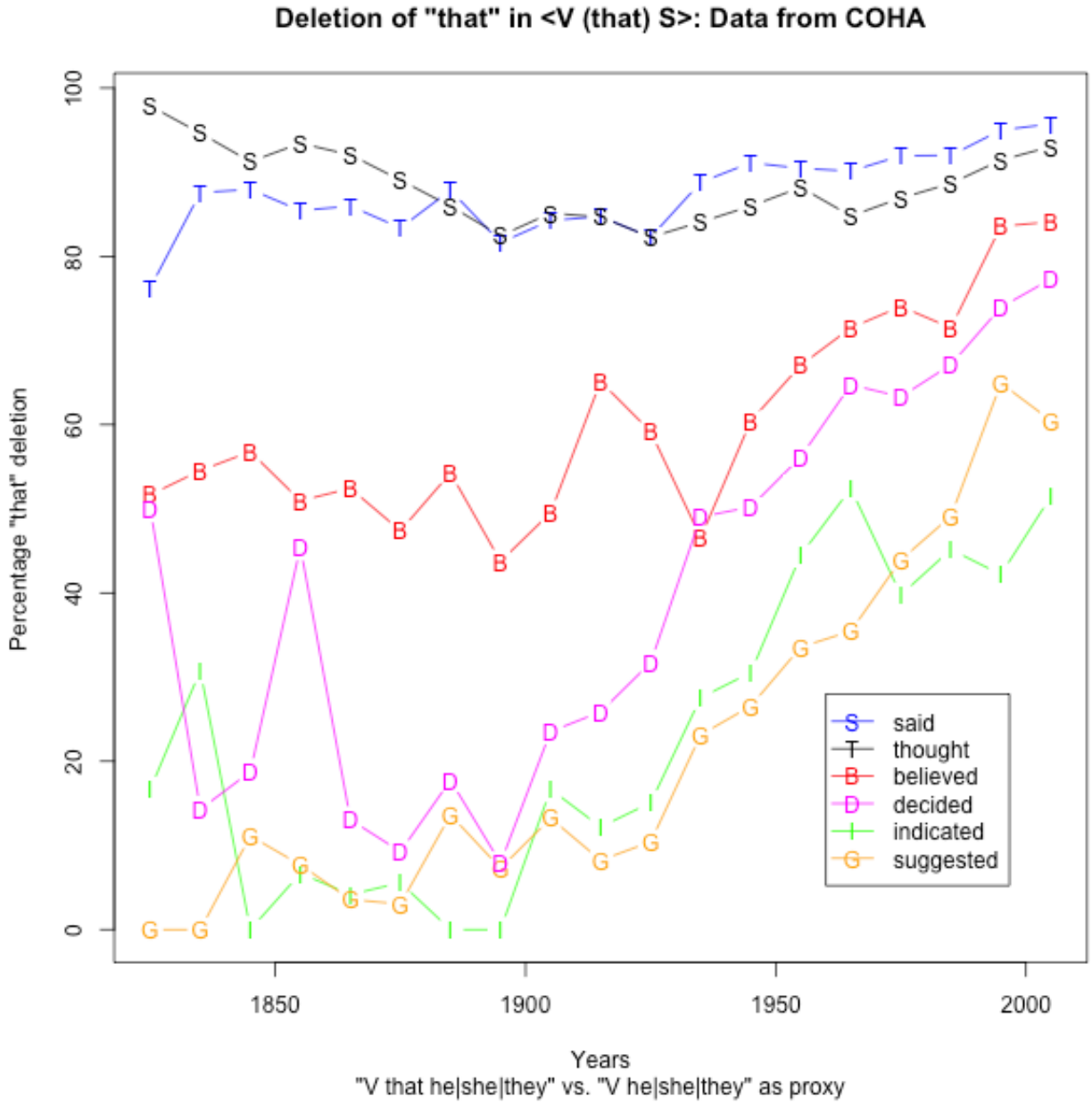
2. Improving productivity and quality
   in semi-automatic annotation
   by better management
   of the human/machine division of labor;

3. Adjusting machine output towards human-annotation norms.

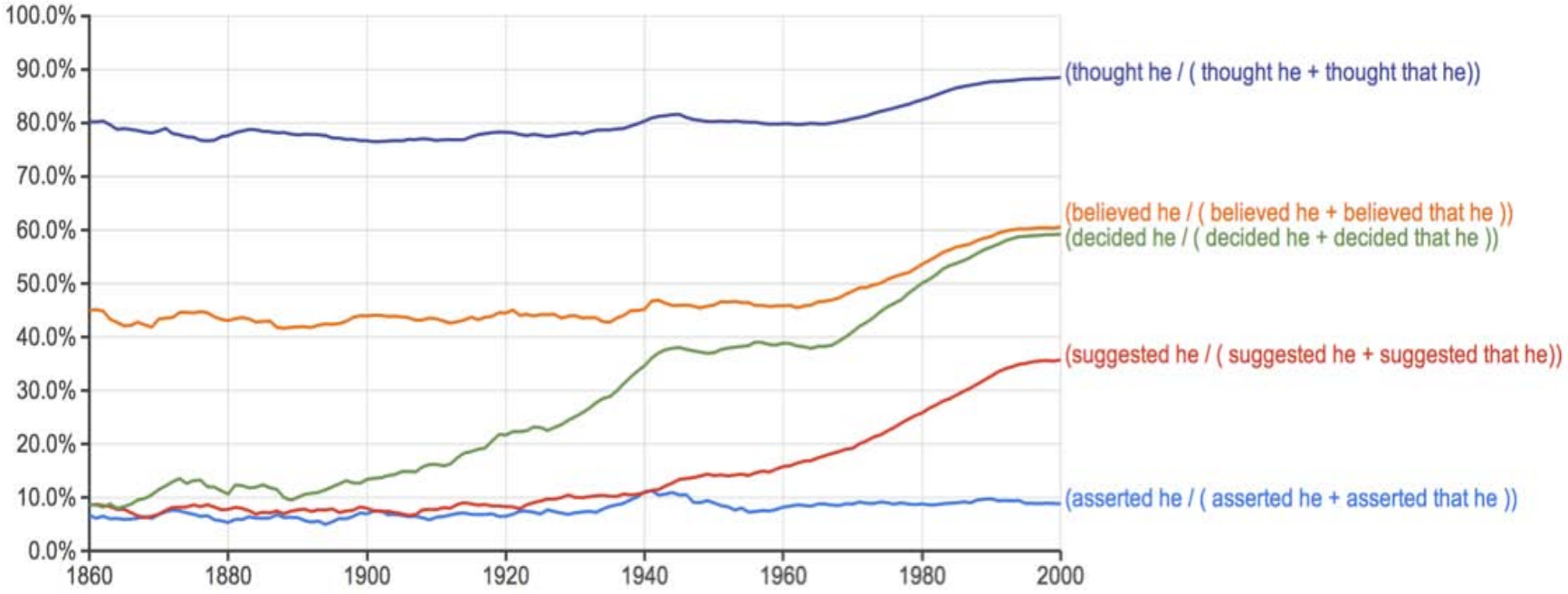
Context: Studies of historical syntax

Why a few million words is not (always) enough…

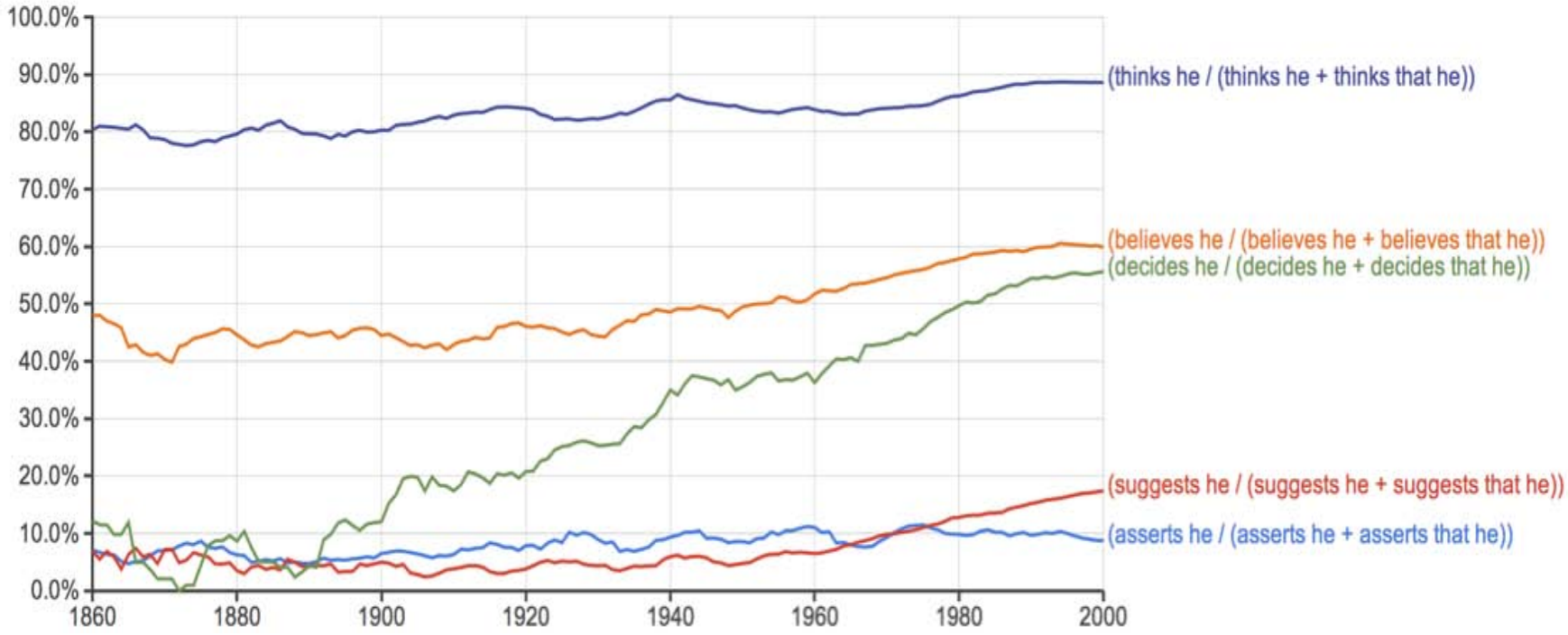
Example #1: V (that) S


Deletion of "that" in <V (that) S>: Data from COHA
Percentage "that" deletion
0
20
40
60
80
100
1850
1900
1950
2000
Years
"V that he|she|they" vs. "V he|she|they" as proxy
said
thought
believed
decided
indicated
suggested

A similar pattern in Google Books dataset:



And also for VBZ inflected forms:



COHA Counts for “suggested (that) he|she|they”:

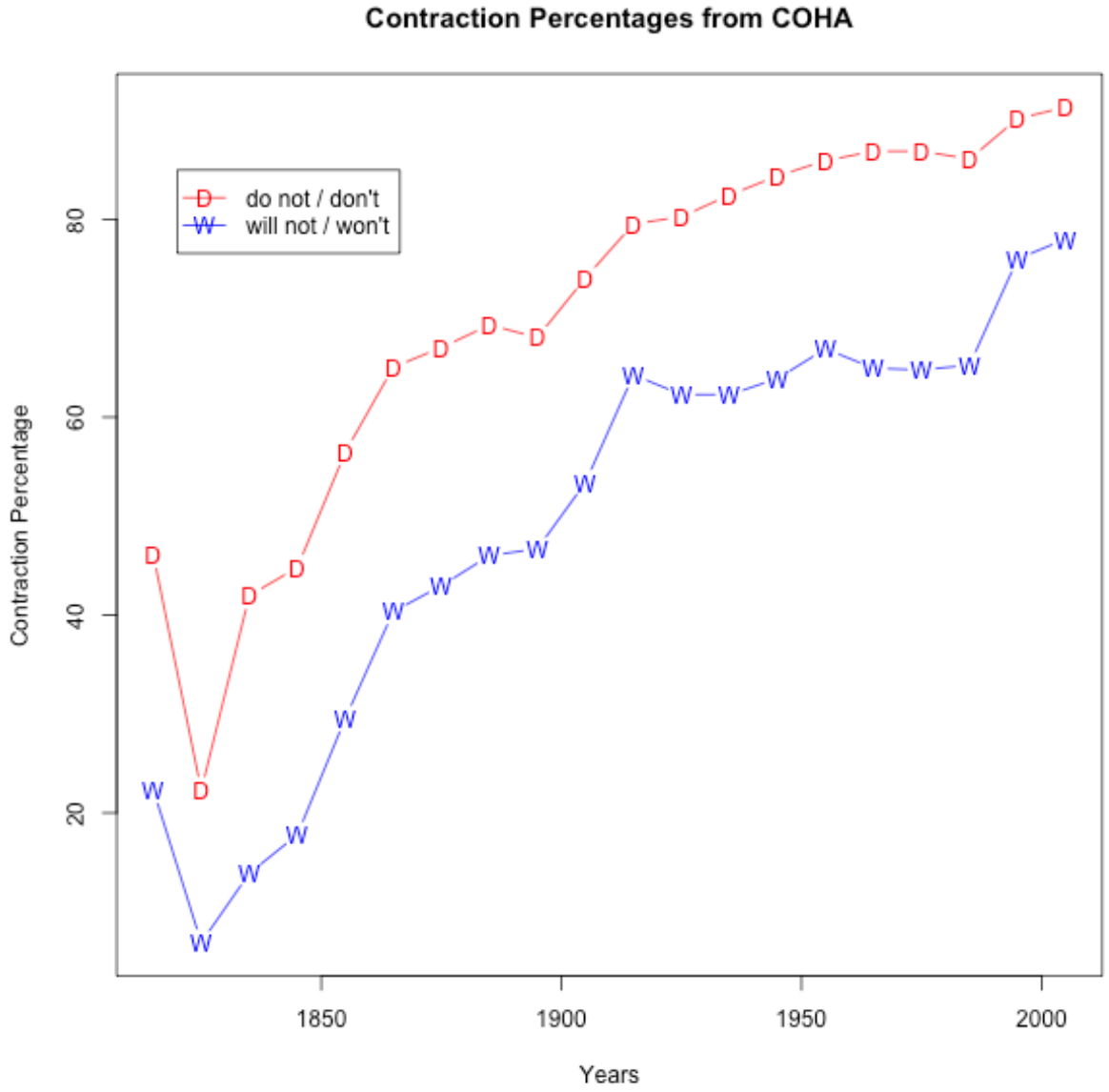
DECADE	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
DEL	0	0	1	1	1	1	6	4	9	7	10	25	30	47	39	58	83	87	101
NO DEL	1	4	8	12	27	32	38	51	59	78	86	83	83	93	71	74	86	47	66
WORDS (M)	6.9	13.8	16	16.5	17.1	18.6	20.9	21.2	22.5	22.7	25.6	24.4	24.1	24.4	23.9	23.8	25.2	27.9	29.5

Conclusion:

For this investigation,

~20 million words per decade is marginal.

Example #2: Contraction of *will not* and *do not*


Contraction Percentages from COHA
do not / don't
will not / won't
Contraction Percentage
20
40
60
80
1850
1900
1950
2000
Years

How about contractions in a set of sources from last week?

Specifically, real-estate listings from trulia.com, e.g.

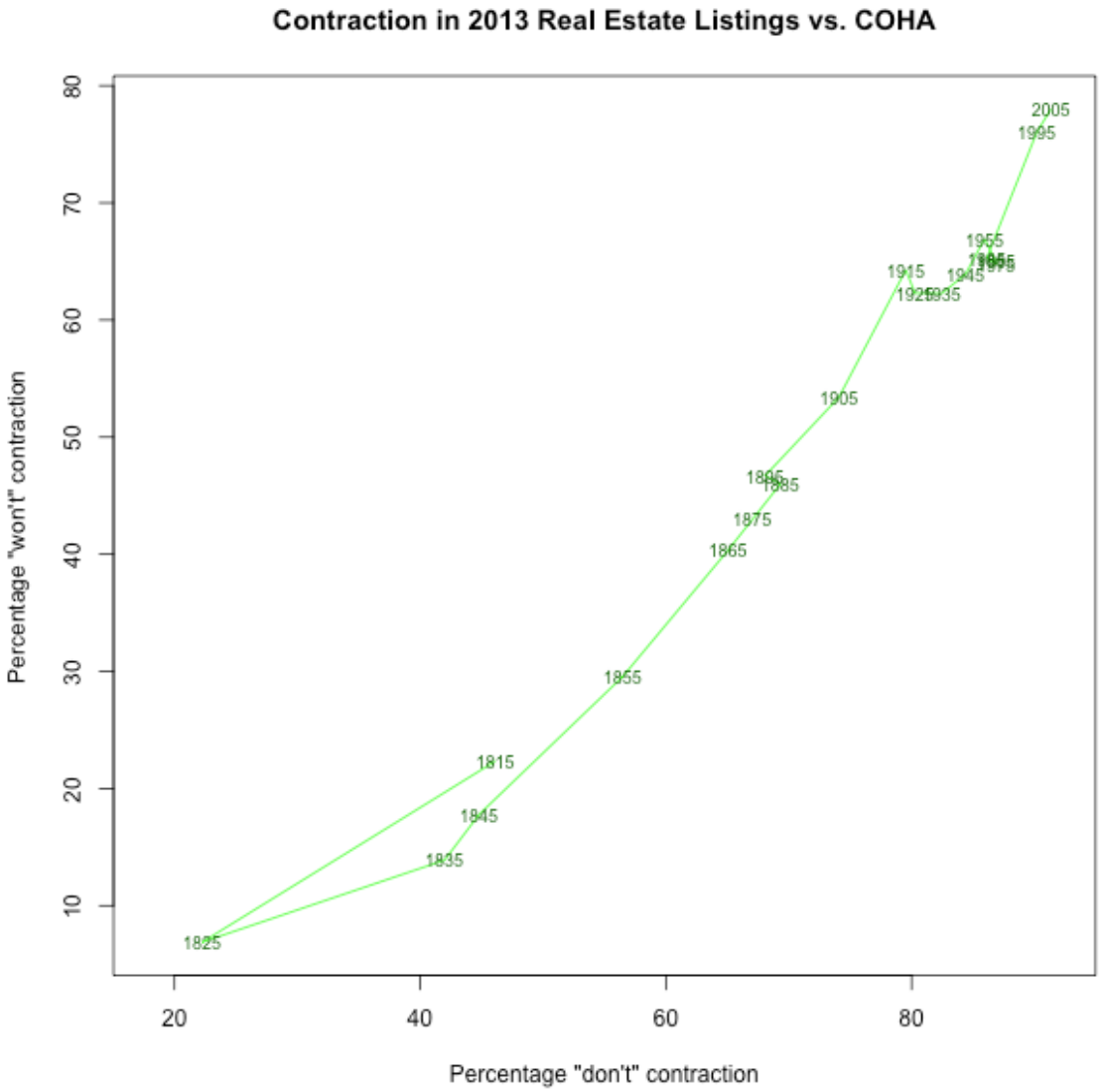
You **will not** want to miss this wonderful home in sought after Martin Manor.
Classic 1920's Brick Bungalow in Historic West End with energy features that **will not** drain your pockets!
Seller **will not** turn on utilities for inspections.

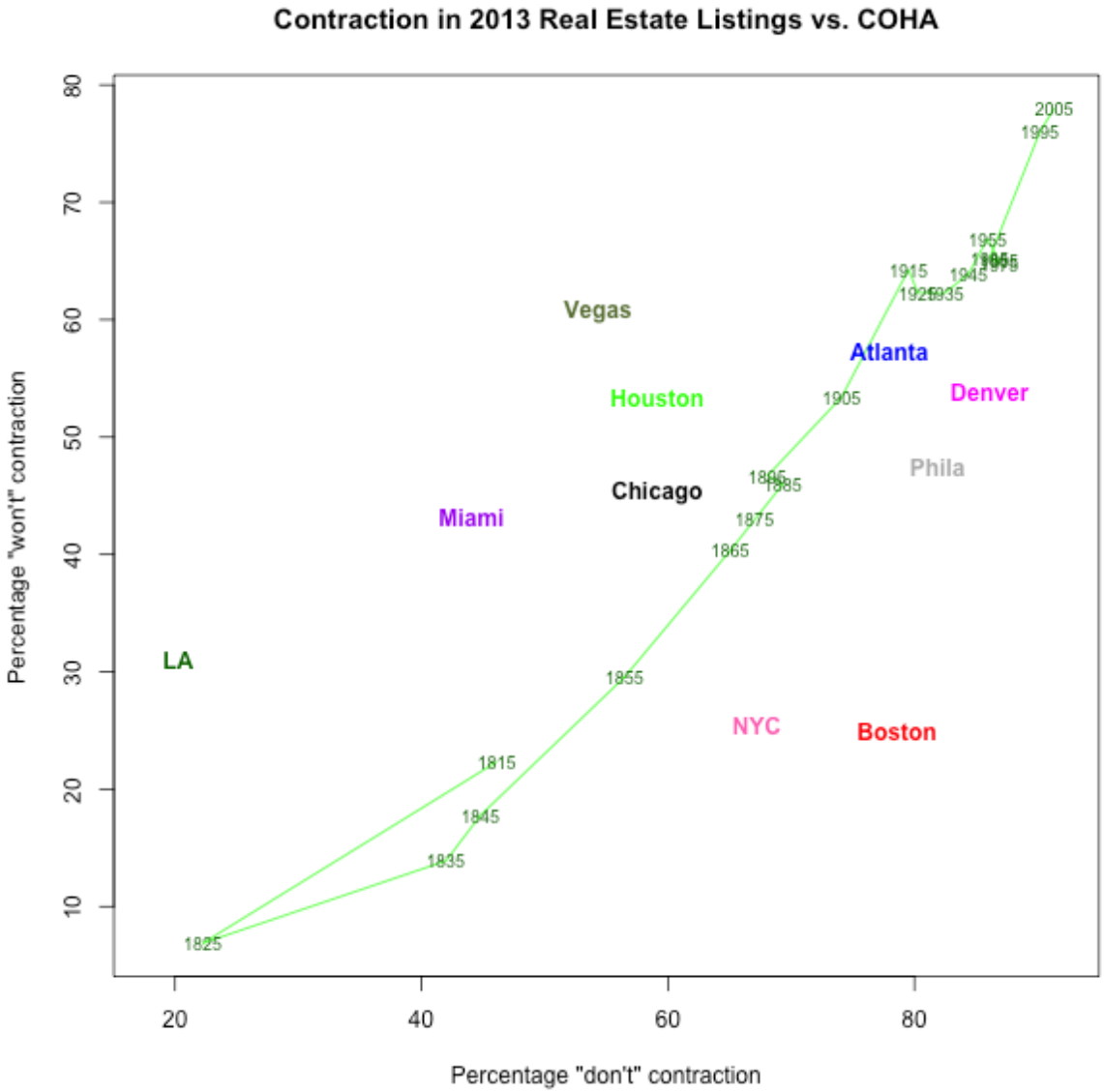
Great price, **do not** miss!
Please **do not** enter the property site without an appointment.
… the master closet has the laundry room, which most units in Foxcroft **do not** have!

Hurry! This one **won't** last long!
You **won't** find a street like this anywhere in Buckhead!
Don't wait. An investment you **won't** regret.

Completion May 2013, but **don'**t wait so builder can customize.
You Will Hate Yourself For The Rest Of Your Life If You **Don't** Buy This Home!
We **don't** work with multiple offers and the buyer must be prepared to wait until bank approval.

10 Cities: Atlanta, Boston, Chicago, Denver, Houston, L.A., Miami, N.Y.C., Philadelphia, Las Vegas


Contraction in 2013 Real Estate Listings vs. COHA
Percentage "won't" contraction
Percentage "don't" contraction
10
20
30
40
50
60
70
80
20
40
60
80
2005
1995
1955
1915
1945
1935
1905
1885
1875
1865
1855
1815
1845
1835
1825


Contraction in 2013 Real Estate Listings vs. COHA
Percentage "won't" contraction
Percentage "don't" contraction
10
20
30
40
50
60
70
80
20
40
60
80
2005
1995
1955
1915
1945
1935
Vegas
Atlanta
Houston
1905
Denver
Phila
1885
Chicago
Miami
1875
1865
LA
1855
NYC
Boston
1815
1845
1835
1825

Contraction counts from trulia.com real estate listings:

	Atlanta	Boston	Chicago	Denver	Houston	LA	Miami	NYC	Phila	Vegas
do not	26	6	92	16	170	374	575	100	74	21
don't	92	22	133	99	246	95	453	205	334	25
will not	38	15	67	47	133	98	323	161	144	18
won't	51	5	56	55	152	44	245	55	130	28
TOTAL WRDS	248K	96K	571K	223K	733K	421K	1.4M	1.7M	754K	147K

Conclusions:

100k words per source is marginal for estimating source effect here
10 sources are not enough to get a stable estimate of the overall pattern

Size of some available collections For studying the history of English

Penn-Helsinki Parsed Corpus of Early Modern English:

1.7 million words from 448 texts over 210 years (1500-1710)

Curated (text & metadata), annotated, published

Early English Books Online / Text Creation Partnership (EEBO-TCP):

125,000 texts over ~220 years (1483-1700)

Partly curated (40,000 done), not annotated, not published yet (though accessible online)

"Phase I" (25,363 texts) to be made available in 2015

"Phase II" (45k more texts) to be made available ~ 2017+

Corpus of Historical American English:

400 mw from 100k texts over 200 years (1810-2009)

Semi-curated, semi-annotated, not published (and will not be)

Eighteenth Century Collections Online (ECCO)

200k texts over 100 years (~1700-1800)

Curated, not annotated, release prospects unclear (to me?)

Hathi Trust:

>10M texts over ~400 years (~31% public domain)

Lightly curated, not annotated, not published (but some can be downloaded)

Internet Archive:

??? Texts(12k on line) over ~400 years

Not curated, not annotated, can be downloaded

. . . etc. . . .

- For Old English, Classical Latin, etc.,
  the extant text is limited
  and all of it is available for historical study
- But for English since ~1500,
  and for many other languages,
  there are 100s or 1000s of books per year
  now available in digital form

So what's the problem?

1. Bad OCR (optical character recognition)
2. Problematic metadata
   (editions, genres, authors)
3. Lack of annotation
   - Headings, captions, marginalia, ...
   - Quotations, dialogue, other languages, …
   - Lemmatization, tagging, parsing, . . .

Old-book OCR is usually bad OCR…

The Internet Archive's text of Henry Caner, *A candid examination of Dr. Mayhew's Observations on the charter and conduct of the Society for the Propagation of the Gospel in Foreign Parts* , 1763

(S)

Thus in page 55", he fays '' the Society hare ma-
*^ nifefted a fufficient forwardnefs to encourage an4
" increafe fmall difaffe<5ted parties in our towns,
" upon an application to theni.'' And in the 5'7th
page he reprefents the Society as hpping that thefe
miali parties will by their influence gradually bring
on a general fubmiffion to an epifcopal fovereign ; and
affirms that this has long been the formal deGgi^
of the Society, and is the true plan and grand
myftery of their operations in New-England."
In his 1 06th page he tells us that the " affair of
Bifhops in America, has been a favourite obje^
with the Society," and in the next page, that
the Society fpare neither endeavours, applications,
nor expence, in order to effe6l their grand defign
" of epifcopizing all New-England," and a few FinesJ
further, " The Society have long had 2, formal deftgn
" to diffolve and root out all our New-England
" churches. — ^I^his (he fays) fully and clearly ac-*
" counts for their being fo ready to encourage fmall
*' epifcopal parties all over New-England, by fend-
'* ing them miflionarics."

Choice of editions and sources:

Thomas Jefferson, *Notes on the State of Virginia*

Written in 1781, updated and enlarged in 1782 and 1783;
printed anonymously in Paris in 1785;
first public edition in 1787 in London.

Hathi Trust:
9 versions, published 1801, 1802, 1803, 1829, 1832 (2), 1853 (3), 1894.

Internet Archive:
12 versions, published 1787, 1801 (3), 1802, 1803, 1829, 1832 (2), 1853, 1955 (2)

Images of the original mss. online at the Massachusetts Historical Society.

Careful e-text version of 1787 edition at UVa Electronic Text Center

Hathi / IA OCR:

Q^ U E R Y I.

AN exafl defcriptlon of the limits andLimks, boundaries of the ftate of Virginia ?
Virginia is bounded on the Eaft by the Atlantic: ,on the North by a line of latitude, eroding the Eaftern Shore through Watkins's Point, being about 37°. 57'. North latitude j from thence by a ftreight line to Cinquac, near the mouth of Patowmaci thence by the Patowmac, which is common to Virginia and Maryland, to the firft fountain of its northern branch ; thence by a meridian line, pafTing through that fountain till it interfed^s a line running Eaft and Wefl:, in latitude 29 . 43^ 42.4'' which divides Maryland from Penn-fylvania, and which v/as marked by MefTrs, Mafon and Dixon; thence by that line, and a continuation of it weftwardly to the com-pletion of five degrees of longitude from the eaftern boundary of Pennfylvania, in the fame latitude, and thence by a meridian line to the Ohio : On the Weft: by the Ohio and Miflifipi, to latitude 2,^". 30^. North : and on the South by the line of latitude laft-
B men-

UVa Electronic Text Center version:

"Boundaries of Virginia"
An exact description of the limits and boundaries of the state of Virginia.
Limits
Virginia is bounded on the East by the Atlantic: on the North by a line of latitude, crossing the Eastern Shore through Watkins's Point, being about 37o.57' North latitude; from thence by a streight line to Cinquac, near the mouth of Patowmac; thence by the Patowmac, which is common to Virginia and Maryland, to the first fountain of its northern branch; thence by a meridian line, passing through that fountain till it intersects a line running East and West, in latitude 39o.43'.42.4" which divides Maryland from Pennsylvania, and which was marked by Messrs. Mason and Dixon; thence by that line, and a continuation of it westwardly to the completion of five degrees of longitude from the eastern boundary of Pennsylvania, in the same latitude, and thence by a meridian line to the Ohio: On the West by the Ohio and Missisipi, to latitude 36o.30'. North: and on the South by the line of latitude last-mentioned.

Opportunity for System / Version Combination:

THE following Notes were written in Virginia in the year 178 1, and Ibme-what corrected and enlarged in the winter of 1782, in anfwer to Queries propofed to the Author, by a Foreigner of Diftindion, then refiding among us. The fubjefls are all treated imperfeflly 5 fome fcarcely touched on. To' apologize for this by developing the circumflances of the time and place of their compofition, would be to open wounds which have already bled enough. To thefe circumftances fome of their imperfedlions may with truth be afcribed ; the great mafs to the want of information and want of ta-lents in the writer. He had a few copies printed, which he gave among his friends ; and a tranllation of them has been lately publilhed in France, but with fuch altera-tions as the laws of the prefs in that country rendered neceffary. They are now offered to the public in their original form and language.

TH E following Notes were written in Virginia in the year 1781, and somewhat corrected and enlarged in the winter of 1782, in answer to Queries proposed to the Author, by a Foreigner of Distinction, then residing among us. The subjects are all treated impersectly; some scarcely touched on. To apologize for this by developing the circumstances of the time apd place of their composition, would be to open wounds which have already bled enough. To these circumstances some of their impersections may with truth be ascribed; the great mass to the want of information and want of talents in the writer. He had a few copies printed, which he gave among his friends: and a translation of them has been lately published in France, but with such alterations as the laws of the press in that country rendered necessary. They are now offered to the public in their original form and language.

What we need

Organized effort to

- Select
  - texts
  - sources
- Correct
  - metadata
  - texts
- Annotate
  - text structure
  - linguistic structure

Luckily,

- Lots of texts in decent shape already exist (EEBO, ECCO, various smaller collections)
- Although OCR for older books sucks,

  It can be improved by better font training and better language models!
- Tagging, parsing etc. are good and improving,

  and there are ideas for making them MUCH better!
- Crowdsourcing often works
- There are other applications and customers for the improvements we need

OCR improvements

It's easy to make BIG improvements in OCR for older texts,
using adaptive language modeling techniques
from ASR and similar areas,
as well as system/version combination techniques.

. . . if you're interested in this
and you know something
about language modeling
and machine learning –
-- or you know someone with skills and interest --

Please contact me!

But meanwhile . . .

Language modeling can distinguish
decent OCR from terrible OCR

… and flush the terrible-est OCR
from our datasets.

This is a new application for language modeling…

1,000 books/year
for 250 years
is 250,000 books

= 7% of Hathi Trust public-domain books

We could eliminate the worst 93% and still have what we need.

1. Eliminating untrustworthy machine output
   from research datasets

2. Improving productivity and quality
   in semi-automatic annotation
   by better management
   of the human/machine division of labor;

3. Adjusting machine output towards human-annotation norms.

“Tree Banking”: Creating parsed text corpora
(with some semantic annotations as well)

Basis for progress in (stochastic) parsing algorithms

Penn [English] Treebank (1993) was an important step

In recent years, parsing is increasingly seen as foundational for

machine translation
information retrieval
information extraction from text

On-going production of treebanks

in new languages
in new genres and registers

BUT . . .

Tree Banking is:

hard to learn

3-6 months of training

only some people ever “get it” (~5%?)

slow even after training

Penn Treebank productivity was ~ 100 words/hour

hard to develop good standards for

Penn Treebank took three tries

Arabic Treebank required a complete revision

Some Historical Tree Banks:

1. The York-Toronto-Helsinki Parsed Corpus of Old English Prose (1.5 MW)
   Authors: Ann Taylor, Anthony Warner, Susan Pintzuk & Frank Beths
2. The Penn-Helsinki Parsed Corpus of Middle English, Second Edition (1.2 MW)
   Authors: Anthony Kroch & Ann Taylor
3. The Penn-Helsinki Parsed Corpus of Early Modern English (1.7 MW)
   Authors: Anthony Kroch, Beatrice Santorini & Ariel Diertani
4. The York-Helsinki Corpus of Early English Correspondence (2.2 MW / 4,970 letters)
   Authors: Ann Taylor, Anthony Warner, Susan Pintzuk, Arja Nurmi, Terttu Nevalainen
5. The Penn Parsed Corpus of Modern British English (1 MW)
   Authors: Antony Kroch, Beatrice Santorini and Ariel Diertani

In this area, the Digital Humanities
(represented by historical treebanks)
are ahead of the engineers.

Factors include:

on-going commitment of dedicated people (>20 years)

not much money

Results:

higher-quality annotation

100 times better productivity!

Thus for the cost of the original (1 MW) Penn Treebank
we could have a (better-quality) 100 MW Treebank

How?

So far:

1. Focused pre-annotation (before automatic parsing)
   Avoids common parser errors – and reduces associated errors as well
2. Batch post-editing of parser output via query-replace tree transducer
   Rapid correction of patterns of error

Some current pre-annotations:

CC (coordinating conjunction) is tagged for what it conjoins
punctuation is treated as CC where appropriate
"that" is tagged to distinguish complement-clause from relative-clause uses:
the fact that/C he learned the answer
the new fact that/WPRO he learned yesterday
"silent that" is indicated and similarly tagged:
the fact 0/that he learned the answer
the new fact 0/WPRO he learned yesterday
parenthetical phrases/clauses are explicitly tagged with virtual delimiters
(certain) PPs are tagged for what they modify

From Beatrice Santorini, a few examples of how she uses batch post-editing ("revision queries"). These are used both to correct parser output and to fix inconsistencies in human annotation:

We distinguish various verb classes, notably ones where a postverbal noun phrase is annotated as the direct object of the verb ("I persuaded him to come to the party") vs. as the subject of a nonfinite clausal complement ("I expected him to come").
The annotation guidelines is not always intuitive (KEEP and LEAVE are particular headaches), so i have several revision queries that flag the two possible error types (objects done as complement subjects, and complement subjects done as objects).

The noun phrase objects of ditransitive verbs are marked as either direct or indirect, depending on the verb. Again, it is easy to forget which class a particular verb belongs to, but relatively easy to enforce consistency with the revision queries.

The attachment of particles preceding PPs is often tricky ("he went up into the attic"). Tt's not always possible to resolve the attachment issues completely automatically, but some subcases always go one or the other, so they can be fixed automatically, and the remainder can be flagged for human correction.

Result: Beatrice can now produce treebank annotations
at a rate of >15,000 words/hour !

(e.g. 1.5 MW of treebanked text in 100 hours of work)

with quality as good or better than before.

Better “error prediction” for pre-editing
and better post-edit propagation
should improve this even further –

A key fact:

Most pre-annotation should be easy for “ungrammarians” to learn to do:

Where do (certain) clauses end?
What do conjunctions conjoin?
What do (certain) prepositional phrases modify?

Even without pre-annotation, modern parsers are about 85-90% accurate.

We believe that no more than 3 human pre-annotations per sentence
(of a type that is cognitively simple and obvious
to any literate person who understands the sentence)
should be enough to ensure treebank-quality automatic parsing
(given a modest amount of batch post-editing
for quality and consistency checking).

1. Eliminating untrustworthy machine output
   from research datasets

2. Improving productivity and quality
   in semi-automatic annotation
   by better management
   of the human/machine division of labor;

3. Adjusting machine output towards human-annotation norms.

Goal: Accurate phone-level segmentation of speech

Applications:

- Speech synthesis
- Pronunciation modeling
- Research in phonetics, sociolinguistics, psychology of language, etc.

HMM “forced alignment” is robust but not very accurate

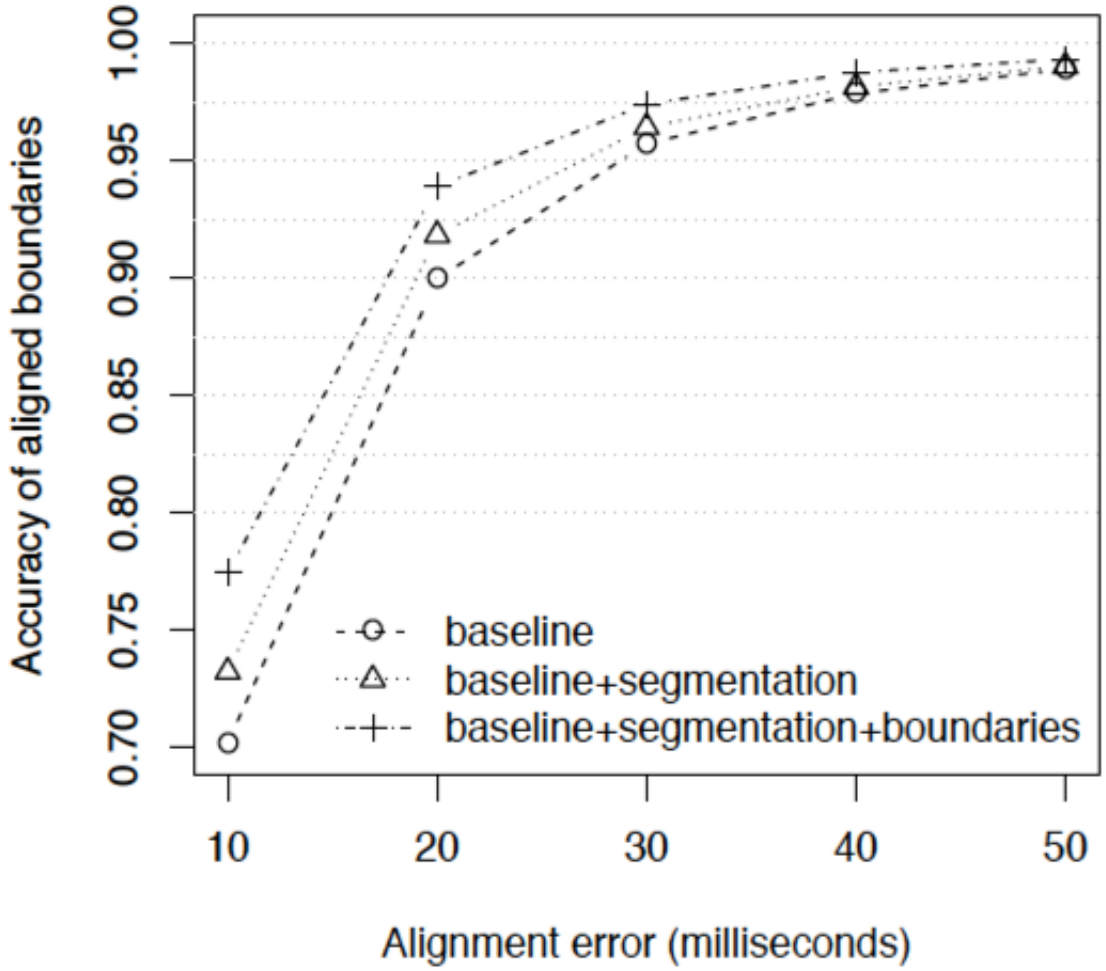
Various methods used to do it better:

- “boundary models”
- New acoustic features
- **“correction models”**
- System fusion

Jiahong Yuan, Neville Ryant, Mark Liberman, Andreas Stolcke, Vikramjit Mitra, Wen Wang, "Automatic Phonetic Segmentation using Boundary Models", *InterSpeech* 2013

Andreas Stolcke, Neville Ryant, Vikramjit Mitra, Jiahong Yuan, Wen Wang, and Mark Liberman, "Highly Accurate Phonetic Segmentation Using Boundary Correction Models and System Fusion", submitted *to ICASSP* 2014

Metric: Percent of boundaries within N msec. of human segmentation



Correction Model	Percent within 20 msec.
None (Window Start)	68.32%
Constant Offset (12.5 msec.)	91.02%
Two linear models: V/Glide vs. Other (*)	93.92%
Regression Tree	93.90%
Neural Network 1	94.17%
Neural Network 2	94.39%
With System Combination	96.77%

Note that improvement due to error-correction model (94.39-91.02 = 3.37%)
is large compared to system-combination improvement (96.77-94.39 = 2.36%)

or boundary-model improvement (not shown – 93.92-91.85 = 2.07%)

This is true even for the simple linear-regression model (93.92-91.02 = 2.90%)

Linear regression for error correction

The boundaries between vowel/glide phonemes are inherently subjective. […]

To compensate for this arbitrariness, we built a linear model to correct the forced alignment boundaries between vowel/glide phonemes. The model predicts manual boundary positions from the forced alignment positions of the two phonemes (phoneme center positions), the identities of the boundaries (the phonemes preceding and following the boundary), and the forced alignment boundary positions. The model was trained on the training data and applied to the test data.

For all other boundaries, the mean difference between manually labeled and forced alignment boundaries for every boundary identity was calculated using the training data, and the forced alignment boundaries in the test set were shifted by these boundary-dependent time differences

Conclusions:

Accurate estimation of error rates can be used to “purify” large datasets

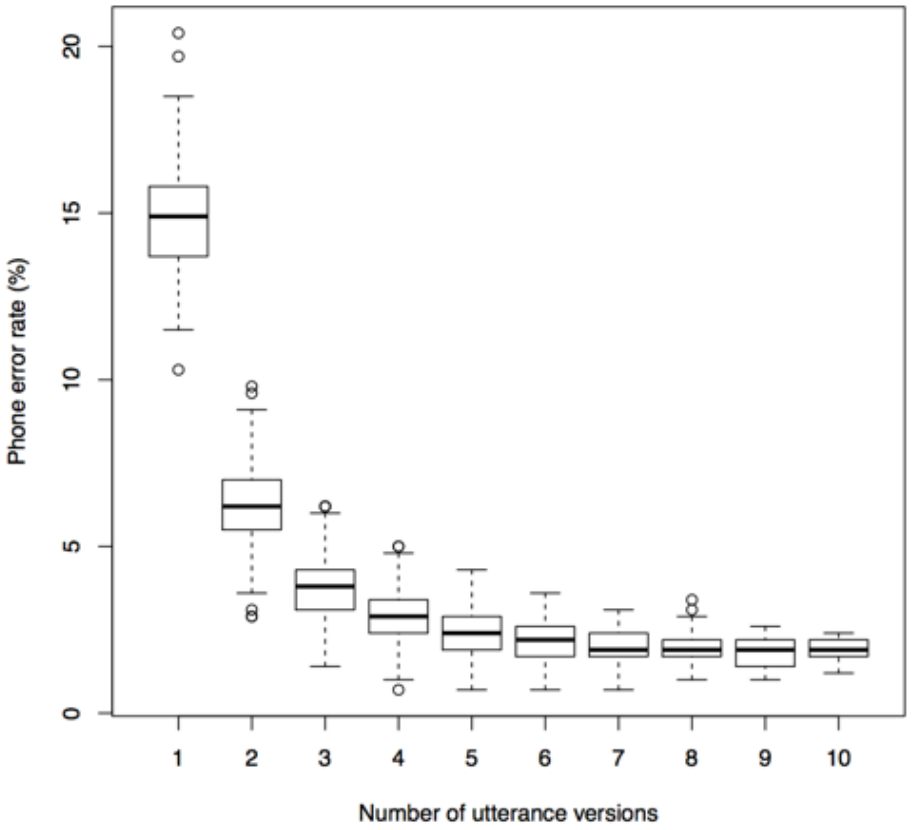
Accurate prediction of error locations can focus human annotation
for large improvements in productivity

Accurate estimation of error magnitudes can improve performance
by automatic correction

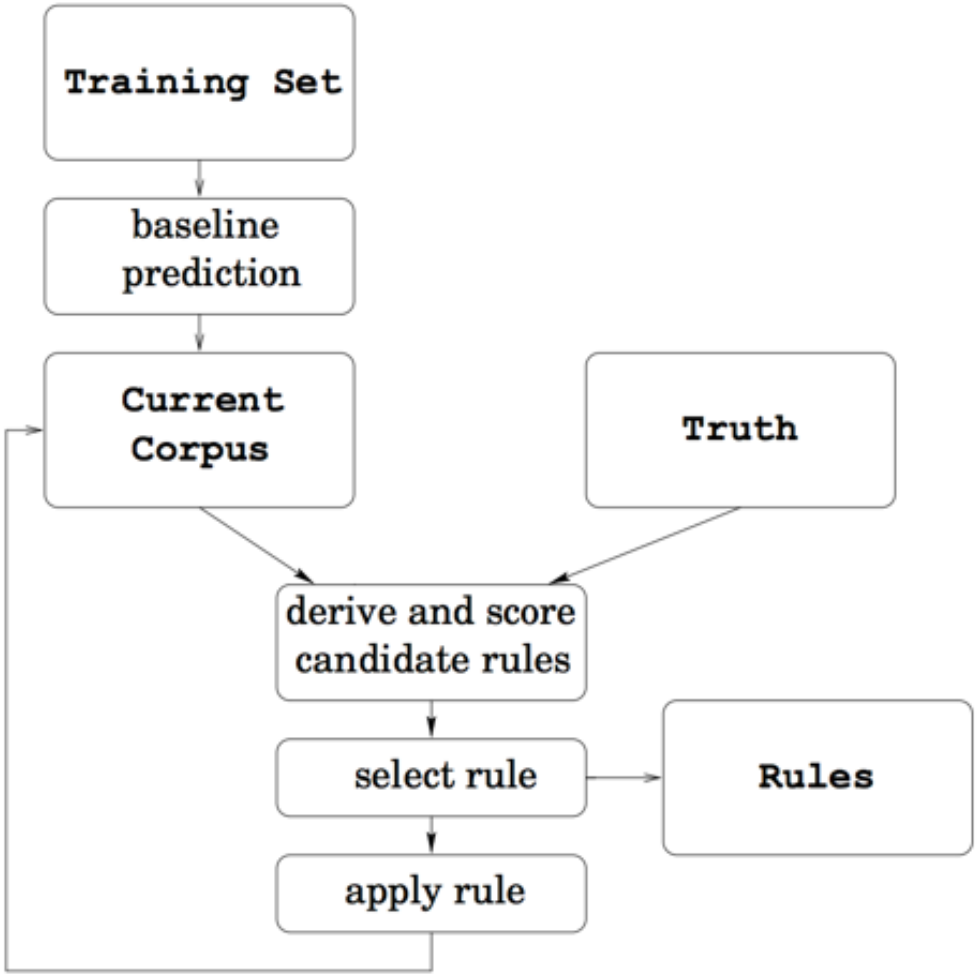
These are all old and obvious points,

but they have new and potentially large implications.

Thank you!



Mark Liberman, Jiahong Yuan, Andreas Stolcke, Wen Wang, and Vikramjit Mitra, "Using Multiple Versions of Speech Input in Phone Recognition", ICASSP 2013



Transformation-Based Learning