Challenges and Opportunities in the Analysis of Large Linguistic Datasets

Mark Liberman

University of Pennsylvania

http://ling.upenn.edu/~myl

Workshop "Linguistique & big data"
Télécom Paris Tech
30 Nov. 2017

The past 50 years
have seen enormous quantitative changes
in the efficiency and reproducibility
of speech and language research,
thanks to advances in digital technology.

The near future will bring even larger changes –
not only quantitative changes in productivity and scale ,
but also qualitative changes in the nature of our research,
enabled by new (semi-)automatic methods.

New sources of data
and new methods of automated analysis
are opening up vast new territories of linguistic research.

We can easily acquire and manage new sources of linguistic data
that are several orders of magnitude bigger than old ones.

Because new methods can do old tasks several orders of magnitude more efficiently,
it's increasingly easy to explore these new datasets in old ways.

We can also easily experiment with completely new approaches to analysis and modeling.

And these new methodologies are rapidly spreading
into all the fields that study speech, language, and communicative interaction,
from poetics, sociology, and politics to psychology and neuroscience.

But major challenges remain.

In some cases, the revolution in data and algorithms is simply incomplete:
There are kinds of data that are not generally available,
or not available at all.
And long before we get to the hypothetical automatic linguist,
there are many simple tasks where the state of the art is shockingly bad.

At the same time, as Human Language Technology gets better and better,
commercial success risks destroying the engines of research progress.

And paradoxically, there some are issues intrinsic to our new research methods
that create new problems at the same time that they solve old ones.

Challenges:

Important types of speech and language data are missing,
and filling some of the gaps requires careful coordinated effort

Unsupervised automatic language learning doesn't work at all,
and (partly) supervised automatic language learning
doesn't work well enough:

Commercial success may risk research failure.

And real-world language is not an orthogonally controlled experiment . . .

What about the opportunities?

They turn out to be pretty much the same as the challenges….

Challenge 1:
Hard-to-get data

Where does linguistic “big data” come from?

A digital shadow universe

increasingly mirrors real life
in flows and stores of bits.

Society is mostly about communication.

And most communication is text

(or talk, which is just text in fancy calligraphy)

. . . more and more often in digital form.

Simple properties of text

(like the words that make it up,
and the ways that they're performed)

are a good proxy for content.

Better than anything else we have, anyhow…

Bigger faster cheaper digital everything

(and better programming languages, and . . .)

make it easier and easier

to pull content out of the flows of text

in that digital shadow universe.

So in that new evolutionary niche:

a host of newly-evolving life forms
have got means, motive, and opportunity
to live off of these flows and stores of text

. . . while adding their digestion products
to the ecosystem.

From that digital ecosystem,
many kinds of text and speech
are easily collected and distributed.

In some cases,
there are intellectual property rights to be licensed,
but this is generally not hard to do.

In contrast, there are some kinds of datasets
where privacy and confidentiality
pose difficult ethical and legal problems,
especially for data sharing across sites.

For example:

Recordings of clinical interviews,
neuropsychological tests,
and similar things.

There are policies, laws, and ethical concerns
that require such recordings to be treated in a special way,
and are widely believed to make cross-site sharing impossible.

Why do we want such recordings for research,
and why do we want to share them?

Because speech and language are often a key behavioral marker,
cheaper and less invasive
than brain imaging, blood tests, or genomic tests,
but also often diagnostically more useful.

And more important, many (most?) relevant problems
are “phenotypically diverse”, in ways that matter –
meaning that we really don’t understand them very well.

With enough data and enough research,
we can hope to find the true latent dimensions
of the relevant behavioral space(s).

But a single site rarely has enough data,
and no single research team is likely to find the answers.

We need to pool data across sites,
and we need a community of researchers
working together to understand it.

Example: “Autism Spectrum Disorder”

It’s clear that Autism is not a “spectrum”, i.e. a single dimension,
but rather a space, with many dimensions –

It’s a space that we all live in,
with some corners that have been medicalized
because they cause serious life problems.

Is there suitable digital data Out There?

Yes –
for instance, the Autism Diagnostic Observation Schedule (ADOS)
is a standard diagnostic tool,
consisting of a multi-part structured interview
which is video recorded and scored from the video,
with a half a dozen scoring rubrics for of the ~12 segments.

For diagnosis, the multiple scores are added up and thresholded.

Order(1,000,000) ADOS recordings are Out There.

An ADOS recording DVD is stored in the patient's folder,
along with many other tests.

We've begun a collaboration
with the Center For Autism Research
at Children's Hospital of Philadelphia,
which has several thousand such recordings.

We selected an initial set of ~100 interviews,
including interviews with neurotypical controls
and with adolescents with other diagnoses such as ADHD.

We did some preliminary work
to persuade the hospital's Institutional Review Board
that it was both possible and worthwhile
to share 20-minute ADOS audio segments for research purposes

-- with appropriate safeguards.

The CAR clinicians contacted the parents and children involved
to get informed consent for sharing
anonymized audio and transcripts with other researchers,
where “anonymized” means that
personal names, addresses, institutional names etc.
are bleeped from the audio
and replaced by generic placeholders in the transcripts.

Nearly everyone agreed –
we ended up with 99 20-minute segments,
which should be published this year by the LDC.

Preliminary research on this small pilot corpus (~33 hours)
suggests that every sensible linguistic measurement
shows some interesting signal.

We hope to persuade other clinical centers
to join us in creating a much large collection.

As Bob Schultz, CAR's director, said:

"With ten thousand interviews,
maybe we can figure out what's really going on."

There are many other kinds of datasets
relevant for ASD research –

And many other possible targets for similar research,

for example, the many diverse varieties
of neurodegenerative disorders,
such as Frontotemporal Degeneration,
Parkinsonism, and Alzheimers.

We're working with Penn's Frontotemporal Dementia Center
on a dataset of picture-description recordings
from ~1000 patients and elderly controls.

And we're working with the Framingham Heart Study
on ~9,000 1-2 hour recordings of neuropsychological testing
from 5,267 subjects,
of whom there are 122 with MCI and 212 with dementia.

Formal aims of the FHS collaboration (from the proposal to the FHS Exec):

1: Generate "gold standard" transcripts of 8000+ existing voice recordings as well as those being acquired through on-going neuropsych testing of all FHS cohorts.
2: Analysis of the 8000+ existing voice recordings obtained between 2005-present using existing voice recognition and voice analysis software.
3: Build additional software to analyze the digital voice signals and generate novel cognitive metrics from latency and other behavioral characteristics.
4: For each neuropsychological tests as well as across tests, identify normative values for e-voice metrics stratified by age, education, sex, both individually and in combination (e.g., age x education; age x sex; age x education x sex)
5: Conduct factor and cluster analysis of e-cognitive metrics across neuropsychological tests to identify domain specific measures.
6: Determine e-voice metrics/profiles that differentiate between those with and without known AD risk factors, including but not limited to, ApoE, family history of dementia/AD, homocysteine, vascular risk factors (including metabolic), inflammatory markers.
7: Determine whether neuroimaging biomarkers are related to e-voice metrics/profiles.
8: Determine whether incident change in neuroimaging biomarkers and neuropsychological tests are related to e-cognitive profiles.
9: Determine whether e-voice metrics/profiles can differentiate participants who are low to high risk for dementia/AD.
10: Conduct data driven analyses to identify e-voice metrics/profiles predictive of AD endophenotypes and risk for dementia/AD, in isolation and in combination with other health, lifestyle, biomarkers and genetic risk factors.

Neither the FTD nor the FHS dataset can easily be shared at present,
though we have hopes for the FHS collection.

In general, there needs to be a painful cultural shift
in the biomedical research community,
where we can hope that the pressing need for reproducibility
will overcome researchers' proprietary attitudes towards data.

Challenge 2:

Inadequate Algorithms

Unsupervised (computational) language learning doesn't work at all.

As a result, most of the worlds languages and and language varieties are resource-poor.

We can build decent acoustic models with easily-collected speech data
(…though multiplied by the number of languages and local varieties it's still a big job,
and unsupervised methods don't work yet here either).

But getting adequate transcribed data for a language model
(much less enough annotated data for syntactic, semantic, and discourse analysis)
remains a massively labor-intensive process.

Even supervised automatic methods don’t work well enough:

ASR transcripts are generally not good enough
for most linguistic research.

General phonetic annotation given a orthographic transcript
is not good enough.

And automatic diarization (who spoke when)
is shockingly bad.

At this point we can probably leave the general ASR problem
to the big companies.

But accurate phonetic annotation
of orthographically-transcribed audio
is not a problem of much current interest to those companies,

And smaller research groups are making progress on it.

Similarly for diarization – for more on this,
see the reports from JSALT 2017.

Challenge 3:

Commercial success

Paradoxically, the commercial success of HLT
can threaten research progress –

-- especially on problems
where commercial cost/benefit analysis
doesn't motivate research investment,
or where current engineering orthodoxy
points in the wrong direction.

Some sources of public funding are starting to take the view
that Google, Microsoft, Facebook, Apple, Amazon, & IBM
have solved all the problems of Human Language Technology,
or are about to do so.

Over time, this risks removing public funding
from process of building “common task” research communities,
and the difficult problem of spreading that methodology
to new areas like clinical research.

Structure of the “Common Task” method

- A detailed task definition and “evaluation plan”
  developed in consultation with researchers
  and published as the first step in the project.
- Automatic evaluation software
  written and maintained by a neutral third party
  and published at the start of the project.
- **Shared data:**
  Training and “dev(elopment) test” data
  is published at start of project;
  “eval(uation) test” data is withheld
  for periodic public evaluations

This method was originally developed
for Human Language Technology,
where it's been strikingly successful –

And it's spread widely to other areas of engineering.

But it's still rare in science,
especially in clinical areas
where the situation is in some ways similar
to HLT research.

Even in when clinical data sharing has been mandated, other part of the structure are missing

e.g. the **Alzheimer's Disease Neuroimaging Initiative** (ADNI) organized in 2004 by NIH

A slide from a presentation by Neil Buckholtz
(National Institute on Aging)
"Transforming Research through Open Access to Discovery Inputs and Outputs"
at the Berlin 9 Open Access Conference, November 2011:

GOALS OF THE ADNI: LONGITUDINAL MULTI-SITE OBSERVATIONAL STUDY

- Major goal is collection of data and samples to establish a brain imaging, biomarker, and clinical database in order to identify the best markers for following disease progression and monitoring treatment response
- Determine the optimum methods for acquiring, processing, and distributing images and biomarkers in conjunction with clinical and neuropsychological data in a multi-site context
- "Validate" imaging and biomarker data by correlating with neuropsychological and clinical data.
- Rapid public access of *all* data and access to samples

BUT ADNI has

- No speech/language data
  (not in 2004, and not added later)
- No well-defined versioning of datasets
- No quantitative evaluation metric
- No focused workshops

Predicting the time course of Alzheimer's Disease
is exactly the kind of problem
("algorithmic analysis of the natural world")
for which the Common Task method has worked in the past.

We should apply such methods
to the large class of similar biomedical problems
(including those where speech and language are centrally involved)

Many scientists will be horrified
(just as speech and NLP engineers were in 1987-1992).
But in the face of the reproducibility crisis,
public funders need to force it to happen.

Challenge 4:

Real-world speech and language

Traditional instrumental phonetics

was usually based on recordings created by “subjects”

reading lists of artificial material in a laboratory setting

(often isolated words, words in a carrier phrase,

or somewhat strange sentences like

“She had my dark suit in greasy wash water all year”).

Advantage: a controlled setting, and to some extent

controlled variation of relevant factors.

Disadvantage: material is not “ecologically valid”,

and many factors are simply absent.

If somewhat more natural material was used
in traditional instrumental phonetics, like a read passage,
the size was necessarily limited.

Now, due to large amounts of available real-world data,
whether conversational or read,
and effective (semi-)automatic processing methods,
we can work with more natural material
at a scale three or four orders of magnitude larger.

Example – Studies of American English /l/:

Umeda 1977 (study of consonant duration for speech synthesis modeling):
20 minutes of essay reading by 1 speaker = **114** instances of /l/
> 80 hours of work just to make ~650 spectrograms

Sproat & Fujimura 1993 (categorical vs. gradient syllabification):
4 speakers x 17 contexts x 4 repetitions = < **272** instances of /l/
X-ray microbeam data

Yuan & Liberman 2009 (acoustic replication of Sproat & Fujimura):
Oral arguments from 2001 term of U.S. Supreme Court
~25.5 hours from 8 justices = **21,706** instances of /l/

LibriSpeech -- 5,832 audiobook chapters, 2,484 speakers, ~1,600 hours
1,567,796 instances of /l/

This is not just a numbers race!

Analysis of /l/ variation across contexts
has strong implications
for theories of English syllable structure
and theories of the nature of allophonic variation.

And we need large amounts of material
in order to control for non-orthogonal co-variation
across all levels of analysis.

Data quantity is not just for bragging rights!

Many linguistic dimensions interact:

- Language + Regional, social, and individual variation
- Register and style
  - Read speech vs. conversation; formality
  - Cultural patterns
- Syllabic and segmental context
- Phrasal structure and position
- Focus, emphasis, redundancy
- Emotion – arousal and valence
- . . .

And therefore quantity potentially turns into quality.

Example:

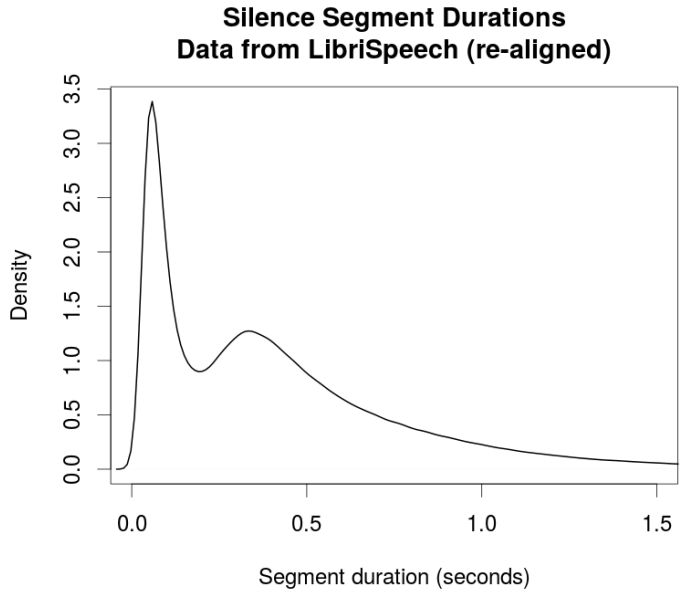
Prosodic differences

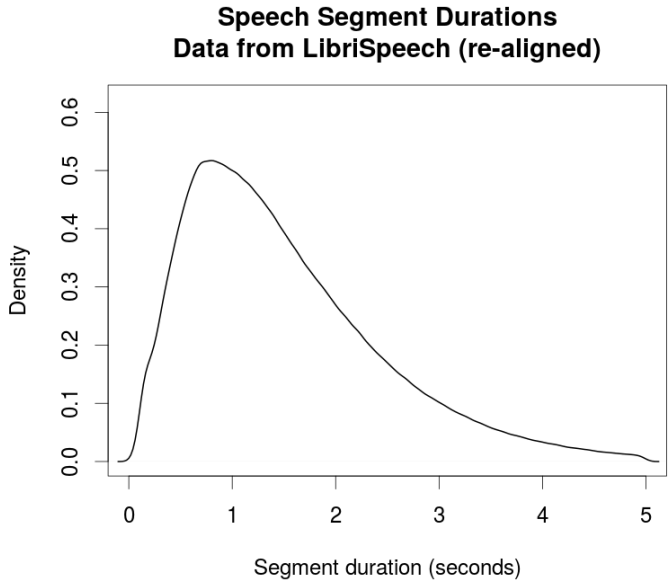
between reading and spontaneous speech

The LibriSpeech dataset consists of 5,832 English-language audiobook chapters read by 2,484 speakers, with a total duration of of nearly 1,600 hours.

[The LibriVox collection as a whole now has more than 50k hours of English – so from 20 minutes to 50,000 hours in 40 years…]

The overall distributions of silence-segment and speech-segment durations look like this:

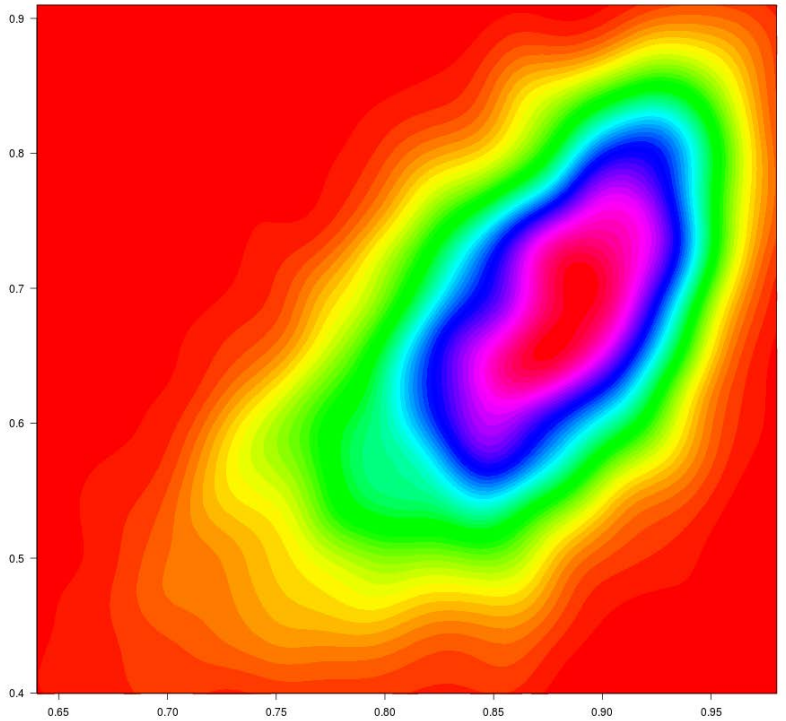




How to model variation among readers?

One simple way to map individual differences
is to plot the proportion of silence segments greater than X1, and the proportion of speech segments greater than X2.

With X1 = 200 msec. & X2 = 600 msec.,
the resulting 2D density plot looks like this:



The x-axis is
the proportion of silence segments > 200 ms.

The y-axis is
the proportion of speech segments > 600 ms.

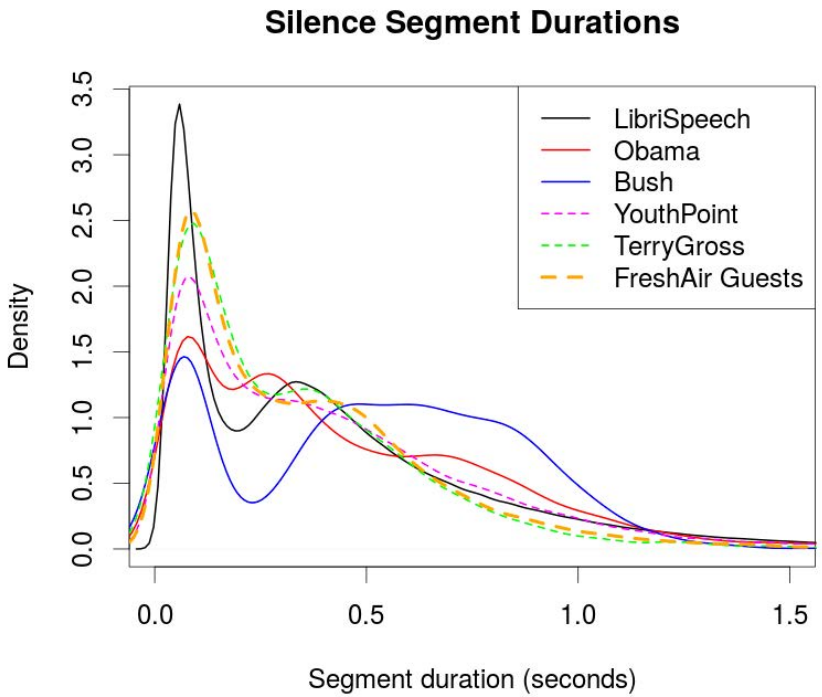
Does this distribution of speaker characteristics mean anything?

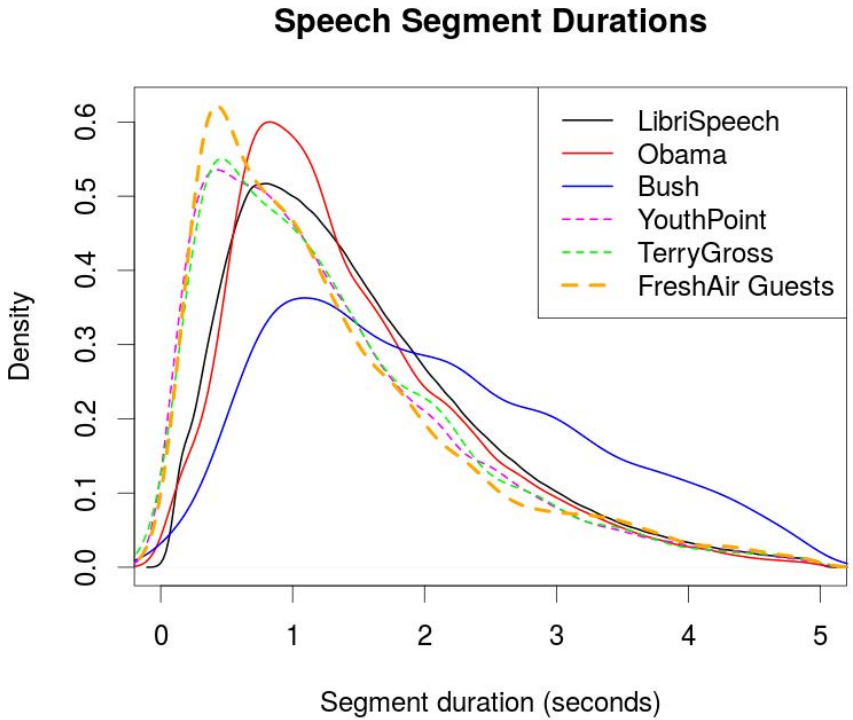
Let's compare some other sources of spontaneous and read speech.

Spontaneous: Fourteen *Fresh Air* radio interviews, involving public figures ranging from Lena Dunham to Stephen King to Gloria Steinem. The host Terry Gross is treated separately from the interviewees.

Spontaeous: A radio program produced by students at the University of Pennsylvania in the late 1970s. Our data set includes a subset of 50 sessions with 57 interviewees, including Ann Landers, Mario Andretti, Francesco Scavullo, Mark Hamill, Annie Potts, Chuck Norris, Buckminster Fuller, Erica Jong, Chaim Potok, Isaac Asimov, Ed Muskie and Joe Biden.

Read: 50 weekly radio addresses given by George W. Bush during 2008, and 127 weekly addresses and prepared statements given by Barak Obama between 2009 and 2011.





The distribution of read speech segment durations
seem to shift towards longer segments
compared to the spontaneous speech segment durations.

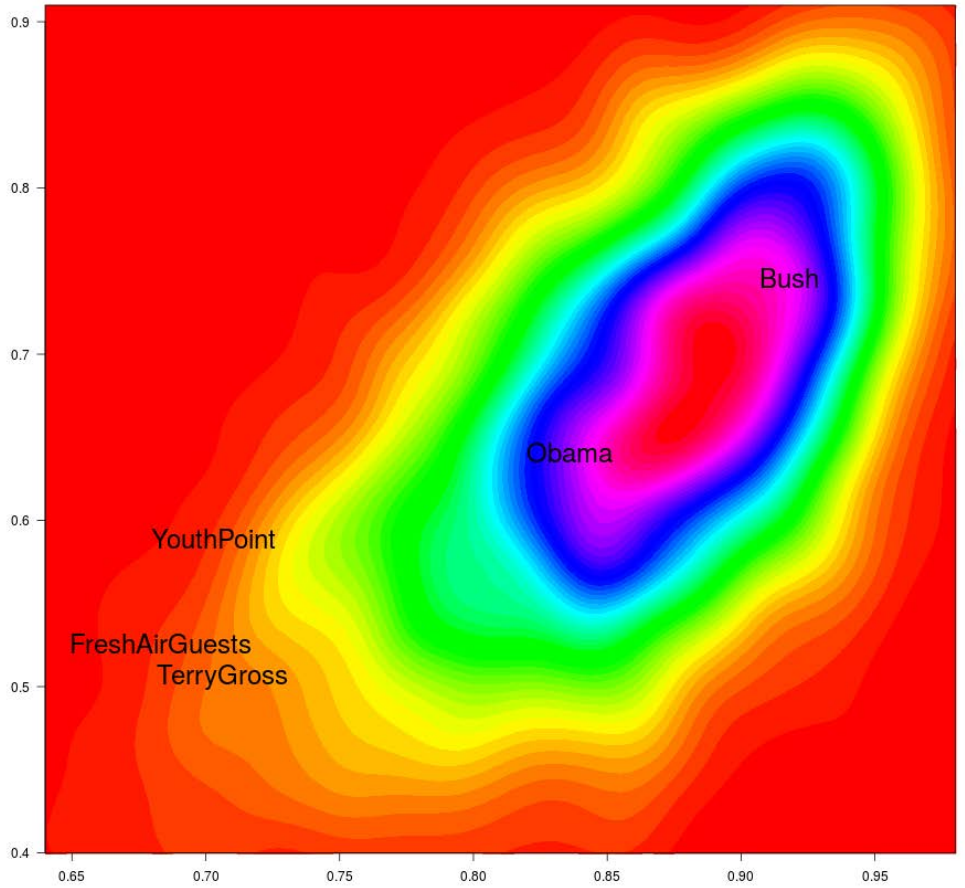
And the 2D density plot
shows this effect clearly:

Obama and Bush
are on opposite sides
of the modal region of readers.

All of the conversational speakers
are down in the tail.

Having background data from
2,484 readers is essential
to establishing this pattern.

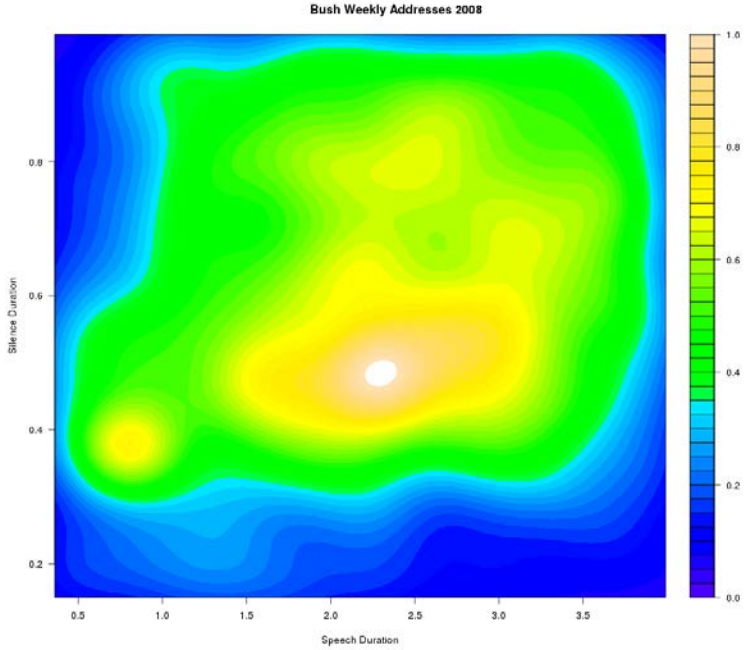
[From Ryant & Liberman, IS 2016]

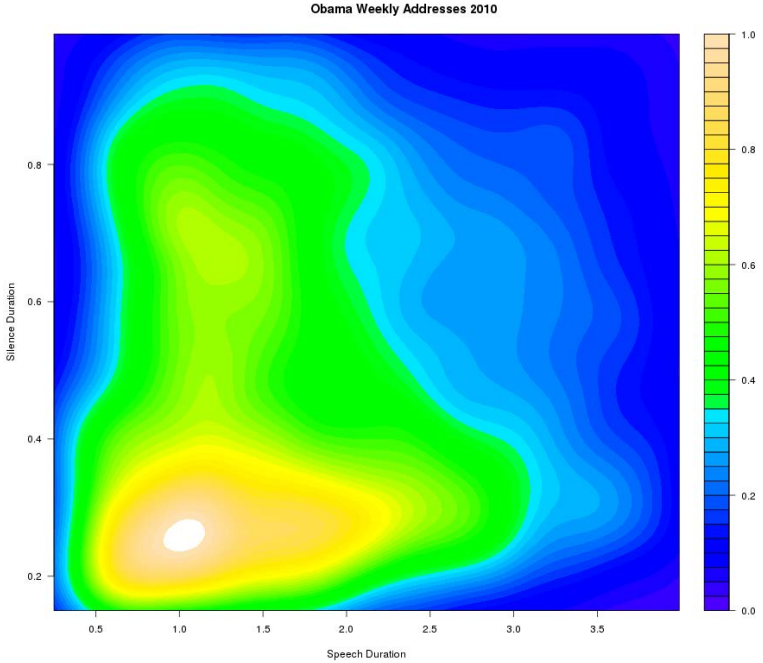


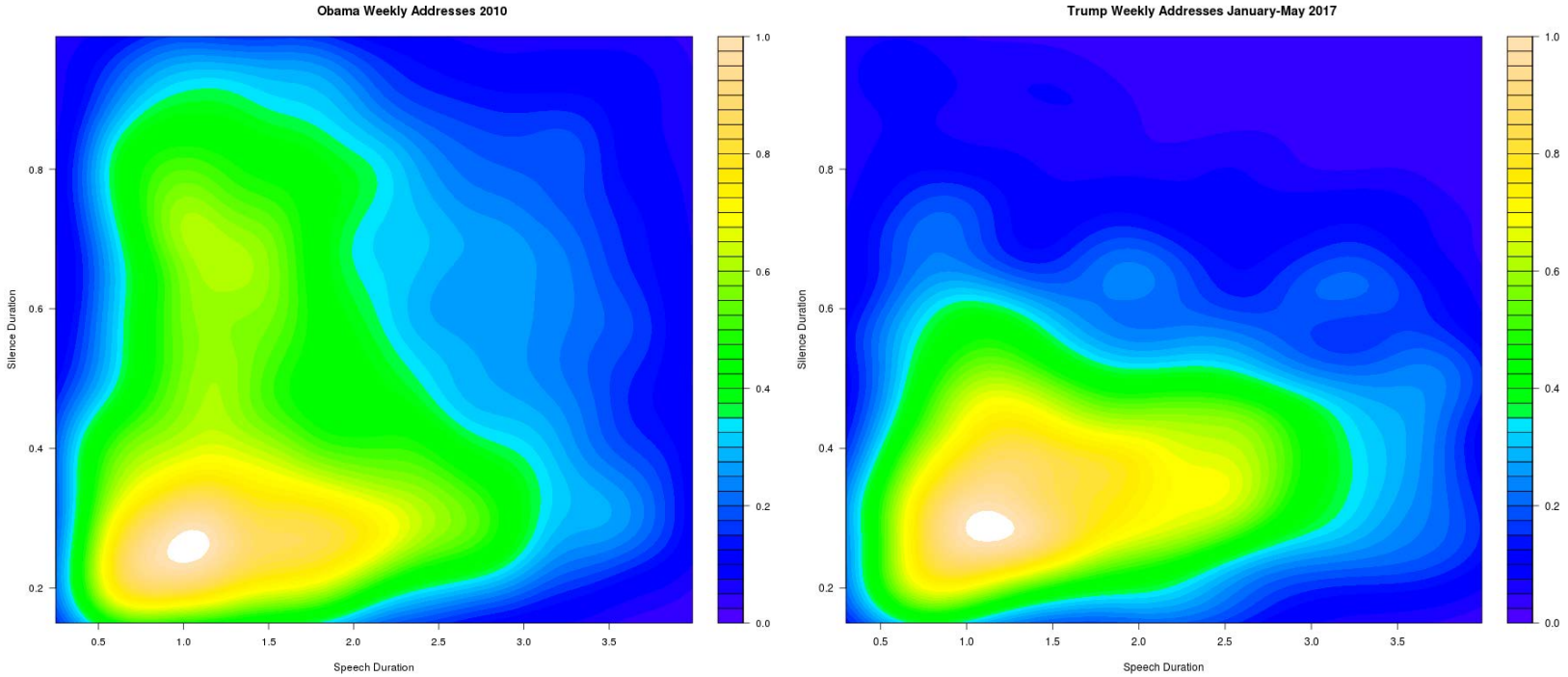
There's much more structure to explore –

Here are 2-D distribution of speech segment durations

and immediately following silence segment durations:






Obama Weekly Addresses 2010
Silence Duration
Speech Duration
0.2
0.4
0.6
0.8
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
Trump Weekly Addresses January-May 2017
Silence Duration
Speech Duration

Increases in scale and efficiency of processing are also important in syntax –

and can help us deal with a sometimes-overlooked consequence of using real-world data.

Laboratory recordings of orthogonally-varied material
have several potential problems:

Reading in an artificial setting is, well, artificial.

Subjects are affected by the orthogonal design of the materials
and speak in ways that are influenced
by the features and dimensions
that are varied or not varied.

These are unintentional experiments in “communicative dynamism”.

But real-world data suffers from the opposite problem.

Its distributions are non-orthogonal at all levels,

so that measures of variation in any given dimension
are likely to be dominated by highly skewed distributions
in other dimensions that are not of interest.

Thus an intended experiment in contextual allophony
may unintentionally be an experiment in word frequency
or syntactic structure differences.

This issue means that we need to
be sensitive to likely co-variates,
and also work at a large enough scale
to contol for them.

An example from syntax:

”Do-support” – Between 1400 and 1800, English changed:

Ate he meat? → *Did he eat meat?*

He ate not meat. → *He did not eat meat.*

Data from Alvar Ellegård 1953:

PERIOD	NEG. QUES. # do	NEG. QUES. % do	AFF. QUES. # do	AFF. QUES. % do	NEG. DECL. # do	NEG. DECL. % do	AFF. DECL. # do	AFF. DECL. % do
1390-1400	0	00.0	0	00.0	0	00.0	0	00.0
1400-1425	2	11.8	0	00.0	0	00.0	11	00.2
1425-1475	2	08.0	6	04.2	11	01.2	121	00.3
1475-1500	3	11.1	10	07.0	33	04.8	1059	01.8
1500-1525	46	59.0	41	22.7	47	07.8	396	01.4
1525-1535	34	60.7	33	32.4	89	13.7	494	02.6
1535-1550	63	75.0	93	44.9	205	27.9	1564	08.1
1550-1575	41	85.4	72	56.3	119	38.0	1360	09.3
1575-1600	83	64.8	228	60.3	150	23.8	1142	06.3
1600-1625	89	93.7	406	69.2	102	36.7	240	03.0
1625-1650	32	84.2	116	82.9	109	31.7	212	02.9
1650-1700	48	92.3	164	79.2	126	46.0	140	01.8

Table 1. The frequency of *do* by environment. [Ellegård 1953:166]

Data compiled by hand, in ~10 years of reading.

Aaron Ecay recently replicated and extended Ellegård's work,
using existing historical treebanks comprising several million words -- in just a few days of programming.

But he ran into a problem:
["Investigating the history of English do-support using automatically annotated corpora", 2015]

> "In the process of investigating the diachrony of do -support in the Penn Parsed Corpora of Historical English (PPCHE), I discovered that there is a difference in the usage of do-support across different argument structure contexts. …
>
> This fact … leads to an account of an intermediate grammar of *d*o where *d*o has been bleached of its causative semantics. … These sentences not be analyzed as tokens of of do-support in the modern sense, but rather tokens of the usage of do as an agentivity marker.
>
> The presentation of the data on argument structure from the PPCHE obscures a fact about the data: it is sparse enough that the so-called argument structure classes are determined by just a few words. Specifically, the experiencer-subject class is dominated by *know*, and the unaccusative class by *come* and (to a lesser extent) *go* . We would like to know whether the properties our analysis imputes to lexical classes are in fact generalizable, or whether they are peculiar to only these lexical items. However, the PPCHE do not contain enough information to investigate the question."

"Thus, I have constructed a new corpus of Early Modern English text, which is much larger than the PPCHE."

- The ***Penn-York Computer-annotated Corpus of a Large amount of English***
- 1 billion words
- Based on the Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO) corpora
- Annotated with POS tags using a 100% automatic process; PPCEME and PPCMBE used as training data

In this source, Aaron was able to disentangle (to some extent) word frequency effects, argument structure effects, word-specific effects, individual author effects, and historical changes.

In these examples, Human Language Technology
brings to scientific or scholarly investigations

- easy re-use of existing digital datasets;
- analysis and tabulation with orders of magnitude less human labor.

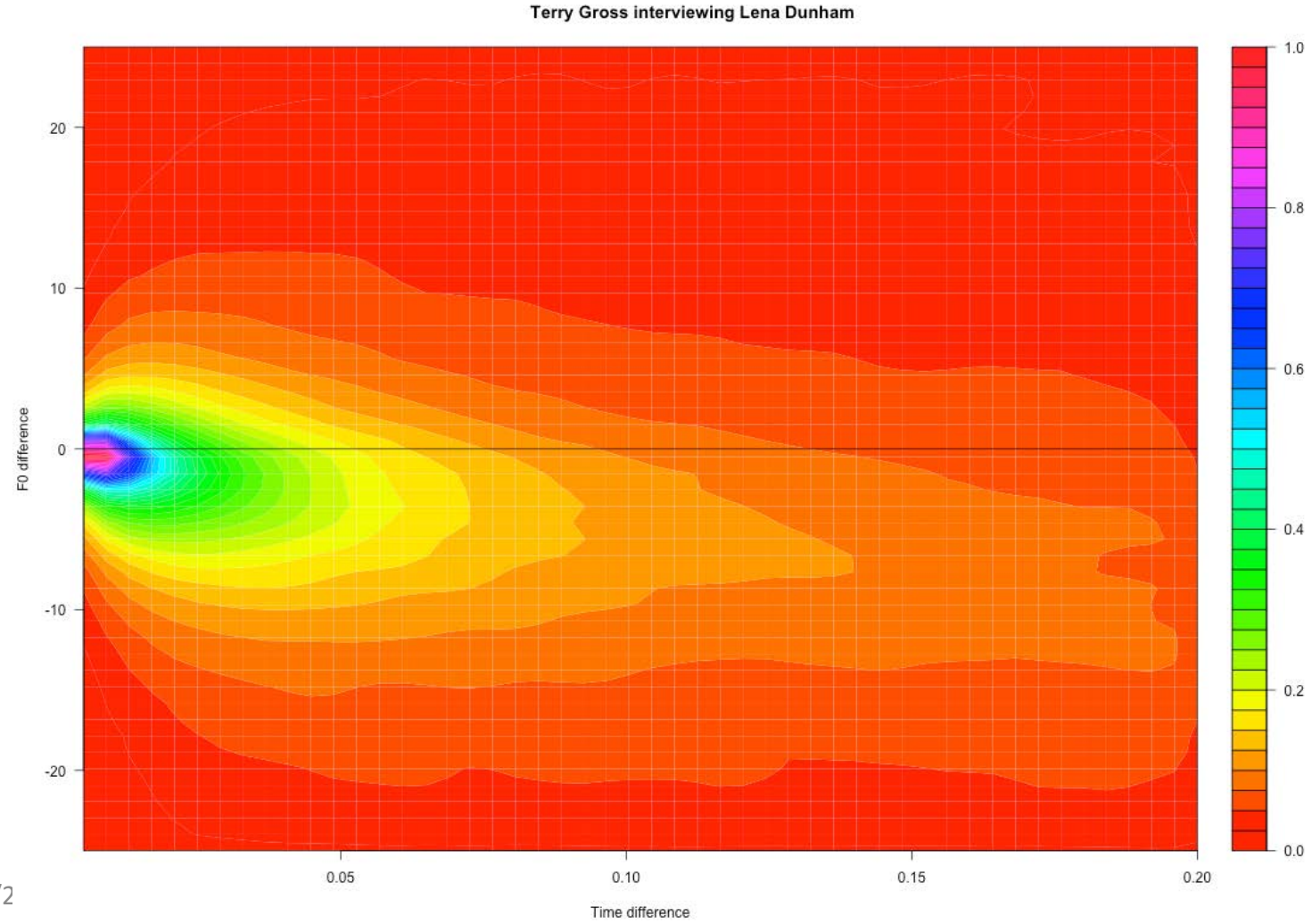
We can work on a scale several orders of magnitude larger than before.

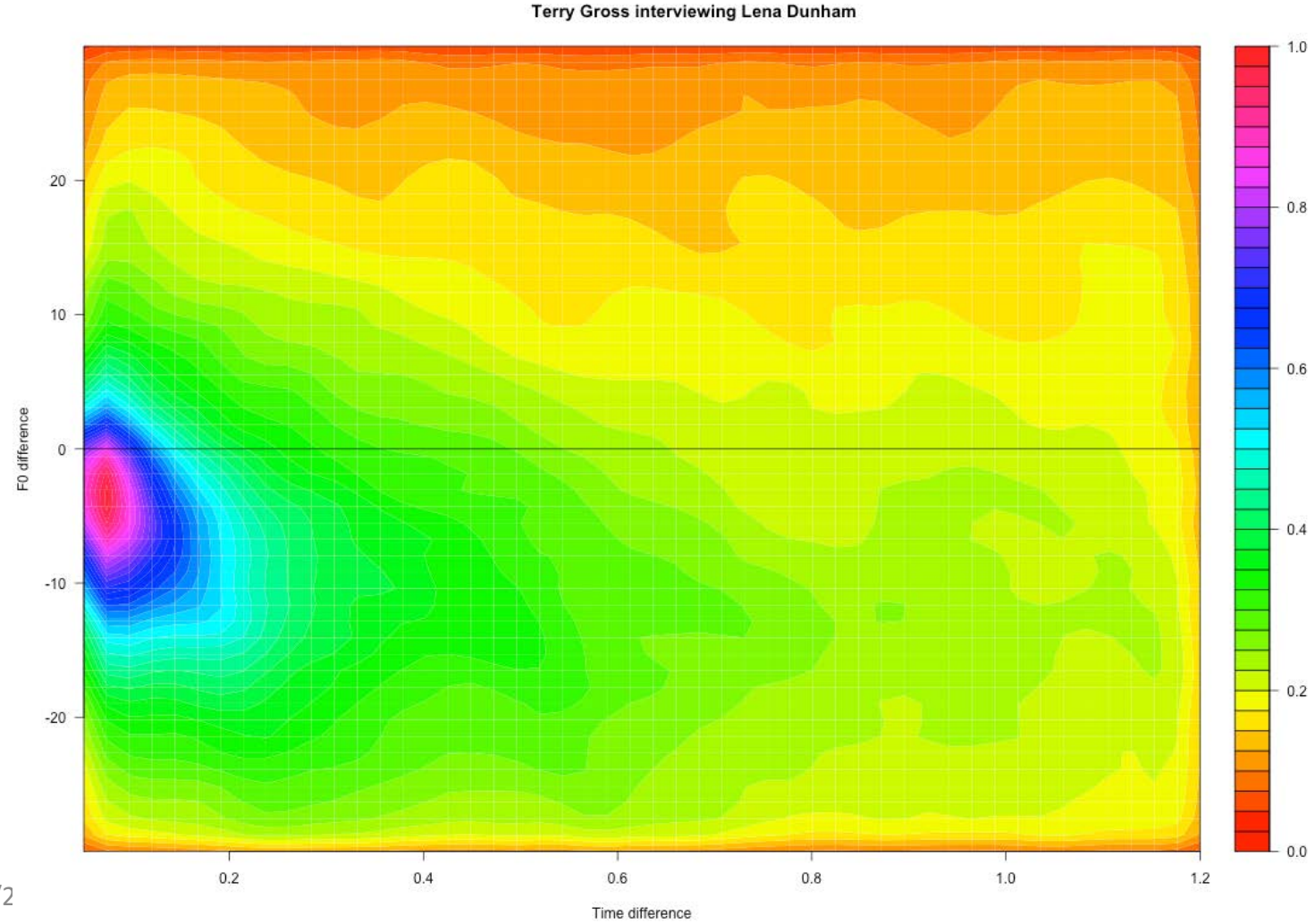
We can test hypotheses in minutes, hours, or days,
rather than weeks, months, or years.

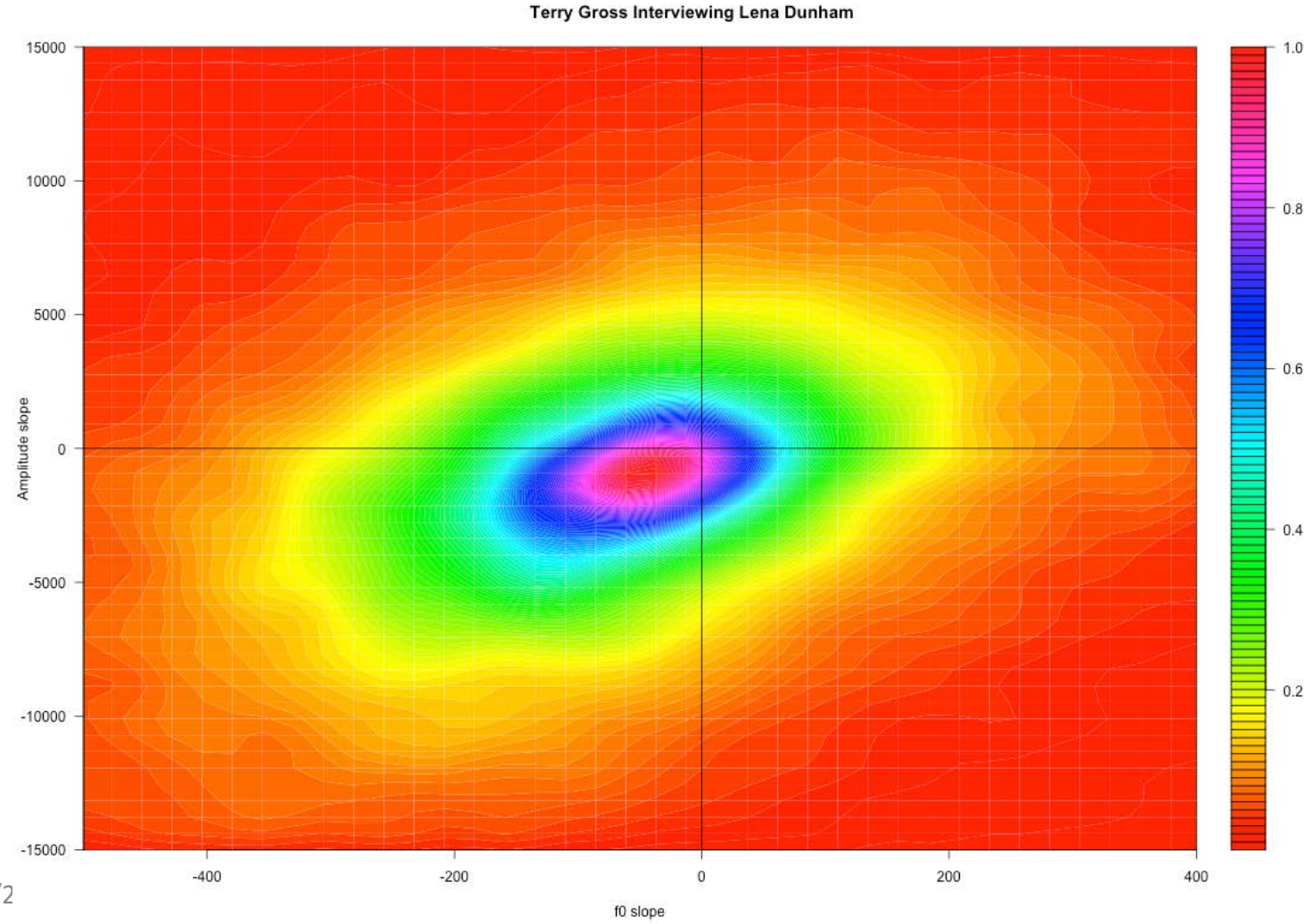
And we can explore large and complex datasets interactively,
to see patterns and generate descriptions.

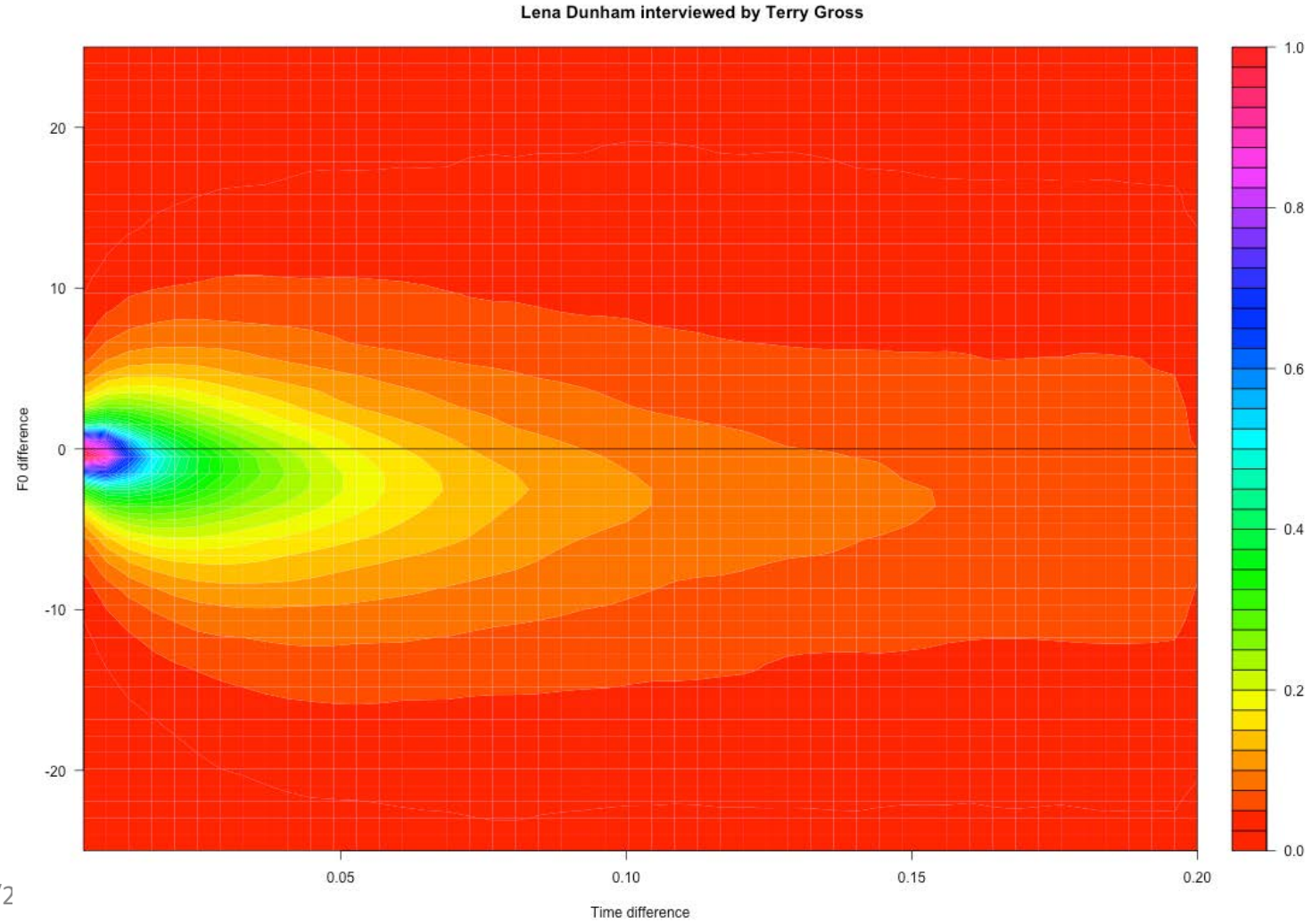
Because HLT (almost) works,
this is an exciting time to do speech and language research!

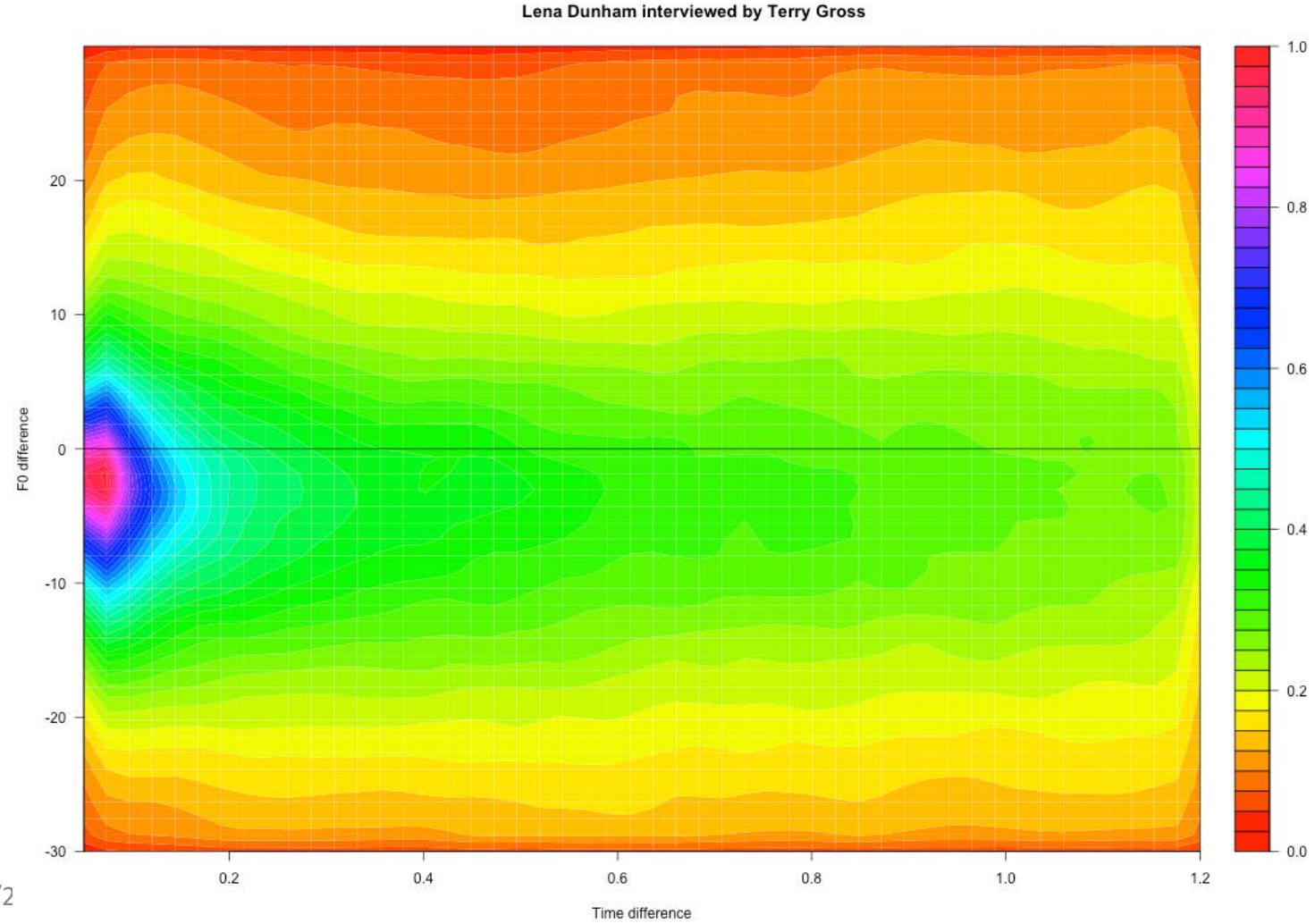
?

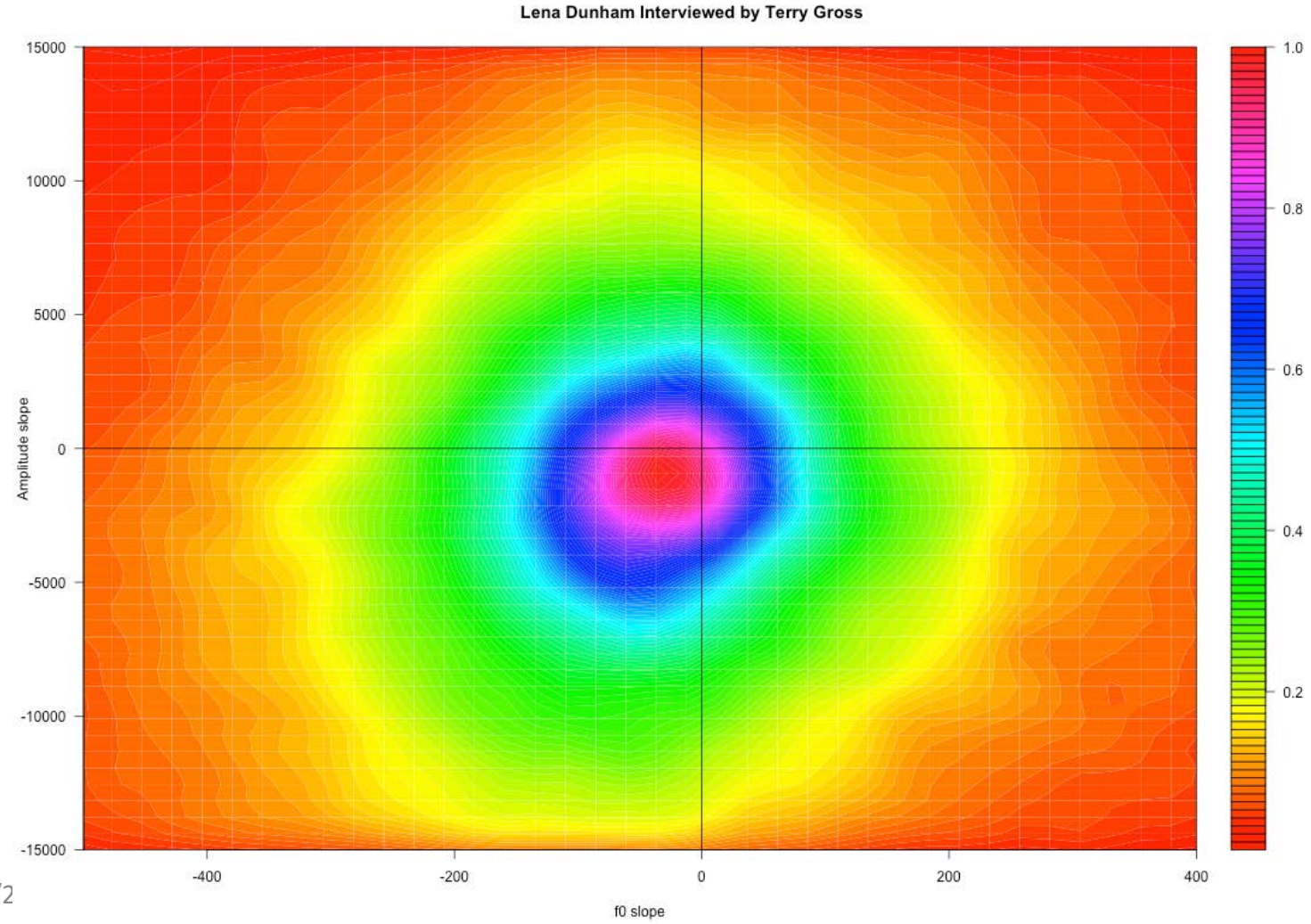

Terry Gross interviewing Lena Dunham
F0 difference
20
10
0
-10
-20
0.05
0.10
0.15
0.20
Time difference
1.0
0.8
0.6
0.4
0.2
0.0


Terry Gross interviewing Lena Dunham
F0 difference
20
10
0
-10
-20
0.2
0.4
0.6
0.8
1.0
1.2
Time difference
1.0
0.8
0.6
0.4
0.2
0.0


Terry Gross Interviewing Lena Dunham
15000
10000
5000
0
-5000
-10000
-15000
Amplitude slope
-400
-200
0
200
400
f0 slope
1.0
0.8
0.6
0.4
0.2


Lena Dunham interviewed by Terry Gross
F0 difference
20
10
0
-10
-20
0.05
0.10
0.15
0.20
Time difference
1.0
0.8
0.6
0.4
0.2
0.0


Lena Dunham interviewed by Terry Gross
F0 difference
20
10
0
-10
-20
-30
0.2
0.4
0.6
0.8
1.0
1.2
Time difference
1.0
0.8
0.6
0.4
0.2
0.0


Lena Dunham Interviewed by Terry Gross
15000
10000
5000
0
-5000
-10000
-15000
Amplitude slope
-400
-200
0
200
400
f0 slope
1.0
0.8
0.6
0.4
0.2