An Infrastructure Network for Interdisciplinary Research in Speech, Language and Human-Computer Interaction

Phil Green

Steve Renals

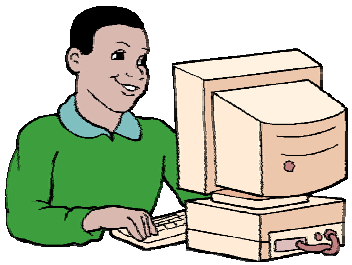
Steve Young

Cambridge University

BACKGROUND

We are a mixed community .... with different ways of working

- design statistical models P(X|Y)
- train parameters on data
- use models to predict Y given X



- design algorithms to map X -> Y
- test predictions on a small corpus
- refine the algorithms to reduce errors

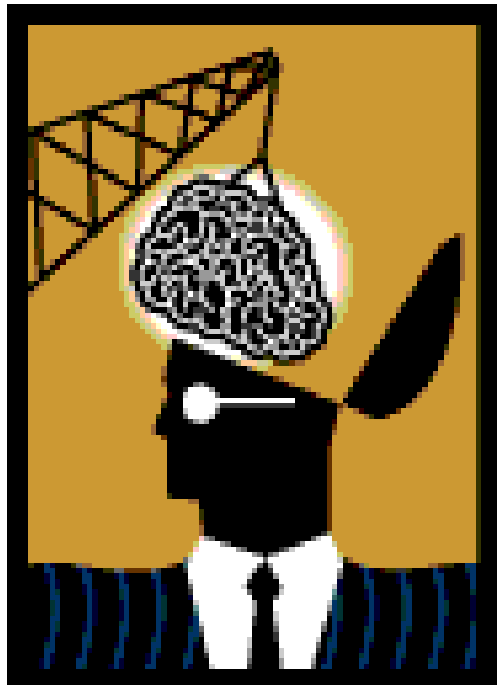
- hypothesise a neural mechanism
- design a controlled experiment to test hypothesis
- perform regression analysis and accept or reject hypothesis

In general, we operate as distinct communities

We have ...

- our own jargon
- our own data sources
- our own journals
- our own conferences
- our own research labs

but we share a common goal



SOME COMMON GROUND

The Need for Data

Of the three communities interested in speech, language and human-computer interaction:

- all rely on data gathered from observing humans
- all need as much data as they can get
- all need the best possible access to new sources of information especially and increasingly brain imaging data

But data is expensive so we need to share it

Many different types of data can be observed during human interactions and human behaviour can only be described accurately by combining observations from multiple sources. E.g.

- speech waveform
- articulatory movement (eg via microbeam x-ray)
- glottal activity (eg via laryngograph, microradar)
- video of lips, facial movement, gestures
- neural activity (fMRI, MEG, etc)

- but very commonly data collection is limited to one layer
- often used for just one experiment and then “archived locally”
- recording conditions rarely standardised

Need to gather data with a view to the wider context

To be really useful, data must be annotated E.g.

- word level orthography
- phonetic transcription, or location of specific phonetic events
- part of speech and syntactic or dependency structure
- tones and break indices
- speech and/or dialogue acts

- annotation effort is rarely shared
- annotations are themselves experimental
- annotation standards/conventions rarely documented

Annotation is expensive but all annotation is potentially reuseable

THE PROPOSED INFRASTRUCTURE NETWORK

SPeech and Language Infrastructure NEtwork

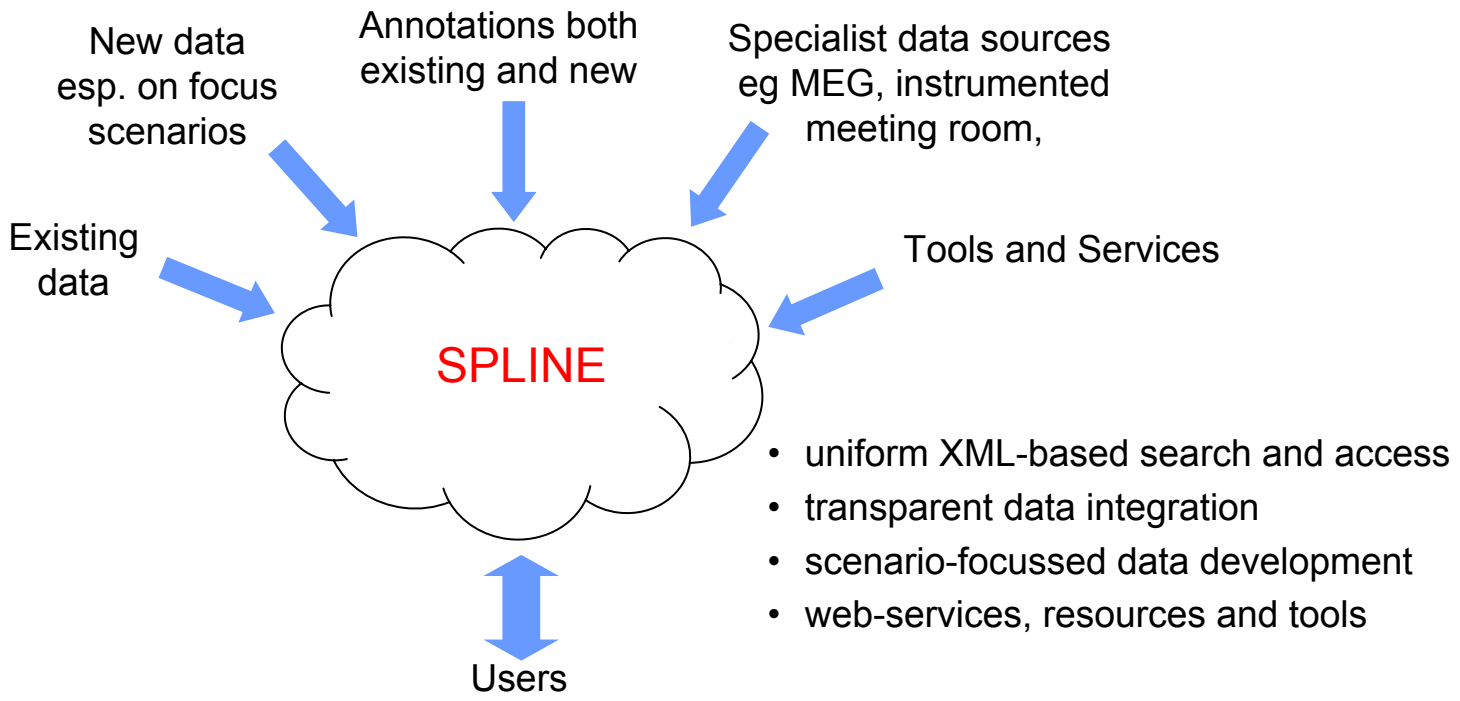
We seek to provide a network infrastructure which

a) extends data access to the widest possible community

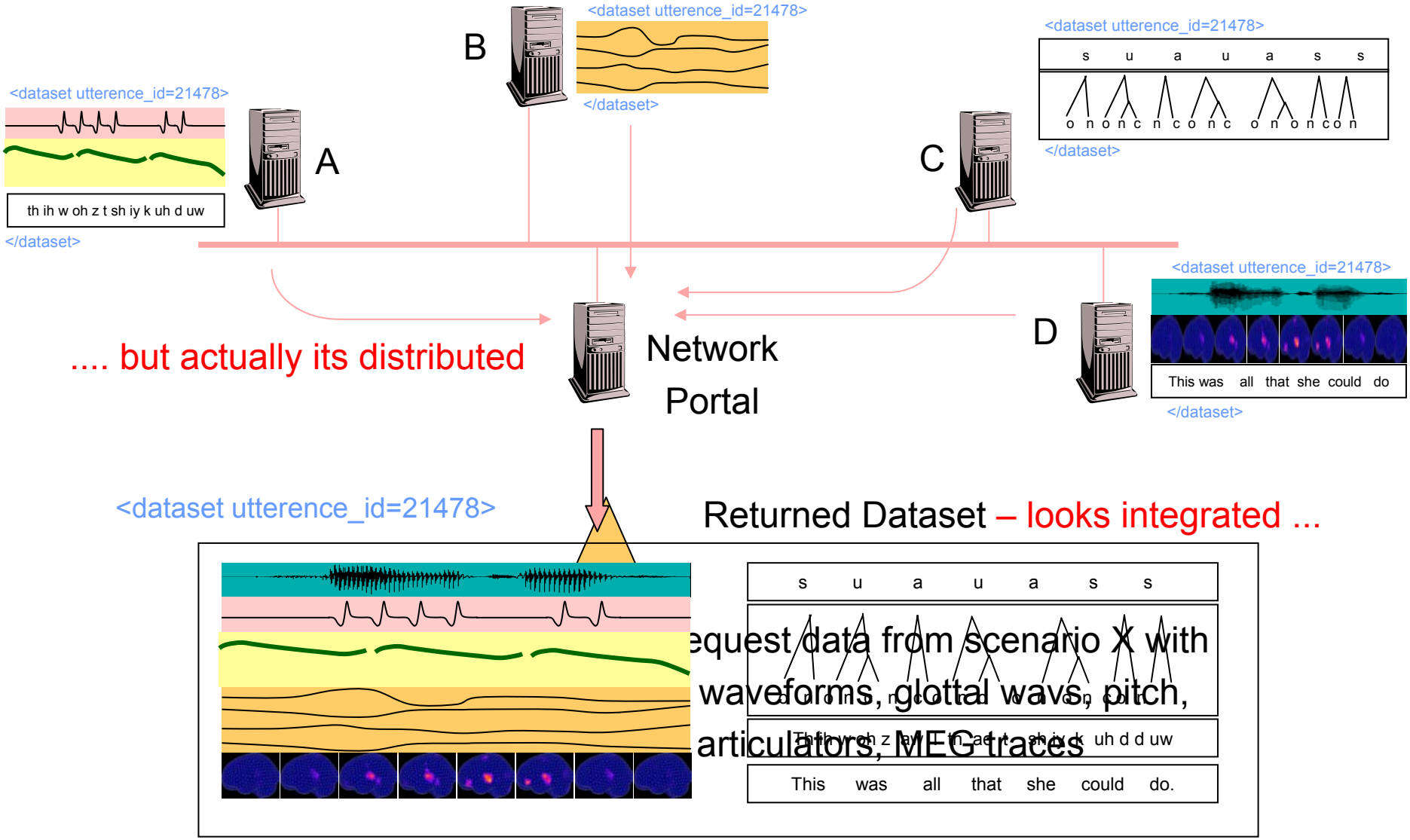
b) extends the utility of data by making it reusable in differing contexts (eg synthesis of parallel data streams from multiple sources)

c) encourages new data collection to take place within a common framework

d) enhances efficiency by providing tools and compute resources applicable directly to the data wherever it is physically located



- uniform XML-based search and access
- transparent data integration
- scenario-focussed data development
- web-services, resources and tools


<dataset utterance_id=21478>
th ih w oh z t sh iy k uh d uw
</dataset>
A
B
<dataset utterance_id=21478>
</dataset>
C
<dataset utterance_id=21478>
s u a u a s s
o n o n c n c o n c o n o n c o n
</dataset>
D
<dataset utterance_id=21478>
This was all that she could do
</dataset>
.... but actually its distributed
Network
Portal
<dataset utterance_id=21478>
Returned Dataset – looks integrated ...
s u a u a s s
This was all that she could do.
equest data from scenario X with
waveforms, glottal wavs, pitch,
articulators, MEG traces
</dataset>

An Example

Why do we need to work on common data?

- A point of research contact for people working in very different areas - serves a valuable ‘social’ function
- Working on the same data provides a solid foundation for collaboration - can give a concrete goal to a set of projects
- Nonlinearly increases the value of the data (exponential rather than log!)
- A good infrastructure makes it possible for new projects to ‘buy in’ to an existing framework
- Allows meaningful comparisons of experimental results
- Both quantity and quality of data grow because it is in the common interest.

NB 11/15 of workshop proposals would fit this model.

Infrastructure Framework

Key ideas

- ❑ individuals publish their own data & tools using very lightweight XML markup
- ❑ data is extended by interested researchers publishing their own additions
- ❑ there is a central data registry but no central control.
- ❑ there are no central archives, but mirror sites encouraged.

- Core is an XML markup language for describing speech and language data resources (cf CML, MathML, etc)
- Data resources classified as:
    - o Primary : original data eg waveform, brain image
    - o Secondary : data derived from other data eg pitch track
    - o Annotation : orthography, break indices, parse, coreferences, etc
- Aim will be to build on existing standards wherever possible (eg ANVIL, MATE, NITE). XSLT translators provided for commonly used formats.
- Primary data is registered and allocated a unique id. All secondary data references this id (or ids)

Scenario-based focus topics for data collection

- Prime aim of network is to integrate existing data and encourage addition of new data
- Coherence will be improved if community could focus on some specific scenarios
- Current projects provide suitable examples:
  - Meetings Data – Human/Human (cf AMI Project)
  - Tourist Information Services – Human/Machine (cf Talk Project)
- Need scenarios which are compatible with current brain imaging data protocols
- Some scenarios may need explicit hardware support eg. MEG, miniRadar, fully instrumented meeting room

Scenario Example 1: AMI

- European Integrated Project based on multimodal meeting recording and analysis, linking:
  - Speech recognition and analysis
  - Organizational psychology
  - Vision
  - HCI
  - Natural language processing
  - Databases
- Many partners, coming together over one data collection (recorded at three sites, many annotations)

Scenario Example 2: Decision-making in cognitive scenes

- Leverage on the M4/AMI meetings room facilities, data and analysis
- Study cognitive systems acting as decision-making agents within multimodal scenes
- Exemplar: the chairperson's problem – how do you control the meeting?
- Existing meetings data used to train recognisers etc
- Use the meetings room facilities, but record scenarios not restricted to meetings.
- New instrumentation needed: better cameras, controlled environmental conditions
- Studies within this project:
  - Models of attention
  - Turn taking and dialogue flow from speech and gestures
  - Embodied conversational agents
  - Cognitive models for the meetings domain
  - Human-like decision-making

Web tools and services

- Initial support will be for browsers and programming interfaces plus encouragement for glossaries, intros, faqs, etc. for cross community education
- Subsequently services will be introduced (eg as a consequence of altruism or a funded research project)
- Example services might include:
  - speech analysis, recognition, synthesis, image enhancement, tagging, parsing, etc
- Heavy compute tasks might rely on grid computing resources
- Other web services might include
  - data mining tools
  - bibliographic search and cross-referencing

OPEN ISSUES

- Is our community ready for this?
- Is there sufficient “bottom-up” demand for it?
- Is the proposal feasible? Especially wrt to brain imaging data.
- Would existing corpora sites such as ELRA and LDC cooperate?
- IP issues: what are they?
- Could it be self-supporting in the long term?