LDC Data Clinic

Christopher Cieri, Denise DiPersio, James Fiumara

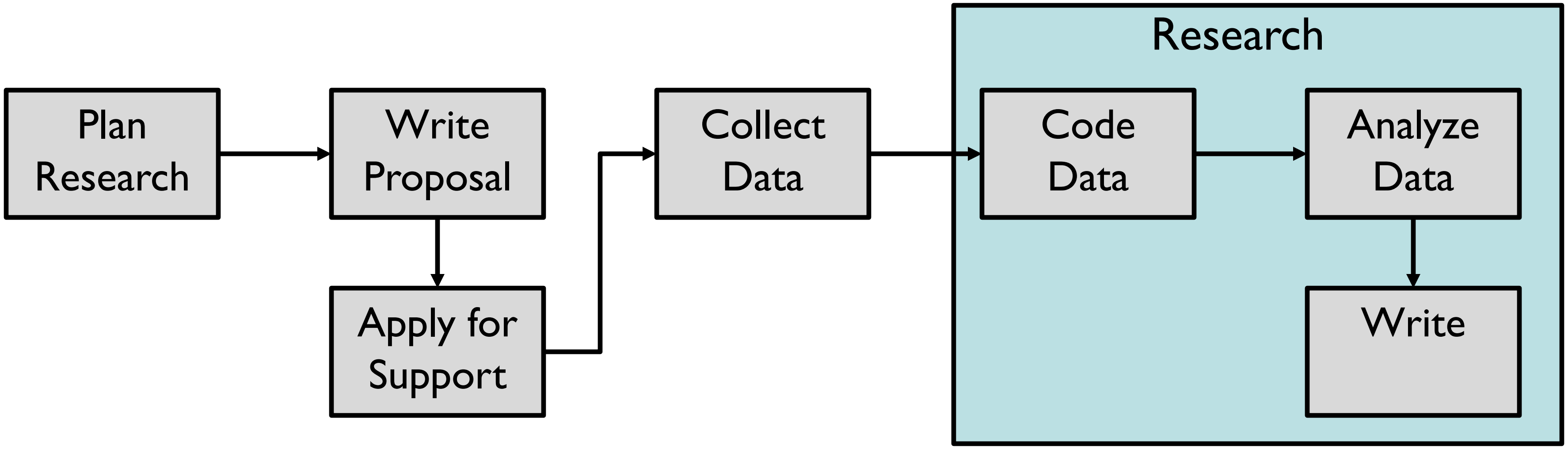
NWAV 48, 2019, Eugene, OR

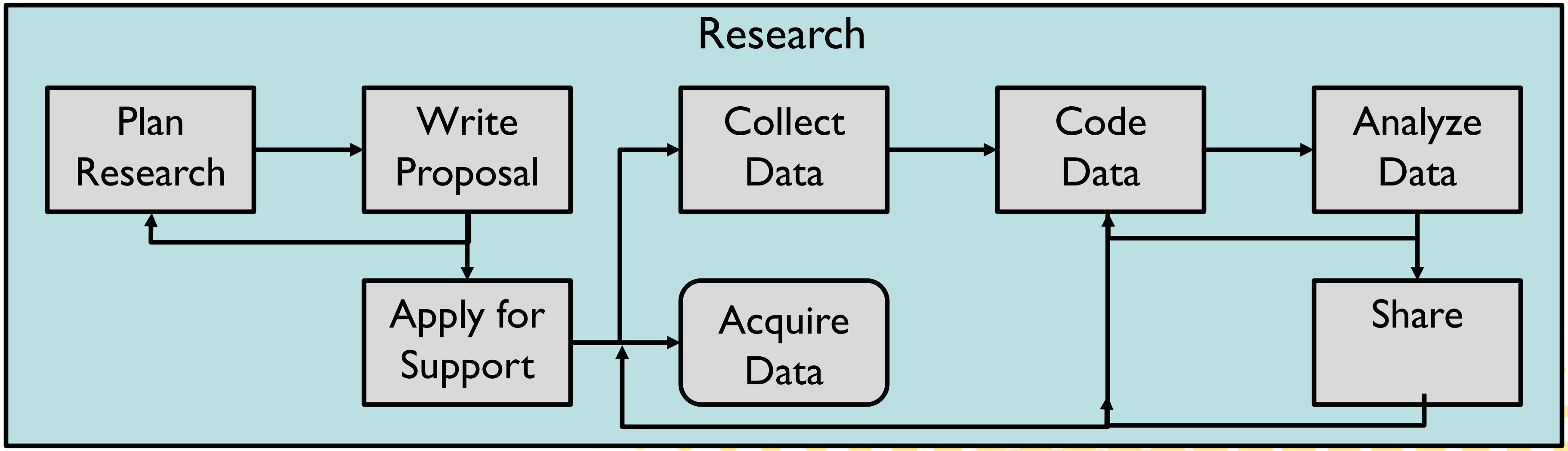
Goals

- Discuss data issues arising from our interaction with this community; intended for students, early careers, people newly facing corpus issues
- Our Credentials
  - LDC = first and largest (also best) data center devoted to language
  - 820 corpora in 90+ languages of which
  - > 187,000 copies distributed to more than 6400 Organizations in 100 Countries
  - Used in > 10,000 Papers
  - Supporting ~10 common task evaluations annually – a model for reuse, replicability
  - Chris
    - PhD in Lx from Penn under Labov, Liberman; LDC Exec. Dir. since 1998, Adj. Assoc, Prof. Lx, 15 additional years exp in LRs, IT
  - Denise
    - JD, 14 Years managing LDC External Relations, (IPR, IRB, distribution, curation)
  - James
    - PhD in English, MA in Lx, 1st Fellow at Penn's Price Lab for DH Fellows, NIEUW project manager

Linguistic Data Consortium

Models of Data Intensive Research





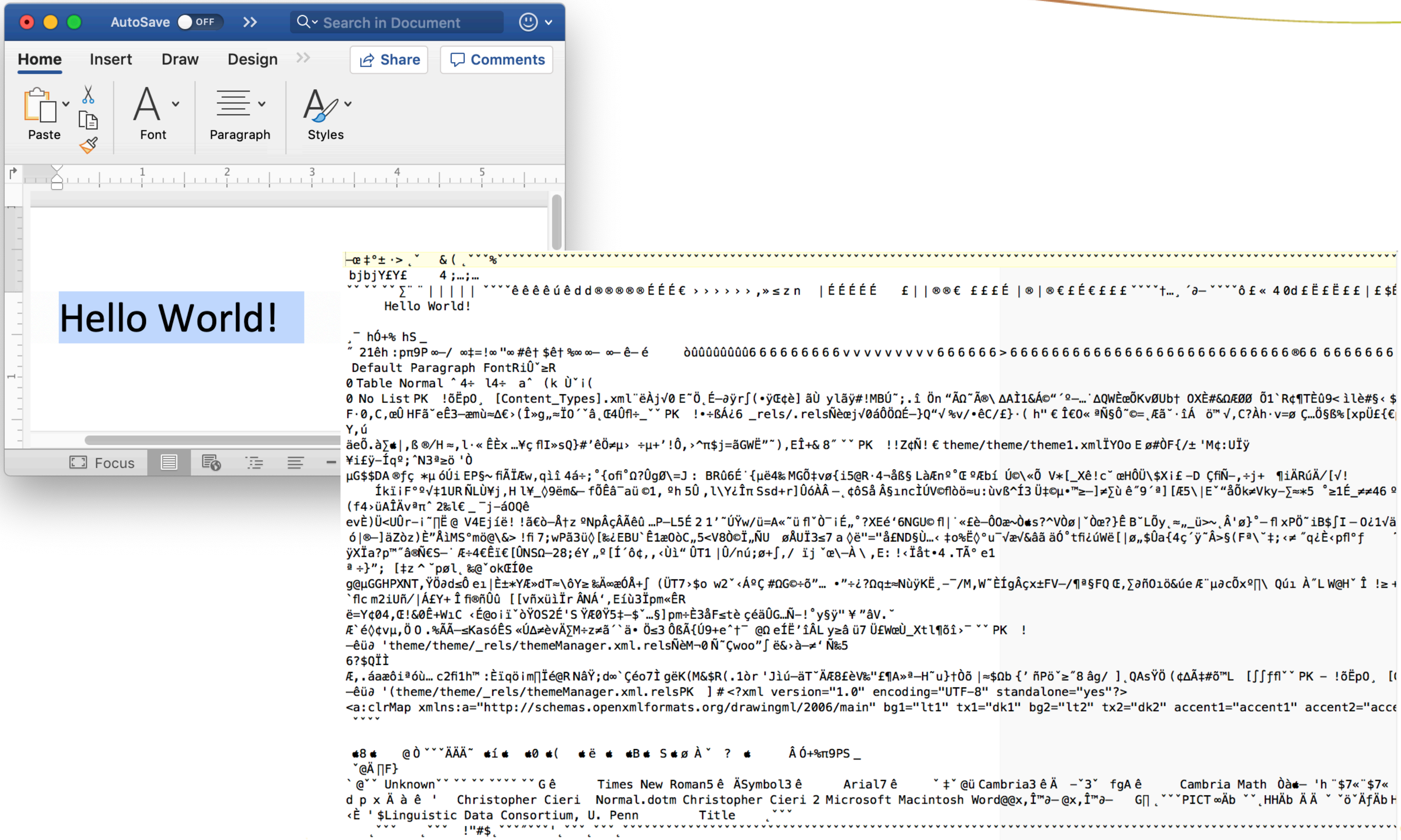
Why Share Data? Why Use Shared Data?

- Funders require it
  - U.S. National Science Foundation, Canada Social Sciences and Humanities Research Council
  - White House Office of Science and Technology Policy (OSTP)
- Journals and Conferences beginning to encourage/require it
- Promotes Good Science
  - replication
  - comparison
  - study of change over time
- Increases Collaboration
- Permits Specialization
- Lower Barriers to Entry
- The research community desires it:
  - The previous, dominant model of considering sociolinguistic data as too valuable to ‘part with’ or to share appears to be giving way to a model where sociolinguistic data is considered to be too valuable not to share. (Kendall, Van Herk 2011: 3)
- Sharpens your own research practices; imposes discipline
- A Good Thing

5 Conditions of Shared Data

- Discoverable
  - user can find, evaluate relevance (absent personal relationship with owner)
- Accessible
  - open terms, procedures stated clearly and in advance for direct access via persistent URL
- Interpretable
  - independently understandable to target community, w/o special resources (e.g. consultation)
- Portable
  - interoperate in user's working environment (hardware, software, community, practices) using open, transparent, widely supported formats
- Preserved
  - faithful copy of original (meta-)data, constantly verified, persists, protected against contingencies (e.g. reliable backup, multiple sites, migration to new media), fixes as patches

(Non-)Portability



Hello World!

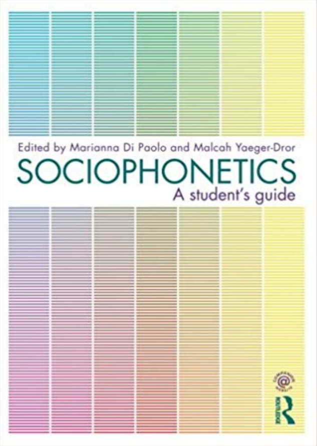
```
–œ‡°±·>˛ˇ   &(˛ˇˇˇ%ˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇ
bjbjY£Y£    4;…;…
ˇˇ ˇˇ ˇˇ∑¨¨|||||ˇˇˇˇêêêêúêdd®®®®®ÉÉÉ€››››››,»≤zn  |ÉÉÉÉÉ   £||®®€ £££É |®|®€££É€£££ˇˇˇˇ†…¸´∂–ˇˇˇˇô£« 40d£Ë£Ë££|£$É
     Hello World!

¸¯ hÓ+% hS_
˝ 21êh :pπ9P∞–/ ∞‡=!∞"∞#ê† $ê† %∞ ∞– ∞–ê–é     òûûûûûûû66666666vvvvvvvvvv666666>666666666666666666666666666®66 6666666
 Default Paragraph FontRiÛˇ≥R
0Table Normal ˆ4÷ l4÷  aˆ (k Ùˇi(
0 No List PK  !õËpO¸  [Content_Types].xml¨ëÀj√0E˘Ö˛É–∂ÿr∫(•ÿŒ¢è] ãÙ ylãÿ#!MBÚ˜;.î Ön "ÃΩ˜Ã®\ ΔAÌ1&Á©"´º–…˙ΔQWÈœÕKvØUb† OXÈ#&ΩÆØØ Õ1`R¢¶TÈû9< ìlè#§‹ $
F·0,C,œÛHFã˘eÊ3–æmù≈Δ€›(Î»g„≈ÏO´˘â˛Œ4Ûfl÷_ˇˇ PK  !•÷ßÁ¿6 _rels/.relsÑèœj√0áÔÖΩÉ–}Q"√%v/•êC/£}·( h"€ Î€O« ªÑ§Ô˜©=¸Æã˘·îÁ  ö™√,C?Àh·v=ø Ç…Ö§ß%[xpÜ£{€
Y,ú
äeÕ.à∑|,ß®/H≈,l·«ÊÈx…¥çflI»sQ}#'êÖ≠µ› ÷µ+'!Ô,›^π$j=ãGWË"˜),EÎ+&8˝ ˇˇ PK  !!Z¢Ñ!€theme/theme/theme1.xmlÏYOo E ø#ÒF{/± 'M¢:UÏÿ
¥i£ÿ–Íqº;ˆN3ª≥ö 'Ò
µG$$DA®ƒç ∗µóÚi EP§~fiÃÏÆw,qìî 4á÷;˚{ofi˚Ω?ÛgØ\=J : BRû6É˙{µë4‰MGÕ‡vø{i5@R·4¬åß§ LàÆnº˚Œ ºÆbí Ú©\«Õ V∗[_Xê!c˘ œHÔÜ\$Xi£–D ÇfiÑ–,÷j+  ¶iÄRúÄ/[√!
    ÍkïiF°º√‡1URÑLÙ¥j,H l¥_◊9ëm&– fÕÊâ¯aü©1, ºh 5Û ,l\Y¿Îπ Ssd+r]ÛóÀÂ–¸¢ôSå Â§ıncÌÚV©flòö≈u:ùvß^Í3 Ü‡©µ•™≥–]≠∑ù ê˝9´ª] [Æ5\|Eˇ"åÕk≠Vky–∑≈∗5 ˚≥1É_≠≠46 º
(f4›üAÎÄvªπˆ 2‰l€_¯j–áOQê
evÈ)Ü<UÛr–i˜∏Ë@ V4Ejíë! !ã€ò–Å†z ºNpÂçÂÃêû …P–L5É 2 1'˜ÚŸw/ü=A«˜ü flˇÒ¯iÉ„˚?XEé'6NGU© fl|˙«£è–Ô0æ~Òs?^VÒø|ˇÒœ?}Ê B˘LÕy˛≈„_ü>~˛Â'ø}˚–fl xPÖ˜iB$∫I – 0¿1√ä
ó|®–]äZòz)È"ÅìMS°mö@\&> !fi 7;wPã3ü◊[‰¿EBU`Ê1æ0òC„5<V8Ò©Ï„ÑU  øÅUÏ3≤7 a ◊ë"="å£ND§Ù…‹ ‡o%Ë◊°u¯√æ√&ââ äÓ˚tfi¿úWë[|ø„$Ûa{4ç´ÿ˜Â>§(Fª\˘‡;‹≠ ˝q¿È‹pfl°ƒ     ˆ
ÿXÏa?p™˝â®Ñ€S–˙ Æ÷4€Êï€[ÛNSΩ–28;éY „º[Í´ô¢,,‹Ùì" ÛT1 |Û/nú;ø+∫,/ ïj ˇœ\–À \ ,E: !‹Ïåt•4 .TÃ° e1
ª ÷}"; [‡z ^ ˇpøl˛ ‰@ˇokŒÍ0e
g@µGGHPXNT,ŸÖ∂d≤Ô eı|È±∗YÆ»dT≈\ôY≥‰Ä∞æÓÅ+∫ (ÜT7›$o w2ˇ‹ÁºÇ #ΩG©÷õ"… •"÷¿?Ωq±≈NùÿKË¸–¯/M,W˜ÈÍgÂçx±FV–/¶ª§FQ Œ,∑∂ñ0ıö&úe Æ¨µ∂cÕxº∏\ Qúı À˝L W@Hˇ Î !≥ +
`flc m2iUñ/|Á£Y+ Î fi®ñÛû  [[vñxüìÏr ÂNÁ',Eíù3Ïpm«ÊR
ë=Y¢04,Œ!&0Ê+WıC ‹É@oiïˇòŸOS2É'S ŸÆ0Ÿ5‡–$ˇ…§]pm÷È3åF≤tè çéäÛG…Ñ–!˚y§ÿ" ¥ ˝âV.ˇ
Æ`é◊¢vµ,Ö O .%ÃÄ–≤KasóÊS «ÚΔ≠èvÄ∑M÷z≠ã´`ä• Ö≤3 ÔßÃ{Ú9+eˆ†¯ @Ω eÍË'îÂL y≥â ü7 Ü£WœÙ_Xtl¶õî›¯ ˇˇ PK  !
–êü∂ 'theme/theme/_rels/themeManager.xml.relsÑèM¬0Ñ˜Çwoo"∫ë&›à–≠' Ñ‰5
6?$QÏÌ
Æ,.áaæôiªóù… c2fi1h™ :Èïqöim∏Ïé@R NâŸ;d∞`Çéo7Ì gëK(M&$R(.1òr 'Jìú–äT˘ÄÆ8£èV‰"£¶A»ª–H˜u}†Òõ |≈$Ωb {' ñPöˇ≥˝8 âg/ ]˛QAsŸÖ (¢ΔÃ‡#õ™L  [∫∫ƒflˇˇ PK – !õËpO¸  [(
–êü∂ '(theme/theme/_rels/themeManager.xml.relsPK  ]#<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<a:clrMap xmlns:a="http://schemas.openxmlformats.org/drawingml/2006/main" bg1="lt1" tx1="dk1" bg2="lt2" tx2="dk2" accent1="accent1" accent2="acce
 ˇˇˇˇ


 8    @Ò ˇˇˇÄÄÄ˜ í  0 (   ë    B  S ø À ˇ  ?      Â Ó+%π9PS _
 ˇ@Ä∏F}
`@ˇˇ Unknownˇˇ ˇˇ ˇˇ ˇˇˇˇ ˇˇG ê     Times New Roman5 ê ÄSymbol3 ê     Arial7 ê     ˇ‡ˇ @ü Cambria3 ê Ä  –ˇ3ˇ  fgA ê     Cambria Math  Òà– 'h ¨$7«¨$7«
d p x Ä à ê '   Christopher Cieri  Normal.dotm Christopher Cieri 2 Microsoft Macintosh Word@@x,Î™∂–@x,Î™∂–   G∏˛ˇˇˇPICT ∞Äb ˇˇ˛HHÄb Ä Ä ˇ ˇöˇÄƒÄb H
‹È '$Linguistic Data Consortium, U. Penn        Title   ˛ˇˇˇ
   ˛ˇˇˇ    ˛ˇˇˇ  !"#$˛ˇˇˇ˝ˇˇˇ'˛ˇˇˇ˛ˇˇˇ ˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇ
```

Collection (e.g. Audio)

- Does collection adequately represent research purpose?
- Underlying assumptions about collection
  - control over self
  - control over speaker, Observer's Paradox, over-sampling
  - control over situation
    - room
    - noise
- Recording Parameters
  - sampling rate, sample size, compression, format
- Microphone / Recorder
    - **usability & compatibility** (especially connectors and power supply)
    - operating parameters such as placement distance & direction relative to type
    - make, model, pick-up pattern
- Comparability

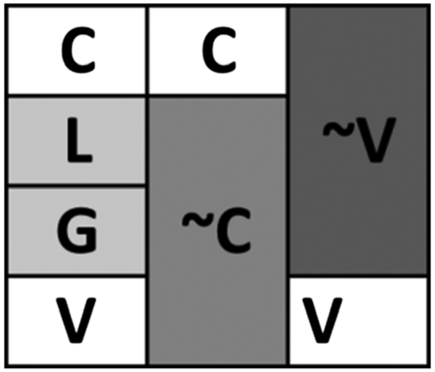


Coding Data (e.g. Speech)

- In additional to shareability, methodology impacts
  - efficiency & scale thus results
  - accuracy & consistency
  - balance, completeness & dynamics
  - (your) reuse for another purpose & replicability
- text, (e.g. transcript, generally human) required for efficient methods
- Decision Points & Comparability
  - nature of dependent variable
  - modeled: continuous/discrete/categorical, separate/part of general process
  - independent variables, or factor groups, considered
  - values (or variants or factors), assigned to each of factor group
  - sampling: talkers, situations, tokens considered/excluded, features by which
  - elicitation method

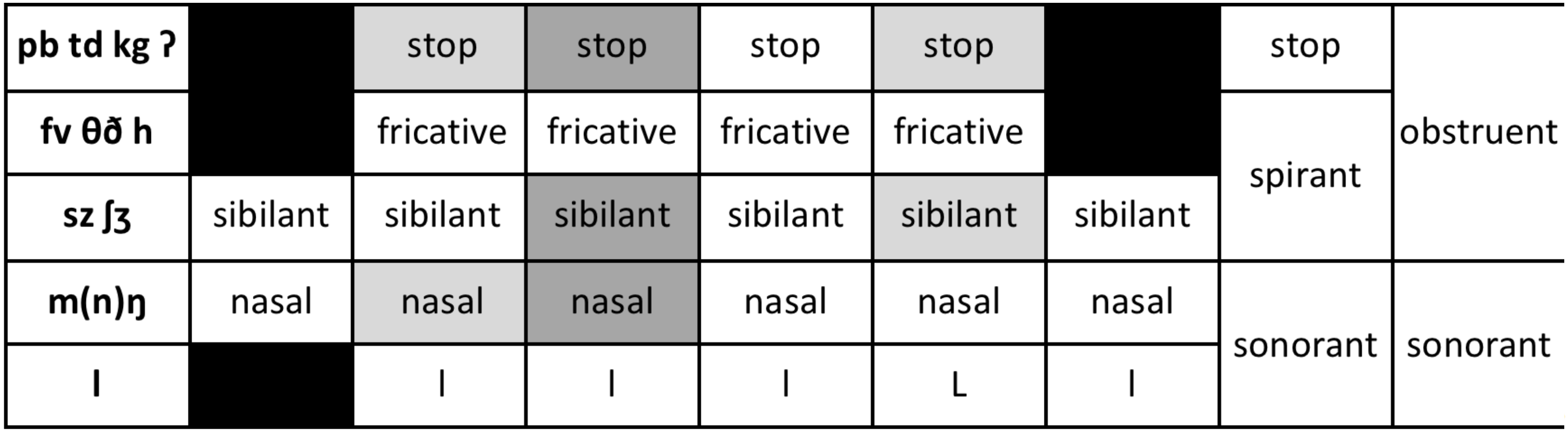
Coding Differences & Comparability

-t/d following environment



C	C	~V
L	~C	
G		
V		V

-t/d preceding environment



pb td kg ʔ	■	stop	stop	stop	stop	■	stop	
fv θð h	■	fricative	fricative	fricative	fricative	■	spirant	obstruent
sz ʃʒ	sibilant	sibilant	sibilant	sibilant	sibilant	sibilant		
m(n)ŋ	nasal	nasal	nasal	nasal	nasal	nasal	sonorant	sonorant
l	■	l	l	l	L	l		

black = not reported same shading within column = not reported as statistically significant.

Coding Metadata

- Demographic
  - association with multiple groups
  - immigrant communities
  - disenfranchised groups
- Situation
- Attitude
- Nature of elicitation matters
  - mined from discussion: in what context
  - elicited by written form or by live questions
    - form of the question, number of categories and which
    - who asked
  - any merging of categories (e.g. due to imbalance in sample)
  - change over time
- When is a simplifying assumption a distorting practice?

Volume 8, Issue 11

Pages: i-ii, 465-628

November 2014

New Methods for Data Collection/Annotation

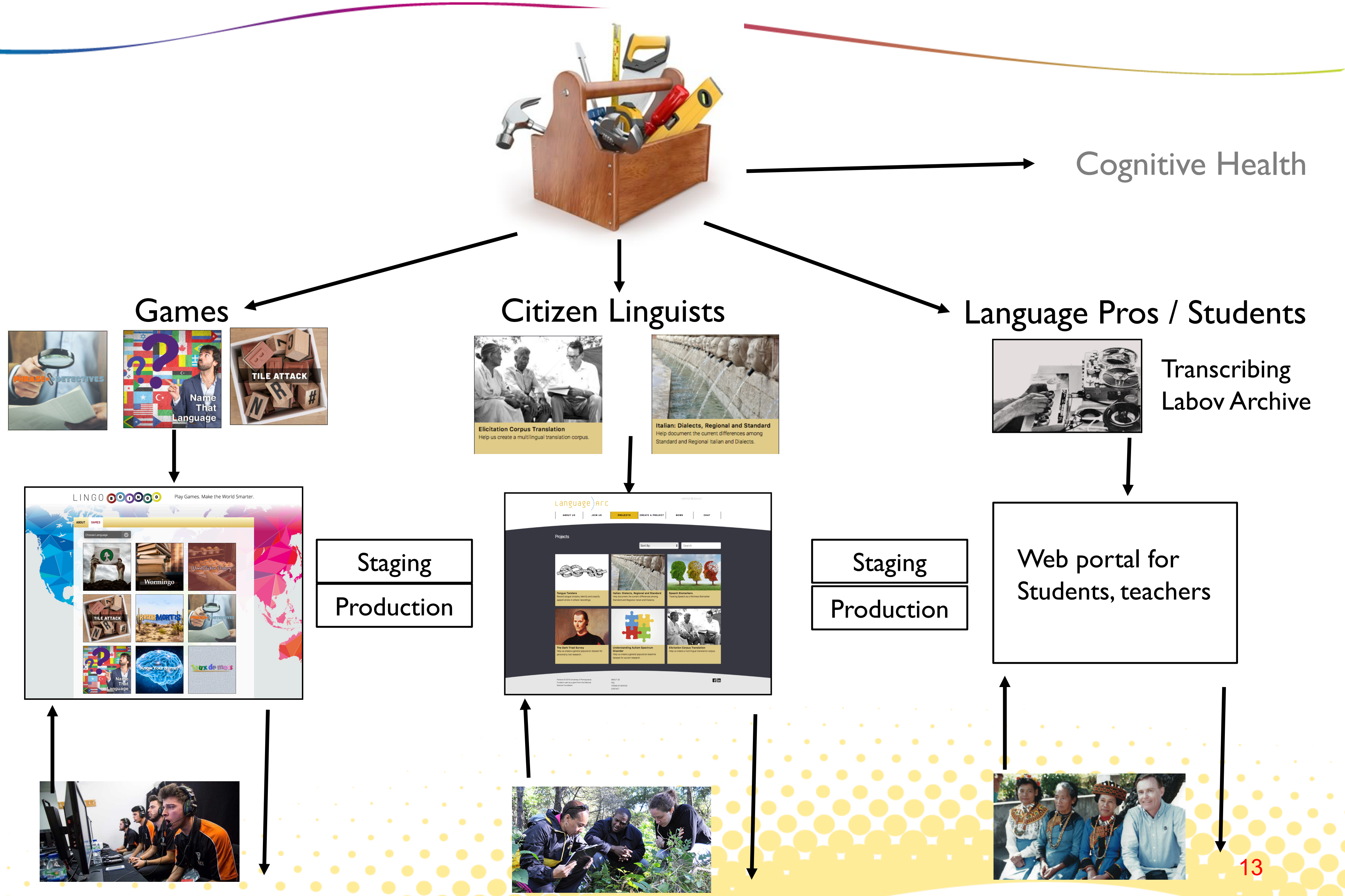
- Creation of LRs limited as it employs finite resources (e.g., grant funds) for nearly infinite problem
- Proposal: use novel incentives to harness the renewable resources of the human drives to learn, compete and make meaningful contributions.
- Novel Incentives: information, entertainment, self-expression, developing skills, socializing, competition, status, recognition, contributing to greater cause or good
- Successful examples:
  - LibriVox: public domain audiobooks, >9,400 volunteers, >13,000 books, 39 languages
  - Zooniverse: Citizen Science portal founded 2009, over 200 total projects (105 active), >1,900,000 contributors
  - Game with a Purpose (GWAP): ESP Game, Google Image Labeler, Great Language Game

New Methods for Data Collection/Annotation

- NSF CISE Research Infrastructure (CRI) grants
  - Planning Grant: 2016 - 2017
  - NEW Grant: July 2017 – June 2020
- Develop and build infrastructure:
  - Software toolkit to build lx activities and projects
  - Web portals for hosting projects, crowdsourcing, community building
- Build upon LDC's WebAnn (web annotation) tools
  - Simplify tools to allow researchers to build lx annotation activities online
  - Simplify annotation task design & interface for non-experts
  - Create multiple portals geared towards different incentives / workforces
    - gamers, citizen scientists, language teachers/students

Linguistic Data Consortium

Multiple web portals



Cognitive Health

Games

Citizen Linguists

Language Pros / Students

Transcribing Labov Archive

Staging

Production

Staging

Production

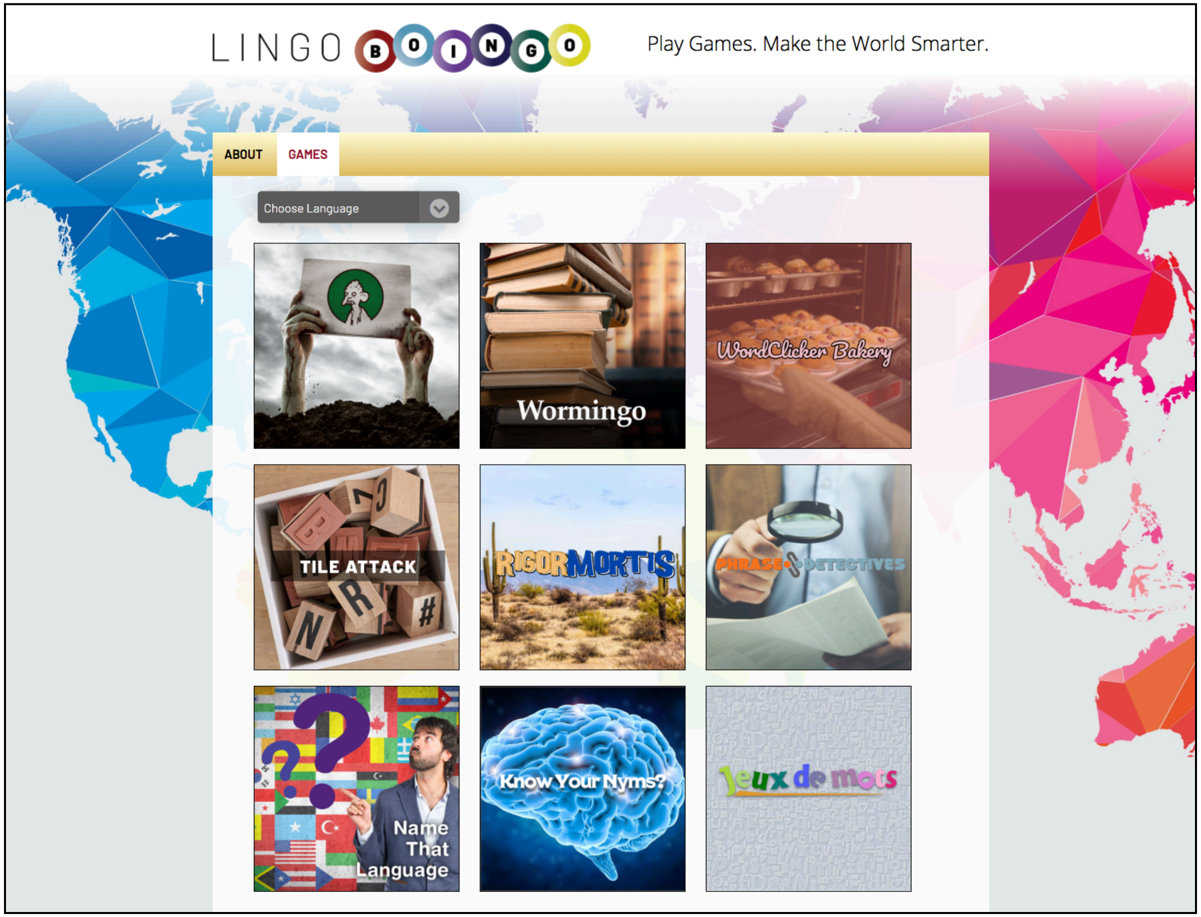
Web portal for Students, teachers

Phrase Detectives
Name That Language
TILE ATTACK
Elicitation Corpus Translation
Help us create a multilingual translation corpus.
Italian: Dialects, Regional and Standard
Help document the current differences among Standard and Regional Italian and Dialects.
LINGO BOINGO
Play Games. Make the World Smarter.
Wormingo
Know Your Rhyme?
Language ARC
Projects

Linguistic Data Consortium

Games Portal

lingoboingo.org



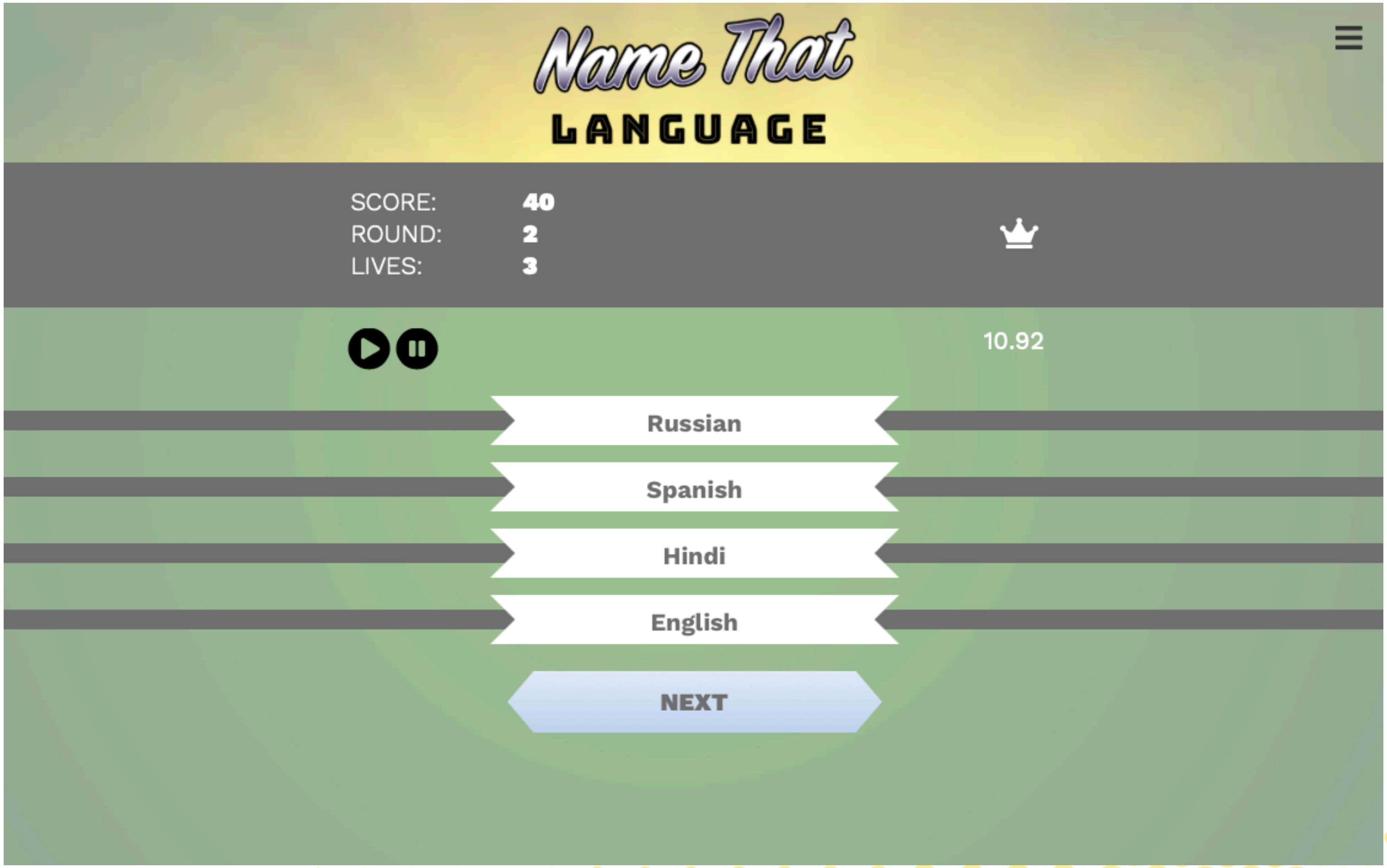
Play Games. Make the World Smarter.

Play Games, Make the World Smarter

The LingoBoingo web pages are Copyright 2017 Trustees of the University of Pennsylvania.
Copyright for individual games is specific to those games.
See individual game sites for more information.

Games Portal

namethatlanguage.org



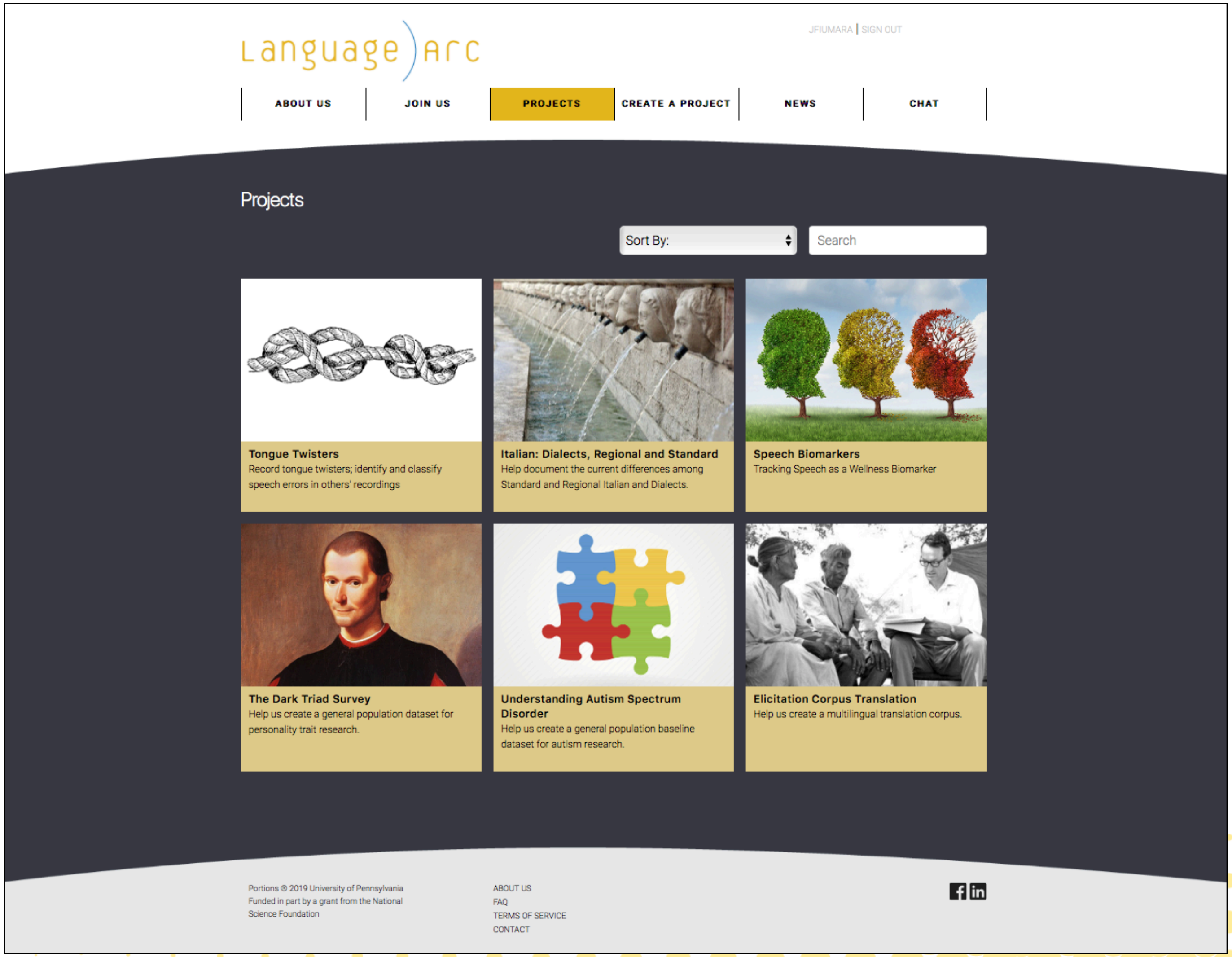
Name That LANGUAGE

SCORE: 40
ROUND: 2
LIVES: 3

10.92

Citizen Science Portal

languagearc.org



Projects

Tongue Twisters
Record tongue twisters; identify and classify speech errors in others' recordings

Italian: Dialects, Regional and Standard
Help document the current differences among Standard and Regional Italian and Dialects.

Speech Biomarkers
Tracking Speech as a Wellness Biomarker

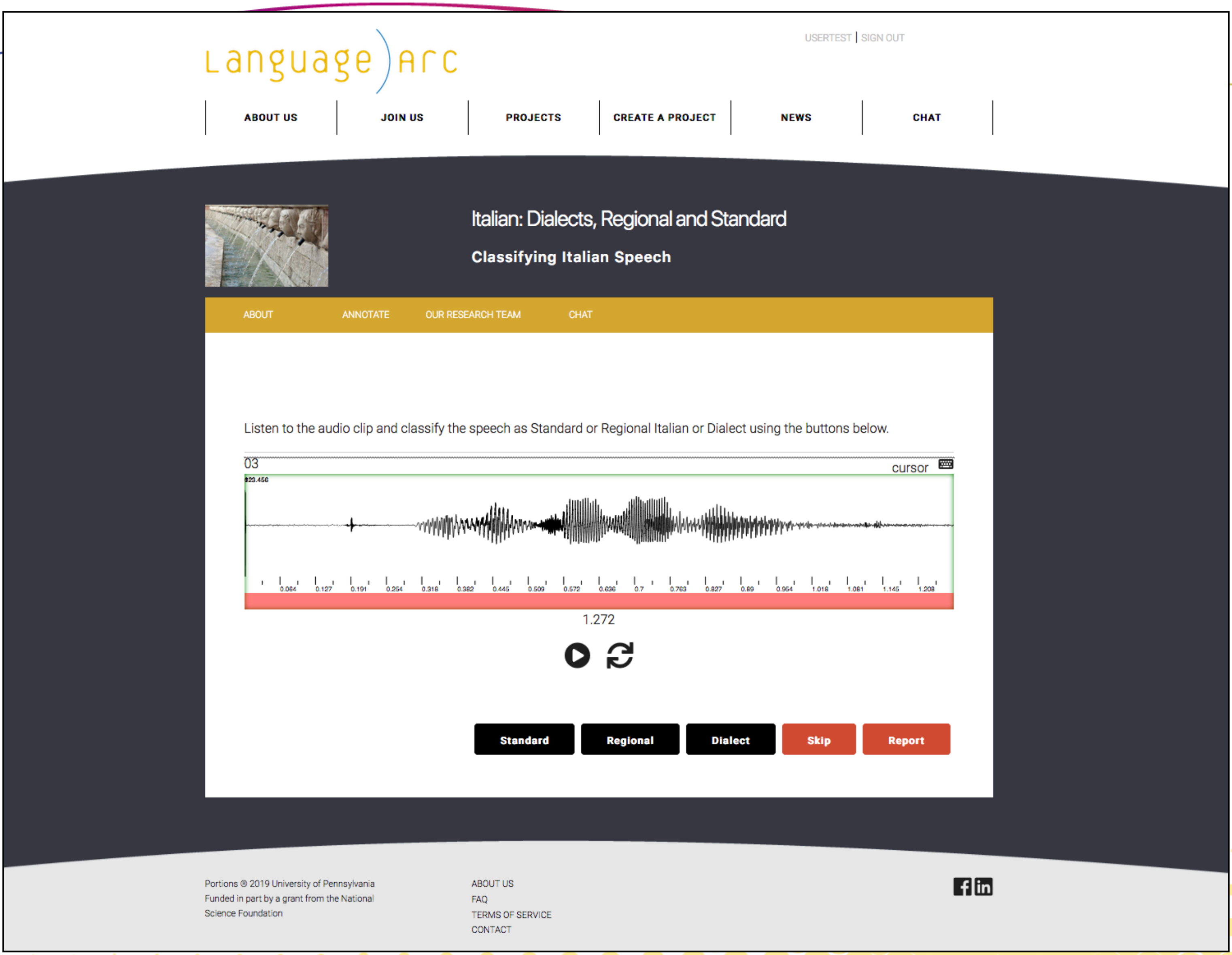
The Dark Triad Survey
Help us create a general population dataset for personality trait research.

Understanding Autism Spectrum Disorder
Help us create a general population baseline dataset for autism research.

Elicitation Corpus Translation
Help us create a multilingual translation corpus.

Portions © 2019 University of Pennsylvania
Funded in part by a grant from the National Science Foundation

Citizen Science Portal



Italian: Dialects, Regional and Standard

Classifying Italian Speech

Listen to the audio clip and classify the speech as Standard or Regional Italian or Dialect using the buttons below.

Portions © 2019 University of Pennsylvania
Funded in part by a grant from the National Science Foundation

Current Status

- LingoBoingo (games portal)
  - deployed
  - link farm to partners' games
  - Name That Language
- LanguageARC (citizen science portal)
  - Deployed soft release
  - Official release within ~ 1 month
- LanguagePro (teachers, students portal)
  - development in progress
  - transcription tool prototyped
  - otherwise used as staging ground for task types
  - CogHealth, developed with non-NSF funds

Navigation Tips

- Human Subjects Collection
- Privacy/Ethics
- Intellectual property
- Finding an archive
- Choosing a license
- Distribution plan
- The Data Management Plan

Human Subjects Collection 1/2

- Triggered by obtaining data from humans by intervention, interaction
  - Requires a protocol approved by an Institutional Review Board
  - Procedures vary by university – check department, research resources
- Based on the Common Rule
  - Respect for persons, beneficence, justice
  - Focus on vulnerable populations, informed consent, confidentiality
- The Linguist Fieldwork Problem: IRBs as research obstacles
  - Survey suggested otherwise (Bowern 2010)
  - Clinical orientation of some IRBs; aversions to data sharing
- Recent changes to the Common Rule
  - Some studies are exempt from human subjects research requirements, although consent should be obtained to confirm data use & sharing
    - "Benign behavioral interventions"
    - Oral histories exempt, but not ethnographic studies
    - Tribal rights – additional review, approval under tribal laws

Human Subjects Collection 2/2

- Most language-related studies are considered minimal risk and subject to expedited IRB review
- IRB submissions should cover:
  - Study description: collection method (field, crowd, campus)
  - How subjects' informed consent will be obtained
    - Disclosure about use in research and how data will be shared
  - How privacy will be protected (e.g., anonymization) if this is a goal
  - How data will be stored, secured, maintained and shared
- Consent methods
  - Written consent, recorded verbal consent, click-through consent on web-based registration system, consent through action
  - Parent/guardian consent, minor assent
- US open data initiatives promote data sharing: impetus for Common Rule changes

Privacy/Ethics

- No single US data protection/privacy law (as opposed to GDPR)
- Sector-specific solutions: credit reporting, health care (HIPAA)
- Personal identifying information
- Anonymization is a good solution for language-related studies
  - Separate personal identifying information from research data; retain for contact, compensation
  - But threat of re-identification within a data set should be assessed
    - A few data points are enough to re-identify (the dark side of algorithms)
    - Potential retribution in minority language communities
- Some communities/participants want credit for their participation

Intellectual Property

- Who owns a language?
  - Language communities may claim ownership rights
  - May include non-language community data
- Copyright rights in found data collections: web-based text, speech, social media
  - Favorable license (creative commons, etc.)
  - Explicit permission from data provider
  - Fair use (apply with caution!)
  - Website terms of use can derail a collection
    - Browser-wrap terms of use
    - Regulate how material can be used, shared, modified; third party data problem
- Seek guidance from university resources

Finding an Archive

- NSF programs may not recommend a repository
  - DEL program requires DSA archive
- A good choice: archives that follow accepted standards, best practices for digital resources (LDC example)
  - Data quality: independent quality checks for deposited data sets
  - Stability: evidence that the repository is established and likely to remain in existence
  - Discoverability: the ease of the finding the resource
  - Standards for metadata, formats, tools, documentation: enhancing usability
- Costs
  - Funding model based on actual costs
  - Flexibility: sponsor-funded, incremental cost sharing
- FAIR: findable, accessible, interoperable, reproducible
- LDC is a CoreTrustSeal repository – WDS/DSA

Choosing a License

- Confusing license nomenclature
  - “Freely available”
  - Open source – may not mean cost-free
    - Restrictions on derivatives, sharing
    - Viral terms
  - “Research only”
    - Consider how data is used to train language technology systems

- LDC license model
  - Language-related research, education and technology development
    - Includes commercial development by for-profit Consortium members
  - Licensees cannot redistribute resources outside their organization
  - Options: research-only, referrals to data owners for commercial licensing
  - Special licenses to support shared tasks

LDC Distribution Plan Options

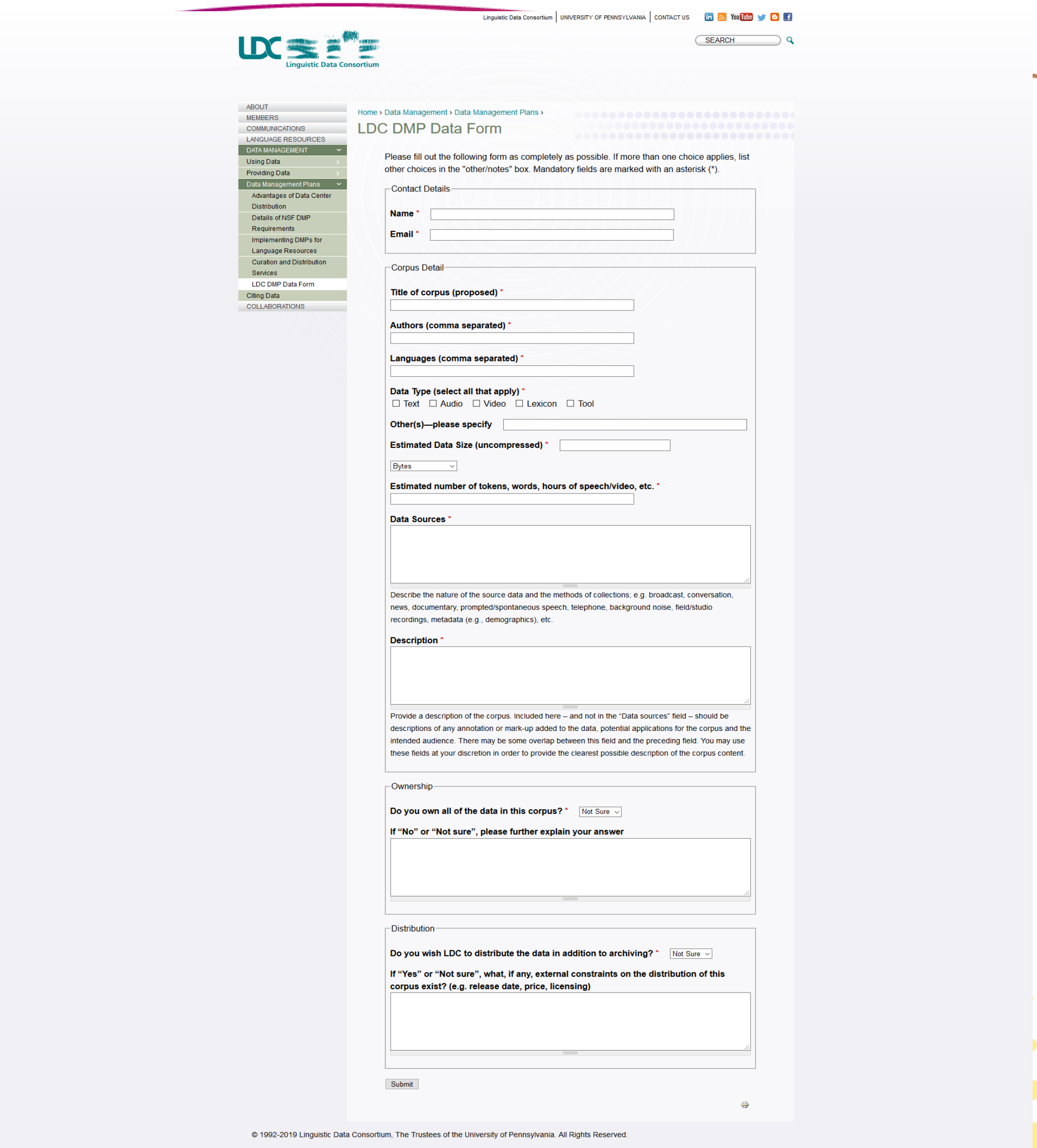
- Archive as host
  - Temporary v. permanent deposit
  - Arrangements regarding curation services: storage, back-up, security
    - Associated costs
  - Hosting for distribution: no distribution services; provided to users "as is"

- Archive as host and distributor
  - Moratorium on distribution pending permanent deposit
  - Distribution-related services: quality control, licensing, customer care
  - Conditions for access: license, fee, other (e.g., organization type)
  - Distribution method
    - Web download, media
    - Archive servers, third party hosted service (e.g., cloud)
    - Associated costs

The Data Management Plan

- The foregoing flows into the DMP
- NSF DMP guidelines for SBE proposals
  - Describe the data expected to be generated
  - Identify a hosting archive
  - Explain data access and sharing
  - Identify other information maintained and shared, such as metadata
  - Pertinent intellectual property rights
  - Ethical and privacy issues
  - Data, tools and documentation formats
  - Plans for archiving and preservation
- Check institutional DMP resources
  - Often maintained by the library
  - Decision trees, tools
- LDC DMP services: DMP submissions form; data sheet

LDC DMP Data Form


Linguistic Data Consortium
UNIVERSITY OF PENNSYLVANIA
CONTACT US
SEARCH
Linguistic Data Consortium
ABOUT
MEMBERS
COMMUNICATIONS
LANGUAGE RESOURCES
DATA MANAGEMENT
Using Data
Providing Data
Data Management Plans
Advantages of Data Center Distribution
Details of NSF DMP Requirements
Implementing DMPs for Language Resources
Curation and Distribution Services
LDC DMP Data Form
Citing Data
COLLABORATIONS
Home › Data Management › Data Management Plans ›

LDC DMP Data Form

Please fill out the following form as completely as possible. If more than one choice applies, list other choices in the "other/notes" box. Mandatory fields are marked with an asterisk (*).

Contact Details

Name *

Email *

Corpus Detail

Title of corpus (proposed) *

Authors (comma separated) *

Languages (comma separated) *

Data Type (select all that apply) *

☐ Text ☐ Audio ☐ Video ☐ Lexicon ☐ Tool

Other(s)—please specify

Estimated Data Size (uncompressed) *

Bytes

Estimated number of tokens, words, hours of speech/video, etc. *

Data Sources *

Describe the nature of the source data and the methods of collections; e.g. broadcast, conversation, news, documentary, prompted/spontaneous speech, telephone, background noise, field/studio recordings, metadata (e.g., demographics), etc.

Description *

Provide a description of the corpus. Included here – and not in the “Data sources” field – should be descriptions of any annotation or mark-up added to the data, potential applications for the corpus and the intended audience. There may be some overlap between this field and the preceding field. You may use these fields at your discretion in order to provide the clearest possible description of the corpus content.

Ownership

Do you own all of the data in this corpus? * Not Sure

If “No” or “Not sure”, please further explain your answer

Distribution

Do you wish LDC to distribute the data in addition to archiving? * Not Sure

If “Yes” or “Not sure”, what, if any, external constraints on the distribution of this corpus exist? (e.g. release date, price, licensing)

Open Clinic Session

- Share your plans, questions, experiences!
- Additional drop-in session today, 2:00-4:00 pm, Oak Room

- Thank you!