Reflections on 30 Years of Language Resource Development and Sharing

Christopher Cieri, Mark Liberman, Sunghye Cho, Stephanie Strassel, James Fiumara, Jonathan Wright

Language Data Center Pre-History

- Mid-1980s: US HLT R&D awakening from the AI winter spurred by the ALPAC report and “W(h)ither Speech Recognition” letter to JASA
- Pierce [28,29] combined laudable goals
  - accurately estimate demand for translation in the market & remove obstructions in supply chain
- with claims that now seem myopic
  - implicitly rejected that information be available regardless of language of publication / readers
  - explicitly rejected the case for voice control of technology
- leading to a 10+ year hiatus in HLT funding at DARPA for example [19]
- Re-awakening included focus on *common task* research management paradigm:
  - multiple teams work in parallel, cooperating and competing
  - well defined, quantifiable goals
  - shared data
  - regular (also frequent) evaluation by neutral party using objective, pre-determined criteria
  - workshop to discuss objectives, challenges, data, approaches, results, evaluation criteria
  - course corrections as needed; virtuous cycle until goals reached or funding spent [19]
  - culture of knowledge, resource sharing attracting research even absent funding [9].

Language Data Center Pre-History

- Early 1990's: Optimism and Recognition of the Critical Importance of Data
  - "useful present-day systems and realistic expectations of progress"
  - "Not even the largest companies can easily afford … data to satisfy their … needs"
  - "… smaller companies and in universities risk being frozen out of the process almost entirely" [18]
  - "growing worldwide awareness of the need for … publicly available common corpora" [14]
  - Existing LRs closely held, unevenly distributed reinforcing schisms, impeding progress
- Solution
  - enable LR sharing at scale; meet current & anticipate future needs.
  - create organization to focus on acquiring, curating and distributing LRs
  - centralize distribution function, technologies, skills to improve quality and reduce cost through scale
  - open RFP; UPenn selected to host, seed funding from DARPA; early support from NSF, NIST

Linguistic Data Consortium

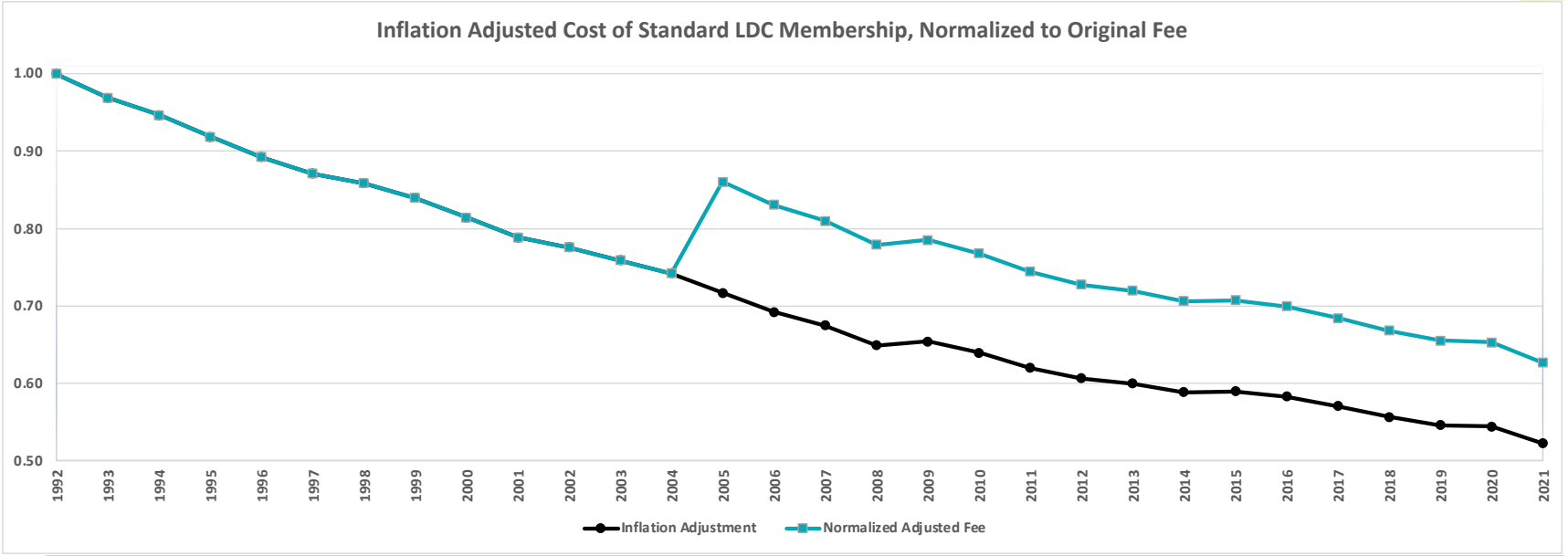
- "*distributing previously created datasets, and funding or co-funding the development of new ones*" [18]
- requirement to become self-supporting through membership, data licensing fees
- 1992-1995: focused exclusively on corpus distribution
- AB with members from the non-profit, government and commercial sectors defined the LDC business model that is still in effect today with small adaptations
- Consortium = a kind of mutual aid society
  - members provide support in the form of membership fees & data contributions
  - receive access to many, many more datasets than any one member could hope to create
  - LDC can also license most corpora but 95% of Members (who embraced Consortium model) report satisfaction [30]
- Catalog >900 corpora in 107 linguistic varieties, including recent additions in Dari, Georgian, Icelandic, Kazakh, Kurdish, Nahuatl, Persian, Pushto, Russian, Turkish Ukrainian, Uzbek, Zulu
- Developed by and/or used within 91 research programs including the following:
  - Large multisite programs sponsored by DARPA, IARPA and other agencies: AIDA, AQUAINT, BEST, BOLT, Communicator, DEFT, EARS, GALE, HARD, HAVIC, Hub4, Hub5-LVCSR, KAIROS, LCTL, LORELEI, Machine Reading, MADCAT, MED, RATS, ROAR, SPINE, TDT, TIDES, Tipster, Transtac
  - NIST evaluation campaigns: LRE, ACE, MT, OpenHaRT, OpenSAD, OpenSAT, RT, SRE, TAC/KBP, TREC, TRECVid
  - community organized evaluations: CoNLL, SemEval, SIGHAN

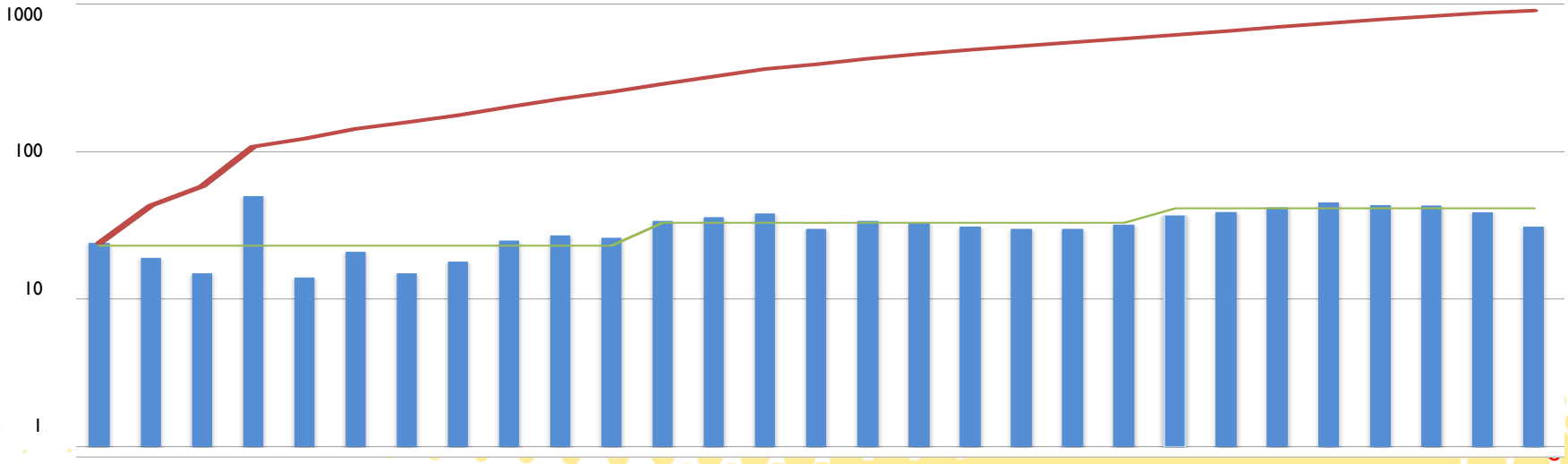
Data Distribution

- to support 75 target applications, the most common of which are:

entity, event, relation extraction & coreference
handwriting recognition
information retrieval
knowledge base population
language identification
language modeling
machine translation
parsing, POS tagging & other NLP
pronunciation modeling
question-answering
semantic role labelling
sentiment detection
speaker diarization, identification
speech activity detection
speech recognition
summarization

Data Distribution





Data Collection & Annotation: Recent Efforts

- Locally implemented, innovated methods for collecting text from:
  - news sources, journals, financial and biomedical documents
  - internet sources including newsgroups, blogs, microblogs, comment threads and discussion forums
  - text interactions via email, chat and SMS
  - scans or images of documents containing printed or handwritten text or both.
- and audiovisual data from:
  - broadcast news and conversation, podcasts
  - conversational telephone speech
  - lectures, interviews, meetings, field interviews
  - read, prompted & task oriented speech, role play
  - speech in noise
  - web video and directly contributed amateur video
  - animal vocalizations
  - digitized analog media including interviews in a variety of tape formats
  - two way radio speech characterized by severe channel noise

Data Collection & Annotation: Recent Efforts

- requiring creation of new hardware software solutions including:
  - satellite downlink node on the Voice of America network to collect multilingual broadcast news [10]
  - fully automated platform for collecting broadcast audio, video & processing through ASR & MT
  - subsequent miniaturization of platform to enable outsourcing of collection to international partners [34]
  - platforms to collect telephone calls; also miniaturized, modularized, deployed and managed remotely
  - interfaces to let users upload SMS archive, remove sensitive messages before contributing
  - digitization station that can accept input from most common legacy analog media players
  - platform to broadcast and receive, and optionally degrade, clean audio for which we had transcripts
- and new annotation tools
  - often highly customized, task-specific local tools [20]
  - but shifting over the past 10 years, to web-based tools, especially LDC webann [35]
  - customized front ends, that appear to users as distinct tools
  - same underlying database schema & project management umbrella
  - allowed remote work when necessary, especially starting March 2020
  - expanded beyond the needs of common tasks programs evolved into Universal Annotator (UA)
  - basis for NIEUW portals, webtrans, a comprehensive audio transcription application
  - used for O(1000 tasks, 1000 users, 10,000,000 annotations)

Data Collection & Annotation: Recent Efforts

- We Can Talk: SR corpus of speech from telephone and video for >200 multilingual speakers of Cantonese plus Mandarin and/or English
- CAMIO (Corpus of Annotated Multilingual Images for OCR): ~70,000 text images, 35 languages, 24 unique scripts, most annotated for text localization yielding >2.3M bounding boxes, transcripts ~16,000 images in 13 languages, yielding >2.4M tokens.
- KASET (Kurmanji and Sorani Speech Transcripts): ~350 hours of broadcast & telephone speech in two Kurdish varieties, plus transcripts of ~ 65 hours
- COSINE (Corpus of Speech in Natural Environments): ~500 hours of audio from multiple genres in Indonesian, Korean, Mandarin, Modern Standard Arabic (MSA), Russian, plus transcripts for ~300 hours & translations for ~75 hours
- AIDA (Active Interpretation of Disparate Alternatives): multimedia documents in Russian, Ukrainian, Spanish, English covering current event scenarios, with focus on disinformation and conflicting claims, ERE annotated with cross-document coreference, and annotation of the relations among events and claims
- KAIROS (Knowledge-directed Artificial Intelligence Reasoning Over Schemas): ~15M-document Schema Learning Corpus, supporting induction of high-level representations of complex events across many domains. The eval set includes documents related to IED attacks, disease outbreaks, etc with annotation of EREs and temporally-ordered components of each incident

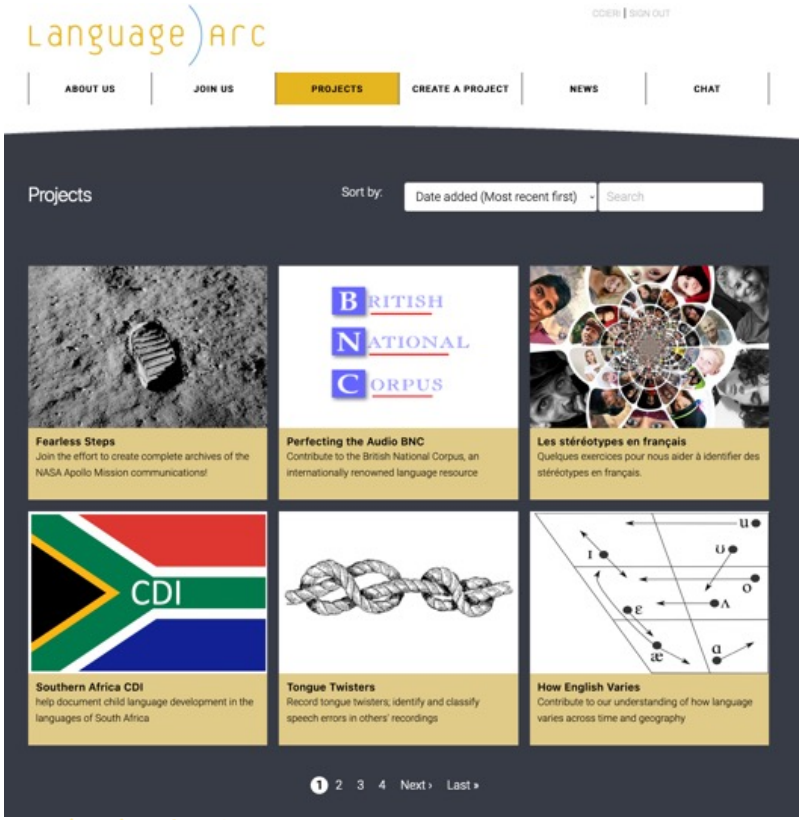
Data Collection & Annotation: Novel Incentives

- Recognizing that LRs remain in short supply after years of concentrated effort, we have begun to work on novel incentives in data collection, with NSF support.
  - online communities show that people will spend time providing language data given sufficient motivation and appropriate tools: Wikipedia, Project Gutenberg LibriVox but also social media sites

- 1st effort: NameThatLangage game eliciting judgements of language spoken in short audio clips
  - input >=80 clips for each of 13 languages plus 5400 clips suspected to be in one of 9 languages
  - to date: results of ~720,000 HITs presented to ~46,000 unique player IDs (86% usable)
  - aggregation of (mostly) non-expert player guesses predicts correct answer >98%
  - when player pool does not converge, clip is not in the suspected language 96% [11]

Data Collection & Annotation: Novel Incentives

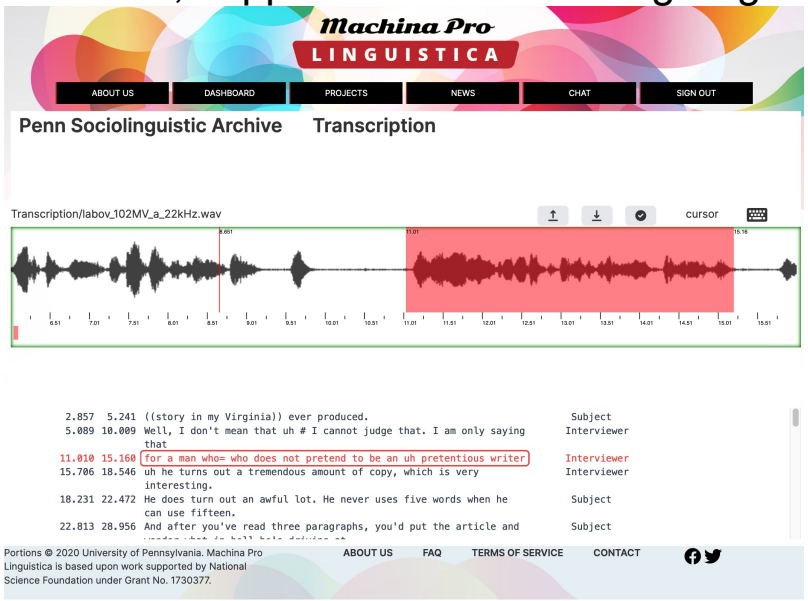
- 2nd effort: LanguageARC Citizen Linguists contribute to data intensive projects [13]
- Documenting Xi'an Guanzhong
  - ~59 native speakers recorded themselves naming objects in 622 images from the MultiPic corpus [12]
  - selected for familiarity to people living in China
  - 34,729 recordings audited each for audio quality, use of the target variety
- Fearless Steps
  - LDC & University of Texas, Dallas eliciting transcripts and diarization of extremely challenging audio in the Fearless Steps corpora [16]
  - communications of NASA Apollo space missions
- Novel Incentives WS, Saturday AM
  - Les stéréotypes en français
  - From Cockney to the Queen



- Principal incentives: opportunities to learn, contribute to social good or reinforce local pride, for example by documenting an under-represented variety.

Data Collection & Annotation: Novel Incentives

- 3rd: Machina Pro Linguistica for linguists and other language professionals motivated to contribute as a way to develop professional skills, supplement their learning or gain access to resources in exchange.



- Penn Sociolinguistic Archive
  - >5800 recordings collected over >50 years by Professor William Labov & students
  - all tool building capabilities of LanguageARC
  - full implementation of LDC webtrans, used within LDC for recent transcription projects

- SpeechBiomarkers
  - LDC presence of researchers working on novel incentives and the language of clinical interactions, led to this portal
  - volunteers do brief exercises (e.g. picture description) to help establish baseline population performance for research involving clinical populations.

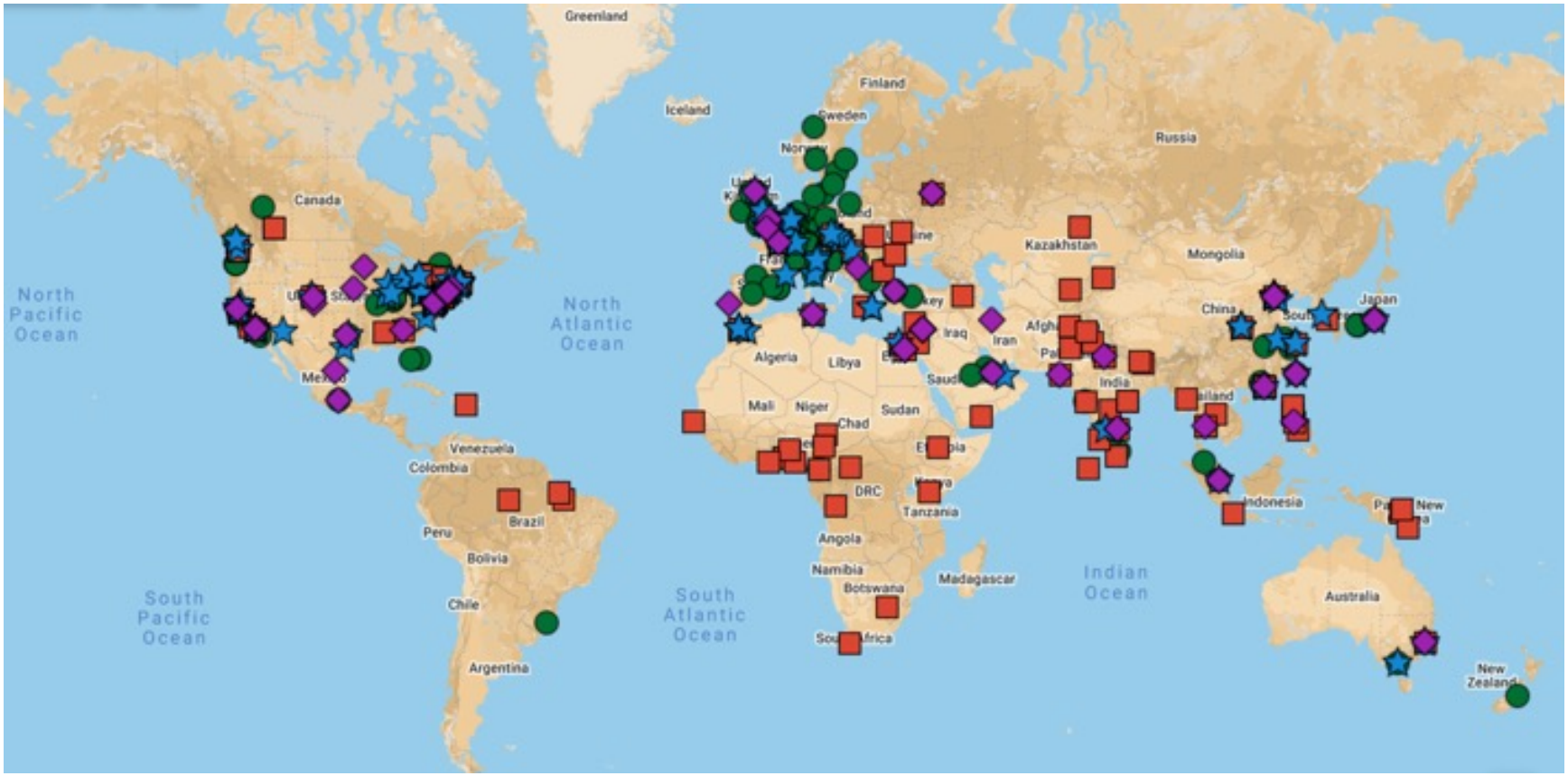
- Novel Incentives let us expand beyond limits of funded programs, document & provide data for technology development in language varieties heretofore under-served

Data Intensive Research

- Data planning, collection, annotation, critical component of research
- Corpus creation for common task program
  - support program needs, cognizant of the tensions present among different stakeholders
  - LDC system developers (isolated from data team) provided insight in their issues [31]
- Other roles could include technology evaluation [8] such as the DIHARD robust diarization evaluations
- Direct Research
  - improving data preparation, analytic pipeline in clinical research with CAR, FTDC, Northwell Health
  - speech/language features distinguish autistic children from neurotypical children in clinical interviews [24, 25] but also when no clinical expert is leading conversation [1]
  - autistic girls use pause fillers (“um”) more like neurotypicals than do autistic boys during natural conversations possibly contributing to a strategy of “linguistic camouflage” [26]
  - speech/language features in brief picture description data useful in better understanding neurodegenerative conditions:
    - frontotemporal dementia [2, 21]
    - amyotrophic lateral sclerosis spectrum disorders [23]
    - progressive supranuclear palsy and corticobasal syndromes [27]
    - Alzheimer’s dementia [4]

- NameThatLanguage, LanguageARC and MachinaProLinguistic are outcomes of the NIEUW project. The Linguistic Data Consortium and the University of Pennsylvania acknowledge the generous support of the US National Science Foundation via the Computer and Information Science and Engineering Directorate's Research Infrastructure program, grant 1730377.

LDC Data Providers



Linguistic Data Consortium

LDC Data Users

Fine

Linguistic Data Consortium

References

1. Cho, S., Liberman, M., Ryant, N., Cola, M., Schultz, R., and Parish-Morris J. (2019). Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations. Proc. Interspeech, 2513-2517.
2. Cho, S., Nevler, N., Ash, S., Shellikeri, S., Irwin, D., Massimo, L., Rascovsky, K., Olm, C., Grossman, M., and M. Liberman. (2021a). Automated analysis of lexical features in frontotemporal degeneration. Cortex 137: 215-231.
3. Cho, S., Cousins, K., Shellikeri, S., Ash, S., Irwin, D., Liberman, M., Grossman, M., and Nevler, N. (in press). Lexical and acoustic speech features relating to Alzheimer's disease pathology. https://www.medrxiv.org/content/10.1101/2021.09.27.21264148v2 (preprint on medRxiv)
4. Cho, S., Nevler, N., Shellikeri, S., Parjane, N., Irwin, D., Ryant, N., Ash, S., Cieri, C., Liberman, M., and Grossman, M. (2021). Lexical and acoustic characteristics of young and older healthy adults. Journal of Speech, Language and Hearing Research 64(2):302-314.
5. Cho, S., Shellikeri S., Ash, A., Grossman ,M., Nevler, N., and Liberman, M. (2020). Automatic classification of primary progressive aphasia patients using lexical and acoustic features. Proc. 12th Language Resources and Evaluation Conference 2020 workshop on Resources and Processing of linguistic, para-linguistics, and extra-linguistic data from people with various forms of cognitive, psychiatric, and/or developmental impairments (RaPID-3).
6. Cho, S., Shellikeri, S., Ash, S., Liberman, M., Grossman, M., and Nevler, N. (2021c). Automatic classification of AD versus FTLD pathology using speech analysis in a biologically confirmed cohort. Alzheimer's & Dementia: the Journal of the Alzheimer's Association 17(S5).
7. Cho, S., Nevler, N., Parjane, N., Cieri, C., Liberman, M., Grossman, M., and Cousins, K. (2021d). Automated analysis of digitized letter fluency data. Frontiers in Psychology 12, 654214.
8. Choukri, K. Mapelli, V., Mazo, H., and Popescu, V. (2016). ELRA Activities and Services. Proc. 10th International Conference on Language Resources and Evaluation: 463-468.
9. Church, K. W. (2017). Emerging trends: A tribute to Charles Wayne. Natural Language Engineering, 24(1): 155–160.
10. Cieri, C. and Liberman M. (2000) Issues in Corpus Creation and Distribution: The Evolution of the Linguistic Data Consortium, Proc. 2nd International Conference on Language Resources & Evaluation.

References

11. Cieri, C., Fiumara, J., and Wright, J. (2021). Using Games to Augment Corpora for Language Recognition and Confusability, Proc. 22nd Annual Conference of the International Speech Communication Association (Interspeech), August 30-September 3.
12. Duñabeitia, J.A, Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., and Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. Quarterly Journal of Experimental Psychology. 71(4):808-816.
13. Fiumara, J., Cieri, C., Wright, J., and Liberman, M. (2020). LanguageARC: Developing Language Resources Through Citizen Linguistics in Proc. 12th Edition of the Language Resources and Evaluation Conference (LREC). CLLRD Workshop: Citizen Linguistics in Language Resource Development. Marseille, May 11-16.
14. Gibbon, D., Moore, R., Winski, R. eds. (1998). Spoken Language Reference Materials. Vol. 4. Walter de Gruyter.
15. Hutchins, W. J. (2001). Machine translation over fifty years, Histoire, Epistemologie, Langage, 22(1):7-31.
16. Joglekar, A, Seyed, O. S., Chandra-Shekar, M., Cieri, C., and Hansen, J.H.L. (2021). Fearless Steps Challenge Phase-3 (FSC P3): Advancing SLT for Unseen Channel and Mission Data Across NASA Apollo Audio in Proc. 22nd Annual Conference of the International Speech Communication Association (Interspeech), August 30-September 3.
17. Krell, R., Tang, W., Hänsel, K., Sobolev, M., Cho, S., Berretta, S., and Tang, S. (2022). Lexical and acoustic correlates of clinical speech disturbance in schizophrenia. In Sharban-Nejad, A., Michalowski M., and Bianco, S. (eds.), AI for Disease Surveillance and Pandemic Intelligence. W3PHAI 2021. Studies in Computational Intelligence, vol. 1013. Springer, Cham.
18. Liberman, M. and Godfrey, J. (1993). The Linguistic Data Consortium. In Chen, Keh-Jiann, Chu-Ren Huang, Proc. ROCLing Computational Linguistics Conference VI, Nantou, Taiwan, September. Association for Computational Linguistics and Chinese Language Processing (ACLCLP).
19. Liberman, M. and Wayne, C. (2020). Human Language Technology, AI Magazine , 41(2):22-35.
20. Maeda, K., Mazzucchi, A. and Cieri, C. (2011) Technical Infrastructure Supporting Large-scale Linguistic Resource Creation in Olive, J. Christianson, C. and McCary, J. eds, Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, New York, Springer.

References

21. Nevler, N., Ash, S., Jester, C., Irwin, D., Liberman, M., and Grossman, M. (2017). Automatic measurement of prosody in behavioral variant FTD. Neurology 89:1-7.

22. Nevler, N., Ash, S., Irwin, D., Liberman, M., and Grossman, M. (2019). Validated automatic speech biomarkers in primary progressive aphasia. Annals of Clinical and Translational Neurology 6:4-14.

23. Nevler, N., Ash, S., McMillan, C., Elman, L., McCluskey, L., Irwin, D., Cho, S., Liberman, M., and Grossman, M. (2020). Auomated analysis of natural speech in amyotrophic lateral sclerosis spectrum disorders. Neurology 95(12): e1629-e1639.

24. Parish-Morris, J., Liberman, M., Ryant, N., Cieri, C., Bateman, L., Ferguson, E., and Schultz, R. (2016a). Exploring autism spectrum disorders using HLT. Proc. 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 74-84.

25. Parish-Morris, J., Cieri, C., Liberman, M., Bateman, L., Ferguson, E., and Schultz, R. (2016b). Building language resources for exploring autism spectrum disorders. Proc. 10th International Conference on Language Resources and Evaluation, 2100-2107.

26. Parish-Morris, J., Liberman, M., Cieri, C., Herrington, D., Yerys, B., Bateman, L., Donaher, J., Ferguson, E., Pandey J., and Schultz, R. (2017). Linguistic camouflage in girls with autism spectrum disorder. Molecular Autism 8, 48.

27. Parjane, N., Cho, S., Ash, S., Cousins, K., Shellikeri, S., Liberman, M., Shaw, L., Irwin, D., and Grossman, M. (2021). Digital speech analysis in progressive supranuclear palsy and corticobasal syndromes. Journal of Alzheimer's Disease 82:33-45.

28. Pierce, J. R., J. B. Carroll, E. B. Hamp, D. G. Hays, C. F. Hockett, A. G. Oettinger, and A. Perlis. (1966). Language and Machines — Computers in Translation and Linguistics. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC.

29. Pierce, J. R. (1969). Whither Speech Recognition? The Journal of the Acoustical Society of America 46:1049.

30. Reed, M., DiPersio, D., and Cieri, C. (2008). The Linguistic Data Consortium Member Survey: Purpose, Execution and Results. Proc. 7th International Conference on Language Resources and Evaluation (LREC), Marrakech, May 28-30.

31. Schultz, J.M. and Liberman, M. (1999). Topic Detection and Tracking using idf-weighted Cosine Coefficient, Proc. DARPA Broadcast News Workshop.

References

32. Tang, S., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R., Bhati, M., Sedoc, J., and Liberman, M. (2021) Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. NPJ Schizophrenia 7, 25.

33. van den Heuvel, H, Oostdijk, N., Rowland, C., and Trilsbeek, P. (2020), The CLARIN Knowledge Centre for Atypical Communication Expertise, Proc. 12th Language Resources and Evaluation Conference, pp. 3312—3316.

34. Walker, K., Caruso, C. and DiPersio, D. (2010) Large Scale Multilingual Broadcast Data Collection to Support Machine Translation and Distillation Technology Development, Proc. 7th International Conference on Language Resources and Evaluation.

35. Wright, J., Griffitt, K., Ellis, J., Strassel, S., Callahan, C. (2012) Annotation Trees: LDC's Customizable, Extensible, Scalable Annotation Infrastructure, Proc. 8th International Conference on Language Resources and Evaluation.