

Development and Distribution of Clinical Speech and Language Datasets

Denise DiPersio, Associate Director

dipersio@ldc.upenn.edu

- ◆ About LDC
- ◆ Language Resources and Clinical Data Sets
- ◆ Source Data for Clinical Resources
- ◆ Clinical Data Available From LDC
- ◆ Clinical Resources Overview
- ◆ Sharing Data Through LDC

- ◆ The Linguistic Data Consortium (LDC) is a non-profit, hosted at the University of Pennsylvania, that creates and distributes language data to universities, laboratories, companies and libraries for language-related education, research and technology development
- ◆ LDC's Catalog has over 750 holdings in more than 90 languages. Three to four data sets – text, speech, video and lexicons – are released monthly
- ◆ Current activities include:
 - Data collection and annotation
 - Software development
 - Research and project support
 - Archiving, publishing and distributing corpora



- ◆ Developing language resources in general requires attention to matters outside the research question such as intellectual property, privacy and ethical issues relating to human subjects collections
- ◆ Corpora in the clinical domain, given their nature, may involve multiple legal and regulatory issues
 - Researchers and regulators may view these as a deterrent to development or data sharing
- ◆ Useful clinical data sets are possible, exist and are accessible
- ◆ Source data implications in clinical resources
 - Copyrighted material – e.g., biomedical journal abstracts and articles
 - Data collection directly from humans – e.g., speech
 - Personal health information – e.g., patient medical records

- ◆ Biomedical journal abstracts and articles
 - Journal abstracts and articles traditionally subject to copyright rights (usually held by the journal)
 - Open science/data movements and funder initiatives have resulted in open access resources
 - PubMed <https://www.ncbi.nlm.nih.gov/pubmed/>
 - US National Library of Medicine, National Institutes of Health
 - 28 million citations
 - Biomed Central <https://www.biomedcentral.com/>
 - Part of Springer Nature – access to journals and other services for researchers
- ◆ Collecting data directly from humans – speech example
 - Regulated by institutional review boards, ethics committees
 - Respect for persons, beneficence, justice
 - Transparency/consent from subjects; do no harm; fairness
 - Parent/guardian assent, minor assent for minor subjects
 - Submission/approval by relevant board/committee in advance of research

- ◆ Personal health information
 - US personal health information protected by Health Insurance Portability and Accountability Act of 1996 (HIPAA)
 - Study subjects agree to waive HIPAA protections for research
 - De-identified records may be used for later research based on prior waiver
- ◆ Other data protection laws (GDPR, national laws)
- ◆ Carrying through to distribution
 - If personal identifying information is collected for a study, it is stored separately from the collected data
 - Personal identifying information from study subjects is not included in final data sets; unique identifiers for subjects used in data (audio files) and metadata
 - Personal health information is de-identified

- ◆ PennBioIE CYP 1.0 LDC2008T20 (LDC, CHOP)
- ◆ PennBioIE Oncology 1.0 LDC2008T21 (LDC, CHOP)
 - PubMed journal abstracts on the inhibition of cytochrome P450 enzymes (CYP) and cancer (oncology)
 - Annotation for biomedical entity, token, POS, syntax, semantic relations
- ◆ BioProp Version 1.0 LDC2009T04 (Academia Sinica)
 - Propbank annotations for 500 English biomedical journal abstracts from the GENIA Treebank (stand-off annotations)
- ◆ TORGO Database of Dysarthric Articulation LDC2012S02 (University of Toronto)
 - 23 hours of English read and elicited speech from speakers with cerebral palsy, amyotrophic lateral sclerosis and a non-dysarthric control group
 - Aligned acoustic data and measured 3D articulatory features
- ◆ Available under standard LDC license terms
- ◆ LDC work in progress

- ◆ Talkbank <https://talkbank.org>
 - Repositories in 14 areas maintained by CMU
 - Data sets of speech and multimodal interactions for dementia, right hemisphere disorders, traumatic brain injury, aphasia
 - Must apply for access to data
- ◆ Curated List, Machine Learning and NLP Resources for Healthcare
 - Moved to github in 2017
https://github.com/isaacmg/healthcare_ml#datasets
 - Licenses may apply; open access for some
 - Datasets include:
 - MIMIC critical care database – de-identified health data for 40k critical care patients (MIT Lab for Computational Physiology)
 - Clinical case reports for machine reading comprehension
 - EBM-NLP – 5000 annotated medical journal abstracts
 - A Large Corpus for Question Answering on Electronic Medical Records
 - OncoKB (oncology knowledge base)

- ◆ University of Texas Center for Computational Biomedicine
<https://sbmi.uth.edu/ccb/resources/> (licenses may apply to some)
 - Resources by project: Deep Learning Package, Clinical Language Annotation, Modeling and Processing Toolkit
 - Mainly tools, some data sets available


- ◆ Health NLP shARe (Shared Annotated Resources)
https://healthnlp.hms.harvard.edu/share/wiki/index.php/Main_Page#Getting_Access_to_the_ShARe_Corpus_and_Gold_Standard_Annotations
 - US universities – multiple collaboration
 - Aim to develop annotation standards and toolkits for clinical material
 - shARe Corpus and Gold Standard Annotations – source is MIMIC data
 - Data User Agreement

- ◆ University of Minnesota Institute of Health Informatics NLP/IE Resources <https://healthinformatics.umn.edu/research/nlpie-group/nlpie-resources>
 - Tools, platforms, data for clinical material including electronic health records
 - Licenses may apply to some resources
- ◆ COLING 2016 Clinical NLP Workshop <https://text-machine-lab.github.io/ClinicalNLP2016/resources.html>
 - Links to clinical resources
 - Includes MIMIC, i2b2, shARe
 - SemEval task sets
 - Analysis of Clinical Text, Clinical TempEval
 - NTCIR MedNLPDoc shared task data
 - Shared tasks processing Japanese medical records for named entity recognition, term normalization, disease identification
 - Data licenses may apply to some resources

- ◆ Max Planck Institute for Psycholinguistics <https://www.mpi.nl/>
 - Coginst: video recordings of medical students in problem-based learning course discussions
- ◆ CLARIN/LINDAT/ELRA/META-SHARE
 - Khoresmoi Query Translation Test Data 1.0
 - Medical search short queries from the public and medical experts
 - Translation between Czech, English, French, German
 - CLEFeHealth 2013, 2014
 - Shared tasks to evaluate information retrieval to address questions from patients reading clinical reports
 - Medical documents (from Khoresmoi), queries, relevance assessments participant results
 - Hungarian Medical Speech Database
 - Pathological speech from individuals with speech disorders
- ◆ CONCLUSION: More text than speech?

- ◆ Strong global network
 - Monthly newsletter announcing new publications reaches over 8000 recipients
- ◆ Archiving and curation best practices followed
 - LDC Catalog is a trustworthy data repository
 - Quality checks, metadata schema, rights management, content delivery, permanent archive, secure storage
- ◆ Expertise in publishing and licensing data
 - Over 150,000 copies distributed under various license arrangements
- ◆ Non-exclusive rights to LDC
 - Consistent with mission to provide broad access to data





Questions? Thank you!