Cornell University
Department of Linguistics

Publish Faster
With Cornell's LDC Language Corpora

A High-Quality and ***FREE*** Resource For

Linguistics

&

Natural Language Processing Research

Want To Publish Faster?

- Then Use Cornell's database of Linguistics Data Consortium (LDC) corpora !
- >800 high-quality text, audio, and video corpora in more than 60 languages
- **Free** to Cornell Linguistics & Natural Language Processing Researchers!
  - Students, Staff, Faculty, Postdocs, Visiting Scholars

FREE Access to >800 high-quality language corpora

- **810** LDC and nine non-LDC corpora as of Feb 9, 2021
- **LDC Corpora (many languages & situations)**
  - 516 standard-license LDC corpora
  - 126 special-license LDC corpora
  - 168 experimental corpora
- **We also have Non-LDC Corpora (some are licensed)**
  - Historical American English (for evaluation only)
  - Spontaneous Japanese
  - Google 1, 2, and 3 ngrams
  - Project Gutenberg English Language Books (downloaded Nov. 2-3, 2016)
  - Buckeye Corpus, 2nd Release (Conversational English)
  - ***BYU Corpora (TV Corpus, Corpus of American Soaps, Corpus del Espanol)***
  - French Treebank

What kind of corpora?

- Text, Audio, and Video
- Multilingual (> 60 languages)
- Written, spoken, signed language (with annotations)
- Sourced from:
  - Newspapers & Newswires
    - (the New York Times Annotated Corpus is ***very*** popular!)
  - Telephone conversations
  - Interviews & meetings
  - Broadcast programming
  - Transcripts, websites, books, periodicals, etc.

Speech Audio Corpora: Not Just Recorded Speech!

- Sound-based Corpora contain sound files (various formats), and ***may*** contain one or more of the following:
  - Word transcriptions
  - Phonetic transcription
  - Annotations
  - Morphology
  - Tokenization

- There are special corpora for study of:
  - Emotion
  - Prosody
  - Anomaly analysis
  - Elephant vocalizations

LDC Corpora are "gold-standard" corpora

- Distributed by the Linguistics Data Consortium
  - An open consortium of universities, libraries, corporations, & govt. research labs
  - Started in 1992 - Creates & distributes language resources
  - LDC has a Core Trust Seal as a trustworthy, secure data repository
  - Issues 36-40 new corpora yearly
  - Every corpora undergoes quality control tests
    - Strict requirements on data quality, organization, metadata, & documentation

LDC Corpora Are Used Worldwide

- LDC has distributed > 140,000 corpora since 1992
  - 36-40 new corpora/year
  - >18,000 citations
- LDC's monthly newsletter has > 22,000 readers
  - Announces & describes all new corpora issued that month.
  - **Sign up NOW!**

Cornell's LDC Holdings

- We have most LDC corpora issued since 1992.
  - Cornell has everything from the last 12 years
  - Contact Bruce McKee (bwm55@cornell.edu) to see if we have a particular corpus
  - If we don't have it, then must purchase "a la carte" from LDC (Cornell gets half-off the list price)
- Individual Corpora size ranges from <1MB to 1.2TB (compressed!)

Linguistics, CS, & InfoSci faculty fund Cornell's LDC Membership

- **Linguistics**
  - Mats Rooth
  - Sam Tilsen
- **Computer Science**
  - Yoav Artzi
  - Claire Cardie
  - Lillian Lee
- **Information Science**
  - David Mimno
- **Yearly membership renewal coordinated by Gretchen Ryan**
  - Gretchen is the Linguistic Department's Administrator
  - AKA "The Herder of Cats"

Limits on Cornell LDC usage

- Only for non-commercial Linguistics & Natural Language Processing Research at Ithaca and at the NYC campuses (Cornell Tech, Weill Medical)
- Typical users are Linguistics, CS, and InfoSci Faculty & Grad Students
- Exceptions made for researchers in other Cornell departments
  - Must be related to Linguistics or Natural Language Processing
  - Dr. Mats Rooth, Dr. Claire Cardie decide on access for LDC corpora
  - Dr. Martin van Schijndel decides access on BYU corpora
- Researchers in these other Cornell departments have arranged research access:
  - Statistical Sciences
  - Electrical & Computer Engineering
  - Psychology
  - Human Ecology
  - The Johnson & Hotel Schools

All LDC Corpora are stored on a Linguistics Department server

- Currently ~7 TB of corpora files (compressed)
- Corpora distributed via Cornell Box service
- Certain faculty/grad students can get :
  - A corpora server account + VPN login
  - Unlimited access to standard license corpora

A Word on LDC Licenses…

Standard LDC License (most corpora):

- Signed by Dr. Mats Rooth for Cornell
- You must co-sign, good for all standard license LDC corpora

Special LDC Licenses (some corpora):

- Signed by Dr. Mats Rooth or other faculty for Cornell
- You must co-sign – license only applies to that corpora
- Each special license is unique – ***read the details – they can affect your publications!***

Experimental Corpora (DEFT agreement):

- DEFT = Deep Exploration and Filtering of Text
- *Not listed on the LDC website*
- Only available through Dr. Claire Cardie (Computer Science)

Freedom *and* Responsibility*

(the four *"Can'ts"* of LDC corpora)

1. ***Can't*** be used for any commercial work.
2. ***Can't*** be shared with colleagues outside Cornell.
3. ***Can't*** be shared with Cornell students, faculty, staff, post-docs, or visiting scholars who have not signed the LDC agreement(s).
4. ***Can't*** take corpora with you after you graduate or end your Cornell employment.

---

* "The Cornell Tradition: Freedom and Responsibility", Dr. Carl Becker's Address on the 75th anniversary of the signing of the Cornell University Charter, April 27, 1940

I want to publish faster!

how do I get corpora access!?

- Visit this page on Cornell's Confluence Wiki:
  - "How To Access LDC (Linguistic Data Consortium) Corpora"
- Different procedures for different people:
  - Faculty active in Linguistics & NLP research
  - Other faculty, staff, grad students & visiting scholars
  - ***"Extremely motivated graduate students"***
- **Everyone** must co-sign licenses & post them to Confluence
- ***Faculty members must supervise Student, Staff, Postdoc, and Visiting scholar research!***

Fast Track Corpora Access

(Only for Linguistics, CS, and InfoSci Students)

- For students working with Linguistics faculty or the following four CS & Information Science faculty members
  - Yoav Artzi
  - Claire Cardie
  - Lillian Lee
  - David Mimno

- Faculty member E-mails Bruce McKee (bwm55@cornell.edu) with:
  - List of required corpora
  - One-sentence description of the research
  - State that the faculty member will supervise the student's research
  - The faculty member's e-mail will be cc'd to Dr. Mats Rooth as a record

"Extremely Motivated Graduate Students"

- Per Dr. Mats Rooth, defined as:

"Students who want to actively explore the corpora database and try new things"

- Such students can get ***unlimited access*** to Cornell's standard license LDC corpora

- Who decides whether you are ***"extremely motivated"?***
  - Dr. Mats Rooth (Linguistics) or Dr. Claire Cardie (CS), in consultation with your faculty sponsor.

- Talk to your advisor today!

The Corpora delivery process…

- Bruce McKee provides corpora access after:
  - Faculty member e-mails approval for corpora access
  - All co-signed license(s) uploaded to Confluence
- Also contact Bruce for:
  - VPN access to corpora server accounts
  - Access to LDC corpora for Compling courses
- Corpora distribution is usually within 24 hours (M-F)
  - Usually a compressed tar/gzip download from Cornell Box
  - Windows Users need the free 7-Zip archiver/decompressor utility

Trends in LDC Language Corpora…

- Machine Understanding/Deep Learning of Speech (**HUGE!**)
  - **LDC has many "parallel text" corpora** – English documents/sentences/words vs. the translation in other languages – **very** useful for training deep learning networks.
  - **IARPA Babel Program**
    - Goal: Generate speech transcription system ***in one week*** for any new language
    - Creating many new LDC corpora in neglected languages
      - Zulu, Swahili, Tamil, Tagalog, Assamese, Haitian Creole, etc.
  - **DARPA DEFT Program – Deep Exploration & Filtering of Text**
    - Automated, deep natural-language processing
    - Many new experimental corpora
- Also check out LDC's collaborative projects

Explore ***YOUR*** research interests ***NOW***!
(Audience Participation Exercise…)

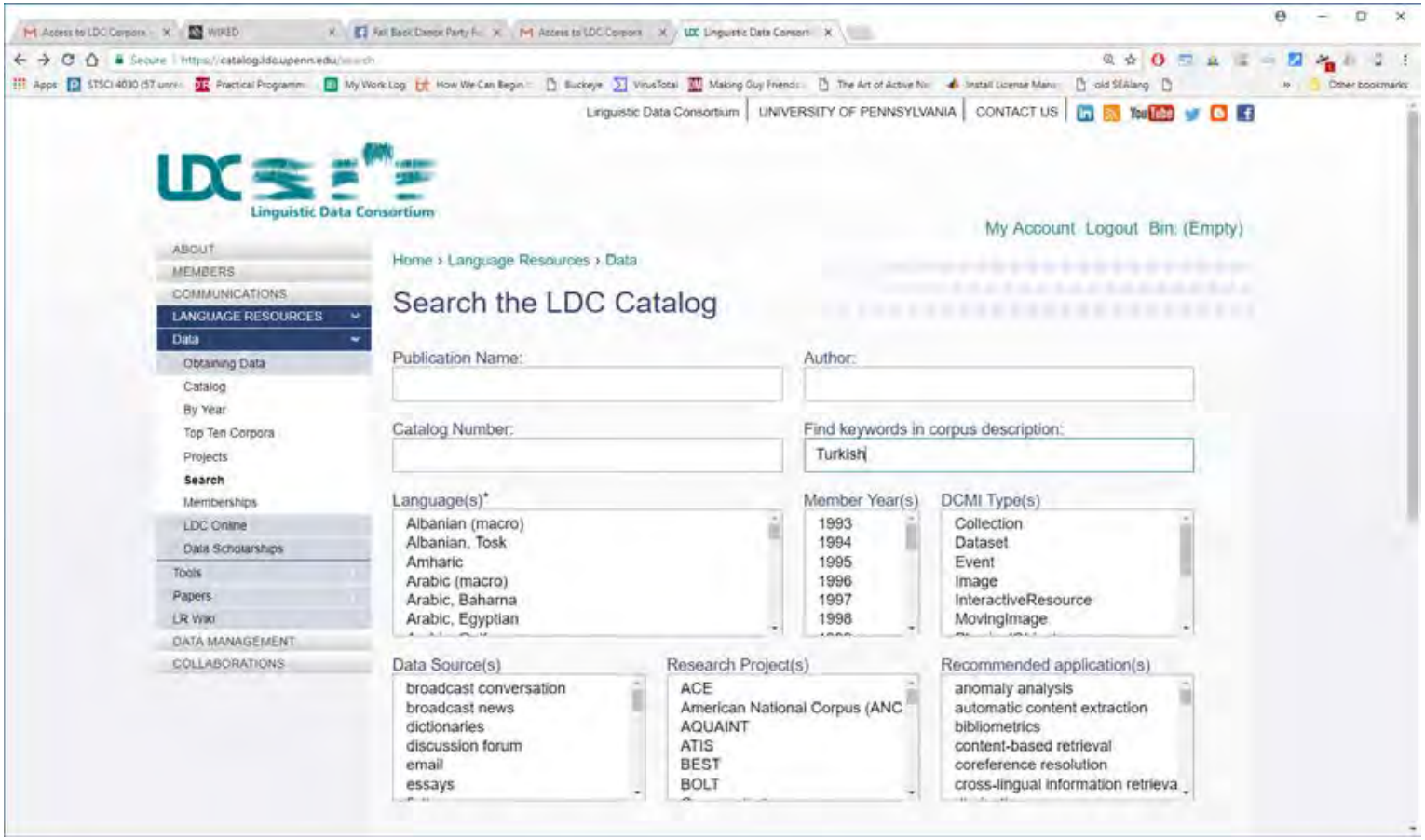
- ***Search the LDC*** Or via a Google Search – for example:
  - site:https://catalog.ldc.upenn.edu Korean
  - site:https://catalog.ldc.upenn.edu Khmer

- **Sign up for the LDC newsletter**

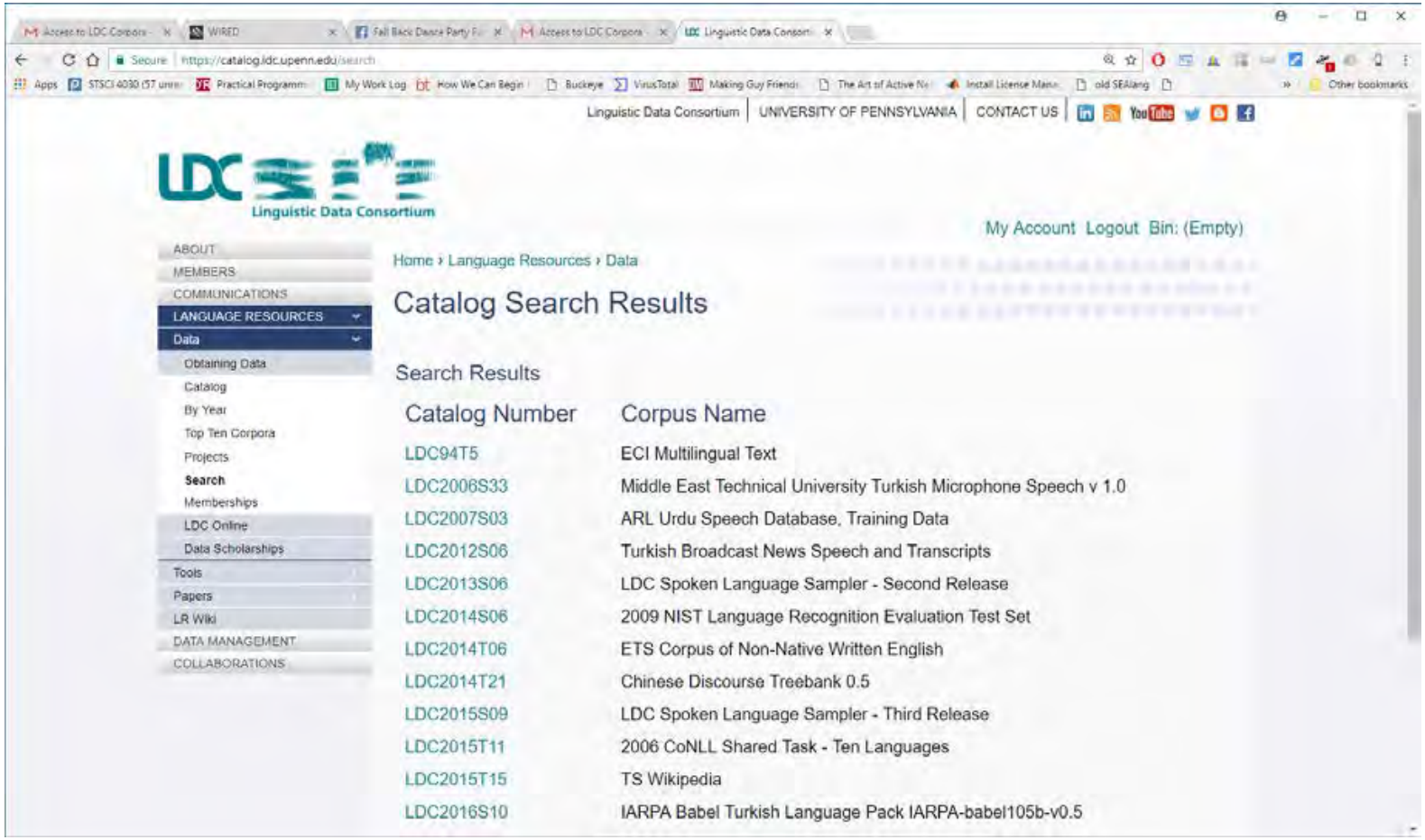
- **Check out the Free LDC-developed Tools -**
  (Includes a tool to convert NIST SPHERE files to other audio formats)

LDC Corpora Search Page

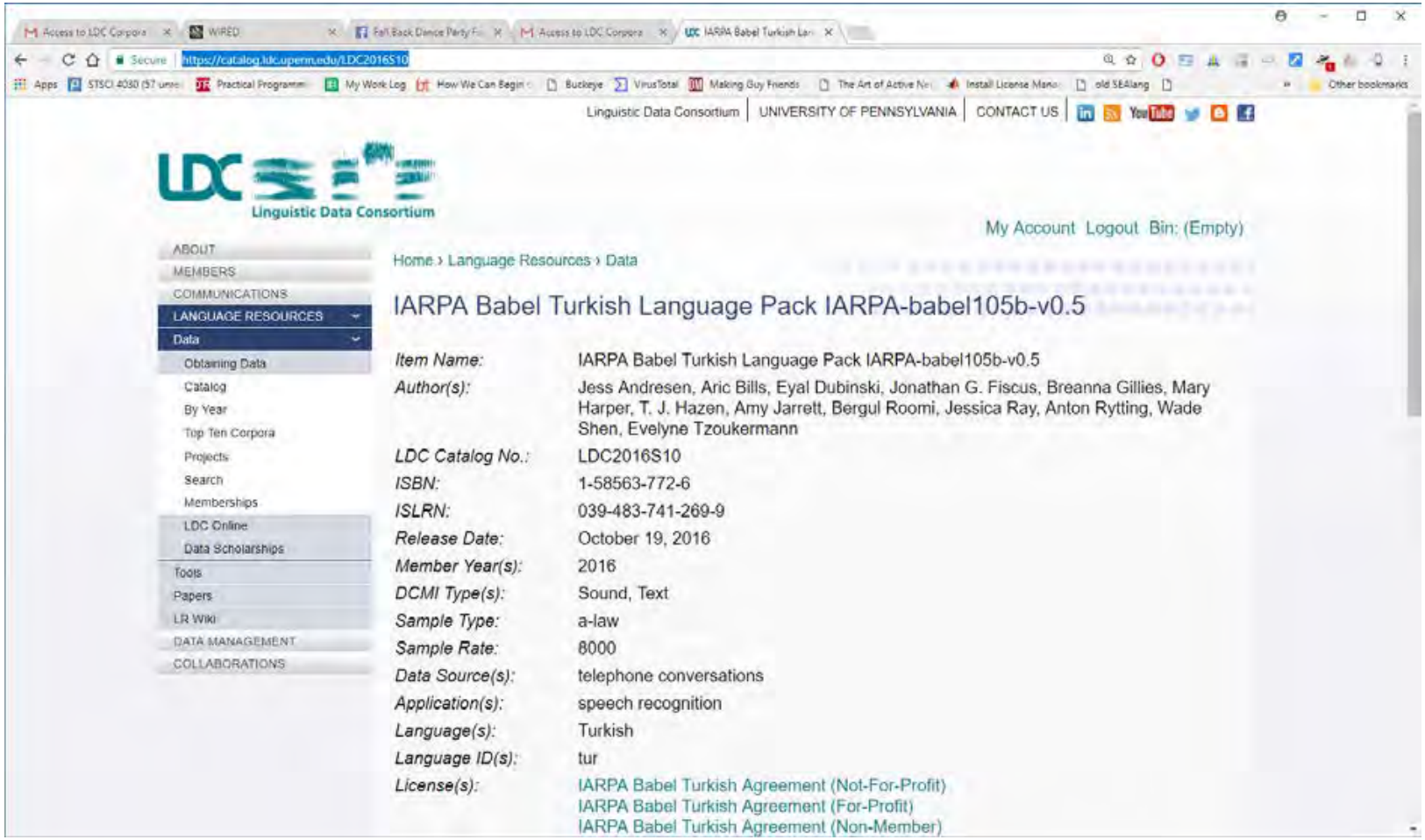
https://catalog.ldc.upenn.edu/search



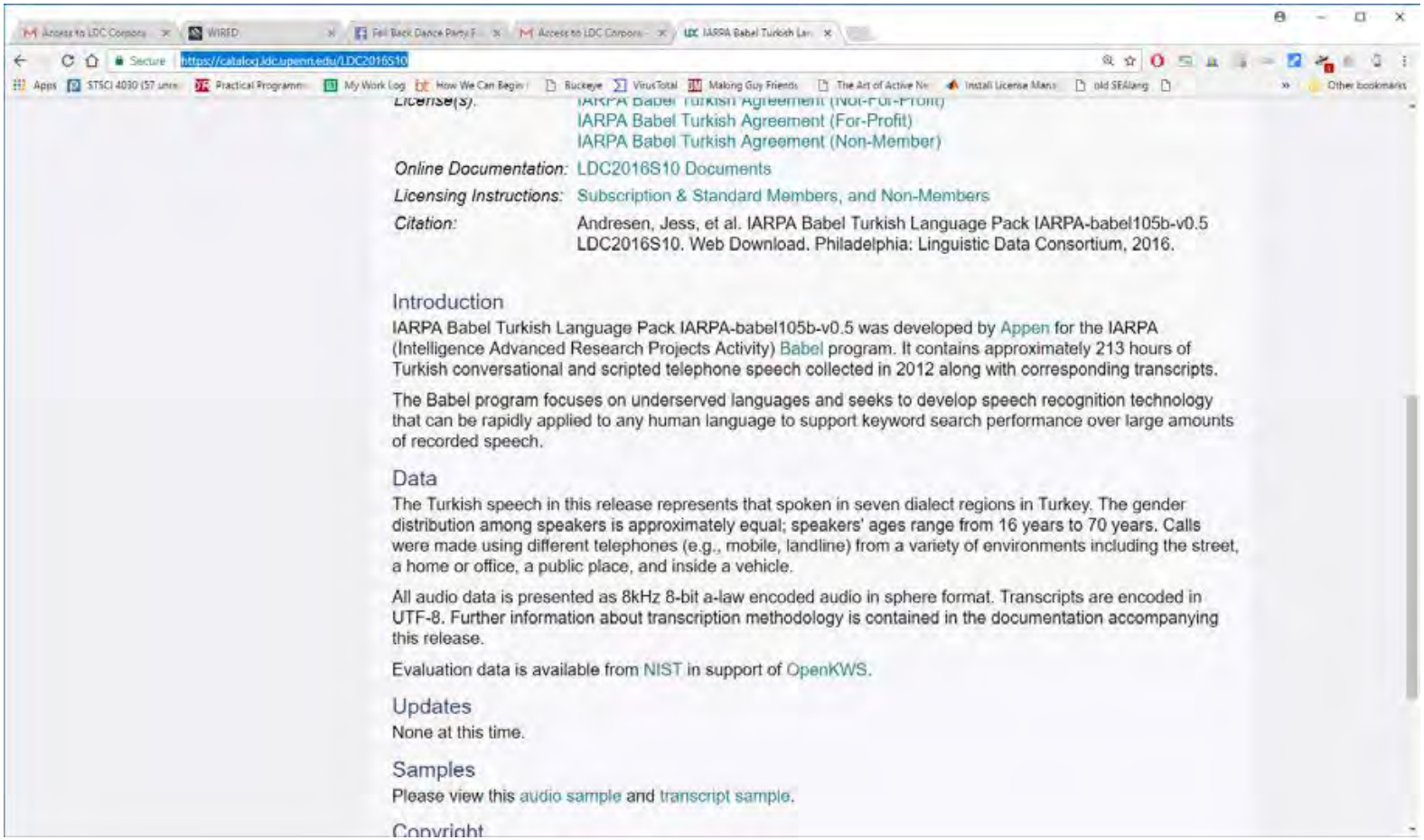
Sample Search Results



Typical Corpora Description (part 1)



Typical Corpora Description (part 2)



License(s): IARPA Babel Turkish Agreement (Not-For-Profit)
IARPA Babel Turkish Agreement (For-Profit)
IARPA Babel Turkish Agreement (Non-Member)

Online Documentation: LDC2016S10 Documents

Licensing Instructions: Subscription & Standard Members, and Non-Members

Citation: Andresen, Jess, et al. IARPA Babel Turkish Language Pack IARPA-babel105b-v0.5 LDC2016S10. Web Download. Philadelphia: Linguistic Data Consortium, 2016.

Introduction

IARPA Babel Turkish Language Pack IARPA-babel105b-v0.5 was developed by Appen for the IARPA (Intelligence Advanced Research Projects Activity) Babel program. It contains approximately 213 hours of Turkish conversational and scripted telephone speech collected in 2012 along with corresponding transcripts.

The Babel program focuses on underserved languages and seeks to develop speech recognition technology that can be rapidly applied to any human language to support keyword search performance over large amounts of recorded speech.

Data

The Turkish speech in this release represents that spoken in seven dialect regions in Turkey. The gender distribution among speakers is approximately equal; speakers' ages range from 16 years to 70 years. Calls were made using different telephones (e.g., mobile, landline) from a variety of environments including the street, a home or office, a public place, and inside a vehicle.

All audio data is presented as 8kHz 8-bit a-law encoded audio in sphere format. Transcripts are encoded in UTF-8. Further information about transcription methodology is contained in the documentation accompanying this release.

Evaluation data is available from NIST in support of OpenKWS.

Updates

None at this time.

Samples

Please view this audio sample and transcript sample.

Copyright

Search LDC for Parallel Texts

- *Parallel Text = English vs. Translation in another Language*
- Type “parallel” in the **keywords** field
- Select one of the following:
  - “machine translation” in the **applications** field, *or*
  - your favorite language in the **languages** field