# Ancestry application, 2023-03-08, Job Descriptions from Previous Work

## CaptionCall

CaptionCall is a provider of captioning services on special telephones available to hard-of-hearing individuals through the Americans with Disabilities Act. They are subcontracted by the FCC, who administer the ADA program for the deaf and hard-of-hearing. CaptionCall is an outgrowth of Sorenson Communications, who provide video-relay services for the deaf through American Sign Language.

Responsibilities of my role can be described as follows.

- Perform literature review and contribute to decisions about research priorities.

- Build a variety of tools for automated speech recognition, natural language processing, metrics for performance evaluation, models underlying speech recognition - including machine-learning models and physical models, and other subjects as the need arises.

- Participate in the creation of Intellectual Property, including patents.

- Find, create, and curate datasets. Take the time to become familiar with the datasets and clean them up as needed so they can be used by other projects.

- Write proof-of-concept code based on your research, which will be either passed on to engineering or be developed by the research team in conjunction with engineering.

- Use and create machine-learning, speech-to-text, data manipulation, and other types of tools, workbenches, software, packages, and native scripting tools on Linux with a view to use the results on mostly Windows machines used by human transcribers.

- Write scoring software that allows consistent and unambiguous representation of different, equivalent transcriptions that will be scored fairly. This involves a normalization of the transcribers' output, whether they be human or machine.

- Contribute to a long-range project combining the best of human transcribers with the best of machine transcribers for increased accuracy.

- Create an iterative break-testing set of text transcriptions - regression testing - for any text normalization.

- Understand issues with encoding and decoding of text and audio files that could prevent optimal performance and analysis.

- Leverage the strengths of both human and machine transcribers in real time to provide increased accuracy.

- Provide quantitative measurement of processes created to present to the research team, senior management, shareholders, government and other organizations who advocate for the hard-of-hearing.

- Determine which groups, including protected groups, are better or worse served by human transcription, machine transcription, or combinations of both. Create a dataset of phone audio for a variety of populations who might be worse served by any of the three transcription strategies. Handle issues of legal consent from and protection of these individuals.

- Create call-recording software with the help of a freelancer to gather data for various needs.

- Develop a robust method of silence detection and removal, then of splicing the mostly-spoken sections into audio files with a specific time length. Administer these audios to different transcribers and analyze the results.

- Broaden the marketable products and services available with machine learning methods and models that relate to speech and language with a special focus on serving underrepresented groups such as the Deaf and Hard-of-Hearing communities.

- Use versioning control, especially git, to work in an Agile software-development environment.

## FamilySearch Data Analyst

FamilySearch is a non-profit organization that provides research and document services in family history. ~~They provide research assistance online and in-person, obtain historical documents for preservation and/or digitization then allow access to many of them, and work to make an ever-increasing number of documents easily searchable and accessible.~~

Responsibilities of my role and my qualifications for it can be described as follows.

- The primary motivation for the creation of the job was the preparation of data truth and test sets along with providing the research team with additional data, documents, annotation, analysis, summaries, and other support as needed.

- Enthusiasm about mining, manipulating, and annotating data from multiple languages.

- Working knowledge and competence in several non-English Language as well as of languages and language patterns.

- Ability to grasp new computational and technical topics as well as self-sufficiency to program in several languages - Java, Perl, and C++, especially - in order to manipulate data.

- Ability to work both with Windows systems as well as Cygwin or Linux.

- Accomplish tasks as an individual or as a team member under time constraints. Brainstorm and identify best solutions in several teams.

- Attention to detail accompanied by patience in teaching technical processes to individuals with varying degrees of technical knowledge, including many with little technical knowledge.

- Data manipulation with Linux tools as well as higher-level programming.

- Provide technical tools, programs, and informed discussion to the research team and volunteer annotators for materials needed to tune machine learning for tasks including 1) zone articles and fields - analyze layouts and text continuity, detect to which of a variety of groups a segmented object belongs 2) index records, 3) transcribe printed sources through Optical Character Recognition, 4) transcribe handwritten source documents through building new Deep Learning models.

- Identify documents as well as types of documents that will allow the expansion of these processes to work on multiple languages.

- Be flexible and organized, since daily activities, including those listed below, vary depend on the latest data needs of the research team.

- Write quick applications that can, among other things

- Aid in data review and selection

- Manipulate catalog source lists in CSV and spreadsheet form to record prioritized microfilm or digital image selection list.

- Analyze data that volunteers have worked on as part of the data preparation process, including deciding metrics for performance and progress tracking

- Perform some of the non-programming work of the data team and its volunteers, including: tagging, zoning, line segmenting, etc. as a way of vetting the quality of work done as well as to help complete part of the needed annotations.

- Locate document images that are potentially pathological for the algorithm, that is, documents with objects in them that could be confused for writing, layouts and certain annotations in the original document that could seem ambiguous to the model while they wouldn't be to a knowledgeable human transcriber, and other things that could confuse the algorithm and/or cause it to struggle.

- Handle the decoding of text files, which can potentially be sent from different regions, older computer software, or those unacquainted with standard encoding of files. Convert all text-file transcriptions to the utf-8 encoding of Unicode.

- Evaluate the text-recognition capabilities of third parties whose technology could potentially give higher accuracy or give a new capability or a capability in a language we have not yet worked on. These could be in other programming languages, in which case an interface should be created or used. Decide on metrics to use for the scoring of such third-party programs.

- Analyze datasets, whether internal that might be part of new indexing projects or those which we might want to acquire, to analyze factors contributing to the decision of whether and how to use them. Factors could include technical details, financial and legal details of the datasets, current or past political factors that could make data sensitive, linguistic and naming details, etc.

- As requested, do visualizations of the data. This could range from viewing the Deep Learning algorithm's output as it would be seen on the original documents to plots of needs and progress.

- Help create guidelines which are concise but cover sufficient detail that the volunteers can use them as a guide in image segmentation and transcription.

## Document Specialist - Granite Mountain Records Vault

The Granite Records Mountain Vault section of the Church History Department of the Church of Jesus Christ of Latter-day Saints has the responsibility of storing historical documents for the church as a whole as well as many large collections of materials useful for family history. One major project has been the digitization of the family history microfilm collection for FamilySearch.

Responsibilities of my role and qualifications for it can be described as follows.

- Digitization of thousands of family-history-related images from microfilm documents

- Image analysis, processing, and manipulation to ensure readability

- Quality assurance for individual images as well as whole-film issues

- Special focus on ensuring that no images missed being captured

- Serve as a paleography and language-recognition resource for other workers

fgadsdfa

<https://ancestry.wd5.myworkdayjobs.com/en-US/Careers/userHome>

<https://careers.ancestry.com/homepage>