Title: MATERIAL Farsi-English Language Pack

Author(s): Aric Bills, Sarra Chouder, Cassian Corey, Marjan Davoodian, Eyal Dubinski, Corinna Ellis, Reza Farnam, Paul Gibby, Luke Hartwig, Dagmara Kalnins, Michael Kazi, Julie Lam, Hanh Le, Nicolas Malyska, Sarah Marvi, Sara McConnell, Jennifer Melot, Alyssa Mensch, Alex Moore, Michelle Morrison, Shelley Paget, Frederick Richardson, Annette Roberts, Carl Rubino, Marjan Sadeghi Moaddel, Bern Samko, Kenneth Saw, Pradeepti Sen, Rosanna Smith, Jonathan Taylor, Brian Thompson, Audrey Tong, Richard Tong, Andrew Weller, Sasha Wilmoth, Jennifer Yu, Ilya Zavorin

Languages:
- speech files and transcriptions: Farsi (FAS)
- translations of speech transcriptions: English (ENG)
- queries: English (ENG)

Introduction

Collected to support the MATERIAL (Machine Translation for English Retrieval of Information in Any Language) program, the MATERIAL Farsi dataset contains Farsi speech files and text documents of several genres, English queries, and various annotations that can be used for ASR (Automatic Speech Recognition), MT (Machine Translation), and CLIR (Cross Language Information Retrieval) research. This pack includes Farsi conversational telephone speech (CS) files. These files have different partitions (e.g., BUILD, ANALYSIS, DEV, EVAL, etc.) to indicate their use in the MATERIAL program. This pack also includes transcriptions and English translations for the ANALYSIS files and English queries and their relevance annotations.

For information regarding how the Farsi files were transcribed and translated as well as annotated for query relevance please refer to, respectively, MATERIALTranscriptionAndTranslationConventions.pdf and MATERIALQueryAnnotation.pdf. Query syntax is described in detail in MATERIALQueryLanguageSpecification.pdf also in the **docs/** directory. This directory also contains metadata information for the speech files.

The size for the speech files are:
- data/build-asr approximately 35 hours
- data/speech approximately 26 hours

This pack is organized as follows:

```
data/
        build-asr/
                speech/                                 speech source and its transcriptions
                        dev/                            recommended dev partition
                                src/
                                transcription/
                                transcription_roman/
                        training/                       recommended training partition
                                src/
                                transcription/
                                transcription_roman/
        speech/                                         speech files (CS)
                src/
                transcription/
                transcription_roman/
```

```
            translation/
        query/
                query_list.tsv                       contains queries that have relevant files in this pack
                query_annotation.tsv                 contains query relevance annotations
docs/
        README.pdf                                   this file
        MATERIALTranscriptionAndTranslationConventions.pdf        describes the transcription and translation
conventions
        MATERIALQueryAnnotation.pdf          describes query relevance annotation
        MATERIALQueryLanguageSpecification.pdf  describes the syntax of MATERIAL queries
        doc_info.tsv                         indicates ID of the language, genre, and partition
        genre_map.tsv                        indicates the genre mapping
        LSDD_3S.pdf                          language specific design document
        analysis_lexicon.txt                 lexicon for the ANALYSIS speech files
        build_lexicon.txt                    lexicon for the speech files in build-asr
        metadata_speech.tsv                  speech file metadata
```

Recommended/Expected use of corpus
- Speech files and their transcriptions can be used to train, develop and/or evaluate Automatic Speech Recognition models, or for similar purposes that require speech recordings accompanied with corresponding transcriptions.
- Queries and speech files with the corresponding relevance annotations can be used to train, develop and/or evaluate Cross Lingual Information Retrieval models, or for similar purposes that require English queries marked for relevance against foreign language speech recordings.

Collection Procedure - format, method, and timespan
The conversational telephone speech (CS) files are natural conversations between two speakers over the telephone. The speakers involved in the collection were required to be native language speakers. They were recruited with the goal of obtaining broad coverage of age, gender, and dialect. They were encouraged to talk about topics they felt most comfortable discussing such as family, friends, sports, movies, etc. These topics were not fixed and varied across languages. Speakers showing distinctive speech disorders were excluded from collection or removed if identified later in the transcription process. All speakers were 18 or older. Dialect regions were defined prior to collection for each language. The number of chosen dialects varied across languages with no dialect representing less than 10% of the collection. There were no restrictions on the acoustic environment (such as whether or not the speaker was indoors, outdoors, driving, etc.), and this information was provided by the speakers. There were also no restrictions on network specifications or telephone models and these values were also noted in the accompanying metadata. Audio was recorded via telephone over an ISDN connection with a terrestrial telephone network. Each speaker was recorded on a separate channel. No post-processing steps were taken to reduce noise or other artifacts of the recording medium at any stage.

Data format specific details
Speech files are in WAV or SPHERE format. Available metadata information is included in the **docs/** directory.