# The Development of ALLSSTAR:
# The Archive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings

Lauren Ackerman , Laura Ann Burchfield, Lisa Hesterberg, Jenna Silver Luque[1], Kelsey Mok, and Ann Bradlow
Northwestern University, Department of Linguistics
[1]SLPJenna@gmail.com

The content of most non-native speech corpora today fall into one of two categories: foreign-accented speech and multi-lingual L1 speech. Despite these types of corpora being extremely valuable resources and the representation of languages being expansive, numerous research limitations arise from restricting corpora to only one category. For example, a corpus that contains exclusively either foreign-accented speech or multi-lingual L1 speech neglects the possibility of comparison between the two categories. There are no corpora that we are aware of that include recordings of speakers in both their L1 and L2. This leaves researchers with unanswered questions such as whether a specific feature is the result of a transfer from the L1, a characteristic of that individual's speech, or a characteristic of L2 speech in general. Moreover, the absence of consistency in materials across corpora and languages is another substantial limitation. ALLSSTAR addresses both of these limitations by including recordings of non-native English speakers in their L1 and in English with comparable materials across languages.

The design of ALLSSTAR features both scripted and spontaneous speech. The multi-lingual scripted materials consist of parallel translations of a paragraph (*The North Wind and the Sun* passage, International Phonetic Association, 1999) and sentences of varying length, register and style (excerpts from *The Little Prince,* Saint-Exupéry, 1946), the *Universal Declaration of Human Rights* (United Nations, 1948), and *HINT* (Hearing in Noise Test) sentences (Soli & Wong, 2008). The spontaneous materials include two types of prompts: open-ended questions and four simple picture stories (Mayer, M. various titles). Though still in development, we currently have recordings from 40 speakers from 15 language backgrounds (German, Hebrew, Russian, Turkish, Mainland Mandarin, Taiwanese Mandarin, Singaporean Mandarin, Brazilian Portuguese, Japanese, Gujarati, Korean, Gishu, Vietnamese, Spanish, and Hindi) producing the scripted and spontaneous materials in English and in their respective L1s.

We have several future projects planned that will indirectly address the advantage of incorporating both foreign-accented speech and non-native L1 speech in a corpus. One project will examine the tones in scripted vs. spontaneous Mandarin speech. Another will compare native and non-native speech recognition in different types of noise, comparing scripted sentences and spontaneous speech. A third will investigate the perceived phonetic likeness of languages using the paragraph recordings. We will also attempt to develop an automatic method of speaker and language identification. A final planned project will analyze the types of errors non-native speakers make on the English HINT.

**References**

International Phonetic Association (1999.) Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge: Cambridge University Press.

Mayer, M.(1974). Bird's New Hat. In *Two Moral Tales*. New York: Four Winds Press.

Mayer, M.(1973). *Bubble Bubble*. New York: Parents' Magazine Press.

Mayer, M. (1974). Just a Pig at Heart. In *Two More Moral Tales.* New York: Four Winds Press.

Mayer, M. (1974). Bear's New Clothes. In *Two Moral Tales.* New York: Four Winds Press.

Saint-Exupéry, A. (1946). *Le Petit Prince.* Gallimard: Paris. Multi-lingual translations taken from various internet sites.

Soli, S.D. & Wong, L.L.N (2008). Assessment of speech intelligibility in noise with the hearing in noise test. International Journal of Audiology, 1-6.

United Nations (1948). The Universal Declaration of Human Rights. Multi-lingual versions available from http://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx