Lifelong Learning Programme

DigLin

Digital Literacy Instructor

Deliverable 3.1

Speech Data and Annotations

(Version 1.0)

Vanja de Lint,
Marta Dawidowicz, Lucy Martin, Helmer Strik, Taina Tammelin-Laine

universität wien

Newcastle University

CLST | Centre for Language and Speech Technology
Radboud University

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

friesland college

northumbria UNIVERSITY NEWCASTLE

This project has been funded with support from the European Commission under project: 527536-LLP-1-2012-1-NL-GRUNDTVIG-GMP. This publication reflects only the views of the project consortium members, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

0 INTRODUCTION

This report is the result of work done as part of Work Package 3 as described in the proposal for the EU Lifelong Learning Project DigLin (the *Digital Literacy Instructor*). In what follows we describe for each of the four languages of the project (Dutch, English, Finnish or German) any pre-existing speech data available from speakers that resemble the target users in the DigLin application (adult immigrants learning to read for the first time in their second language).

1 DUTCH

1.1 WHAT KIND OF DATA?

For Dutch there are data available from a corpus called LESLLA (Low-Educated Second Language and Literacy Acquisition). In the recordings there are 1 or 2 people present (of which 1 non-native speaker), and the non-native speaker performs sentence imitation tasks and story retelling. In total there are 5 different tasks, each of which was performed by the non-native speaker at 3 separate points in time. All audio has been cut into utterance-length files, which hardly ever exceed 1 minute.

1.2 HOW MUCH DATA?

In total there are 14289 audio files, with accompanying annotation files. It concerns speech from 8 Turkish speakers and 7 Moroccan speakers.

	Moroccan	Turkish	**Total**
Discourse	630	700	1330
FatherDaughter	1492	1934	3426
Quest	809	857	1666
Sentence Imitation	628	720	1348
Snowman	2898	3621	6519
Total	6457	7832	14289

1.3 WHAT FORMAT?

The audio files are in wav format. The transcriptions/annotations were done in PRAAT and are split into a Textgrid file and an ELAN file.

1.4 ANY ANNOTATIONS?

Ortographic transcriptions are available for all utterances. The full database (including audiofiles, transcriptions and metadata) can be accessed through the following link: https://corpus1.mpi.nl/ds/asv/?openpath=node:1893295.

2 GERMAN

2.1 WHAT KIND OF DATA?

The available English data have all been obtained in academic context:

1 exam situation (examiner, second assessor and candidate)
2 expert talk/ discussion (moderator, speaker, various numbers of panellists)
3 student presentation (speaker, lecturer).

2.2 HOW MUCH DATA?

There is a total of 58.18 hours of audio material, which can be divided over 189 “communications”. The data comes from a total of 257 speakers.

2.3 WHAT FORMAT?

All audio files are in mp3 format.

2.4 ANY ANNOTATIONS?

All transcriptions and annotations can be found on the following website: https://gewiss.uni-leipzig.de/index.php?id=home. Access is secured.

3 FINNISH

3.1 WHAT KIND OF DATA?

Non-native Finnish data can be obtained in the form of recordings of the national spoken language test (YKI). Each session is about 2 minutes long. These recordings unfortunately have a rather strong background noise.

3.2 HOW MUCH DATA?

There are at least 1600 recording sessions. Every recording session is from a different speaker so the number of speakers is the same as the number of sound files, namely 1600.

3.3 WHAT FORMAT?

The audio files are in mp3 format.

3.4 ANY ANNOTATIONS?

There are some student annotations available through the University of Jyvaskyla.

4 ENGLISH

For English there are several databases available with non-native speech data. Below is a table which presents an overview.

First language	Nature of data	Number/nature of speakers	Format	Whose data
Arabic (several varieties)	1 Orthographically transcribed picture descriptions 7-10 minutes each 2 Repetition of 45 sentences to elicit past tense; only past tense forms transcribed	71 adults	WAV	Kahoul
Arabic Bengali, Dari, Persian	Individual sample page (69 words of same paragraph), on-line access to audio file, brief biodata, Doulos SIL IPA transcription of the sample and set of each speaker's phonological generalizations. Site also includes separate contrastive analyses of typical errors for every L1 group.	67 Arabic, 12 Bengali, 5 Dari, 16 Farsi, 10 Kurdish, 8 Panjabi, 8 Pashto, 6 Somali, 5 Tamil, 28 Turkish, 12 Urdu (and most other US immigrant languages). Samples are mostly US English.		Open source, Speech Accent Archive http://accent.gmu.edu/
Panjabi	Longitudinal data orthographically transcribed, with audio file accessible on-line	Two speakers in the UK	MP3	ESF - TalkBank
Arabic	Longitudinal data orthographically transcribed, with audio file accessible on-line	18 speakers	MP3	Qatar - TalkBank

5 REFERENCES

Campbell, G. 1991. *Compendium of the World's Languages.* London: Routledge.

Chavarria-Aguilar, O. 1962. *Pashto Basic Course*. Ann Arbor: University of Michigan Press.

Doke, C. 1954. *The Southern Bantu Languages.* London: Oxford University Press.

Ferguson, C. A. and M. Chowdhury. The phonemes of Bengali. *Language* 36: 22-59.

Kalelkar, N.G. 1965. *Marathi.* New Delhi: Indian Council for Cultural Relations.

Kaye, A.. (Ed.) and P. T. Daniels. 1997. *Phonologies of Asian and Africa (Including the Caucasus).* Volumes I and II. Winona Lake, Indiana: Eisenbrauns.

Kelkar, A. R. *Studies in Hindi-Urdu*. Deccan College Building Centenary & Silver Jubilee Series: 35.

Kelly, J. 1974. *Phonology and African Linguistics in African Language Studies,* vol.15. London: University of London.

Krishnamurti, B. 2003. *The Dravidian Languages.* Cambridge: Cambridge University Press.

Ladefoged, P. 1964. *A Phonetic Study of West African Languages.* Cambridge: University Press.

Ladefoged, P. and I. Maddieson. 1996. *Sounds of the World's Languages.* Oxford: Blackwell.

Mann, M. and D. Dalby. 1987. *A Thesaurus of African Languages: A Classified and Annotated Inventory of the Spoken Languages of Africa.* London: Hans Zell Publishers.

Meinhof, C. (trans. NJ van Warmelo). 1932. *Introduction to the Phonology of the Bantu Languages.* Berlin: Dietrich Reimer/Ernst Vohsen.

Okell, J. 1994. *Burmese: An introduction to the Spoken Language.* Dekalb: Northern Illinois University.

Shamal, A.R. 1965. *A Contrastive Study of the Segmental Phonemes of English and Dari* (Afghan Persian). (M.S. Thesis). Washington, DC: Georgetown University.

Steever, S. (ed). 1998. *The Dravidian Languages.* London: Routledge.

Swan, M. and B. Smith. 1987. *Learner English.* Cambridge: Cambridge University Press.

Turnbull, A. 1982. *Nepali Grammar and Vocabulary.* New Delhi: Asian Educational Services.

Urbanska, I. (Ed.). 1977. *A Handbook of Polish Pronunciation for English Learners.* Panstwowe Wydawrictwo, Naukowe:Warszawa.

Watson, J. 2002. *The Phonology and Morphology of Arabic*. Oxford: OUP.

Westermann, D., and I. C. Ward. 1990. *Practical Phonetics for Students of African Languages.* London: Kegan Paul International.