## Sharing speech data - notes from a small language

by Ingunn Amdal, NTNU
CLARIN and FLaReNet workshop at KTH 25 and 26 Nov 2009

CLARIN is about sharing language resources targeting users in Humanities and Social Sciences. For small languages like Norwegian, the need for sharing is also recognized in the speech technology community. There is far from enough data as stated in several BLARK reports on Norwegian. This talk will address several issues in sharing speech data; the (lack of) connection between the CLARIN goals and the speech technology community priorities, the challenges of sharing existing speech data, and guidelines for sharing future database collections. I will include examples from corpora showing the heterogeneous nature of speech (and multimodal) corpora we have to deal with in standardizing annotations as well as in defining metadata structures and taxonomy.