Possible contributions of the zero resource approach to SLT for underesourced languages

Emmanuel Dupoux, EHESS

1. Algorithms
2. Datasets

1. Algorithms: unsupervised and weakly supervised

- Subword representation learning
  - unsupervised rivals or beats multilingual features
  - weakly supervised use same word/different word information (Chen et al, 2015)
  - weakly supervised can scale up: Siamese networks trained on 360 hours of Librispeech beat supervised DNNs on TIMIT (Zeghidour, submitted).

  → could be applied to re-speaking datasets: Aikuma [Bird], BULB [Besacier]

- Spoken term discovery
  - integration of NLP methods (hierarchical non parametric Bayesian models; Lee et al. 2015)
  - integration of speech prosody (Ludusan et al)

  → could be used to improve audio keyword search [Tsakalidis]

- Picture captioning (challenge V2.0!)
  - speech based: Harwath & Glass (2016)

  → could be used with ethnographic-like datasets (to be constructed)

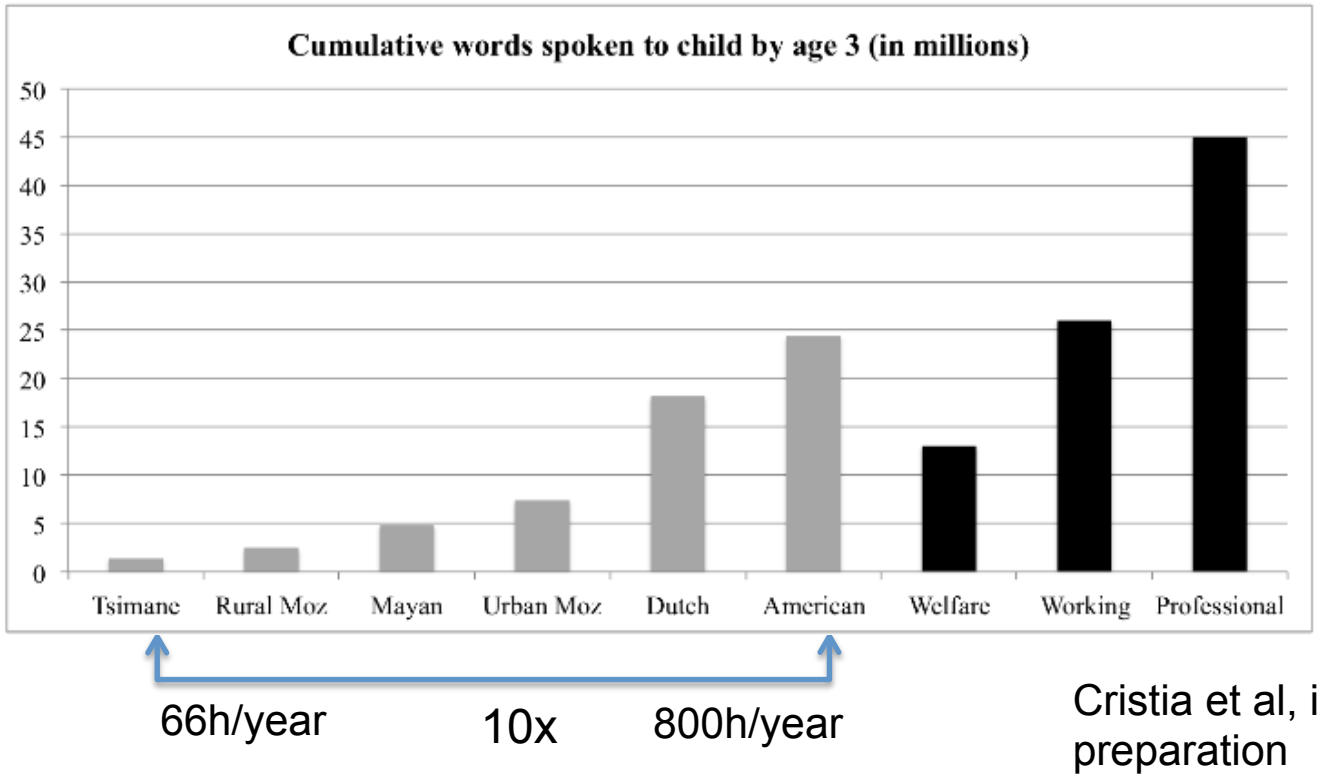
→ all of these algorithms should be open source and checked for robustness: Kaldi system [Khudanpur]

→ it would also be nice to have the participation of companies with large scale computing facilities [Jansche]

2. Data: dense home recordings

home recording (LENA device)

dense multimedia recording (Roy et al)



Cristia et al, in preparation

- reconstruct the sensory input of the infant.
- establish lower and upper bounds on the minimal amount of data sufficient to trigger language learning.
- document the range of socio-linguistic variation (amount of speech, number of speakers, multilingualism, language use) that underpins the transmission of (as yet) living languages [Lim].