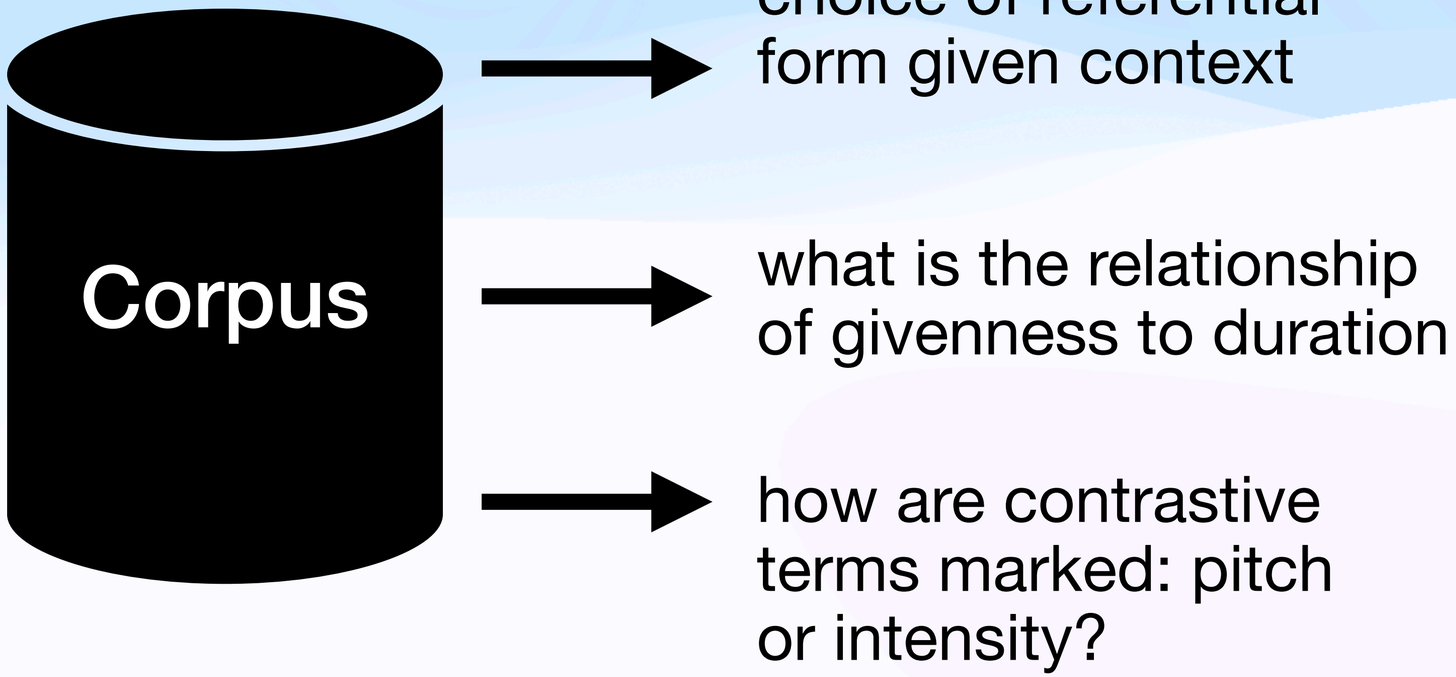
LLMs and Annotation

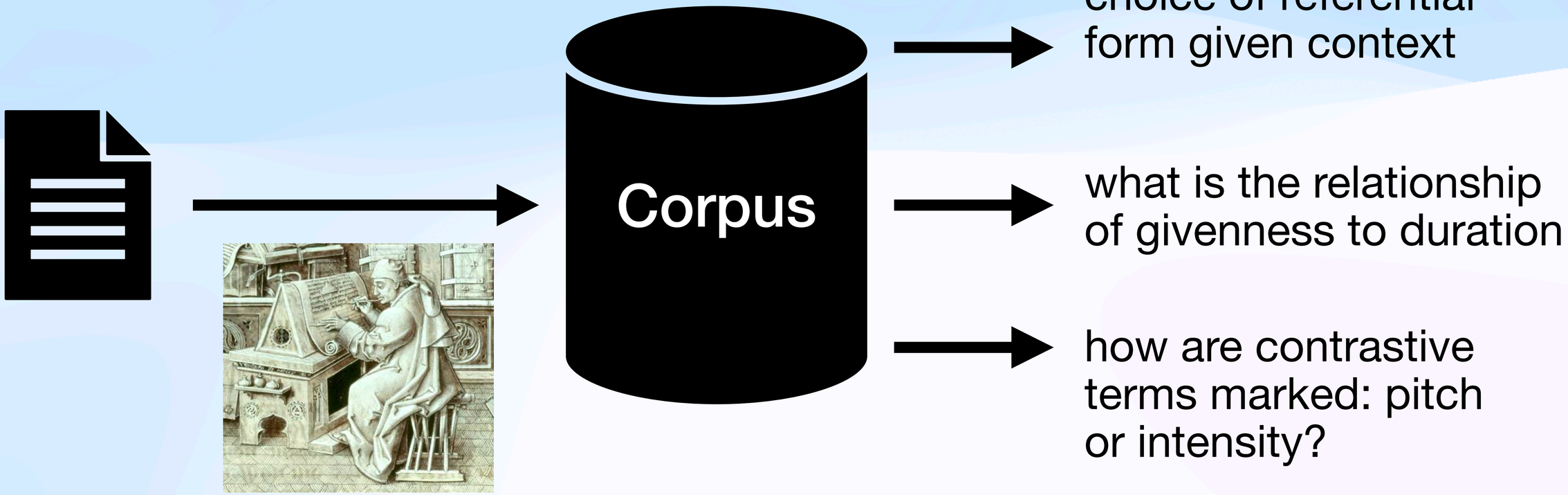
Reconciling Imperfect Solutions

2025-07-09 T. Mark Ellison on work with Amalia Canes Nápoles

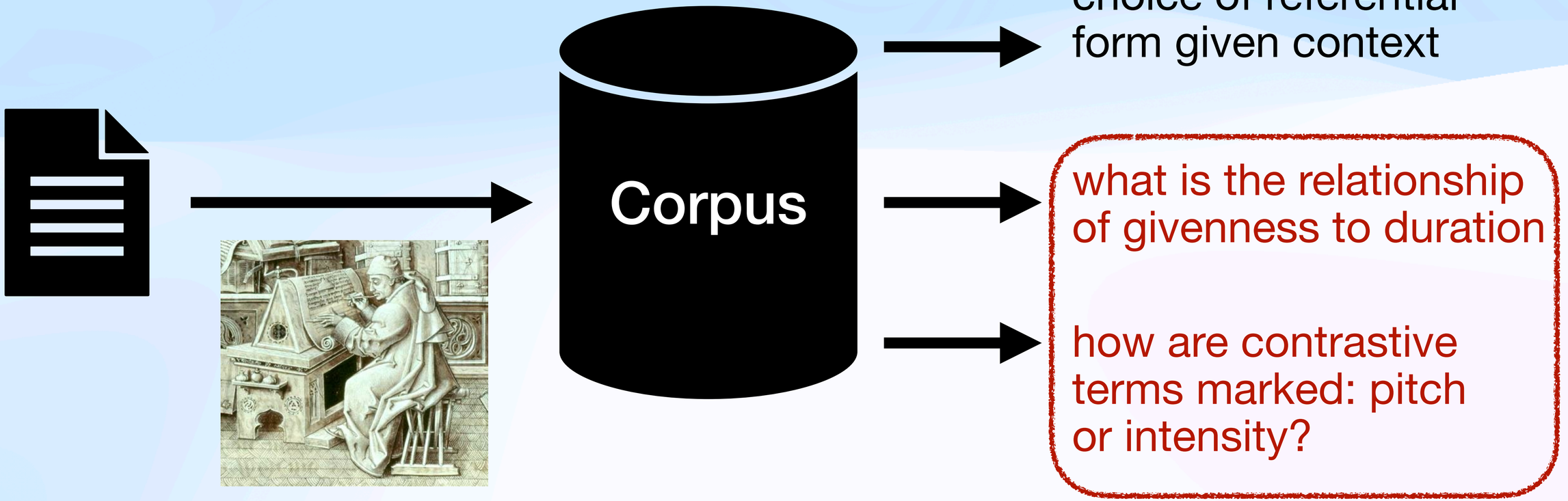
Corpora for Exploring Prominence in Natural Speech



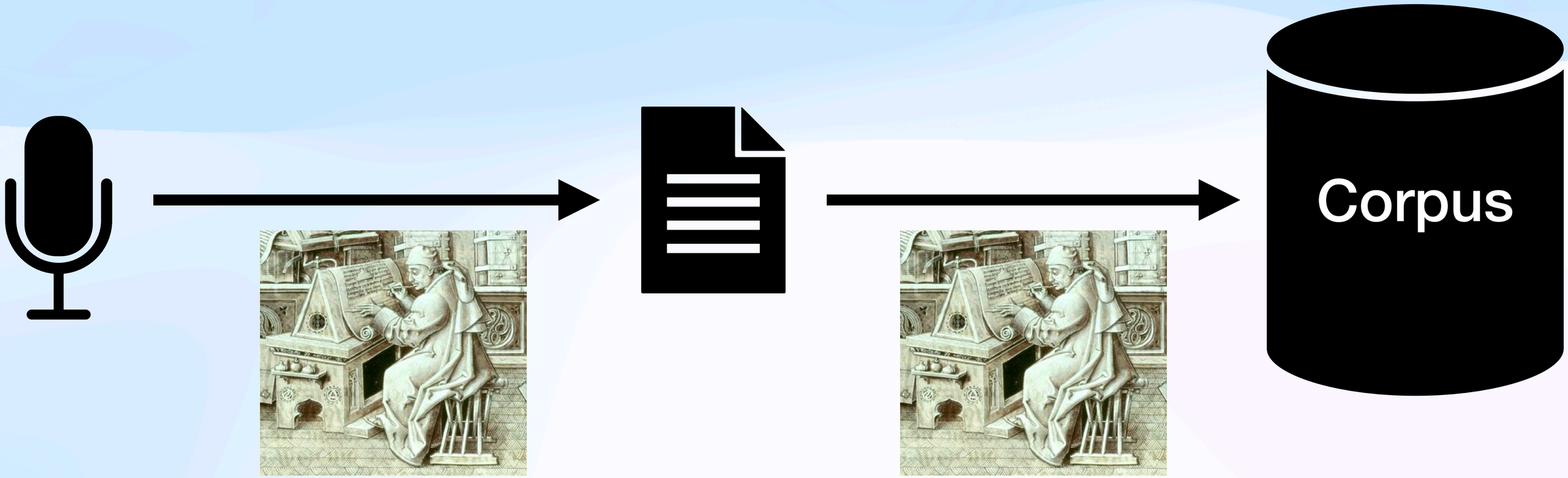
Corpora for Exploring Prominence in Natural Speech



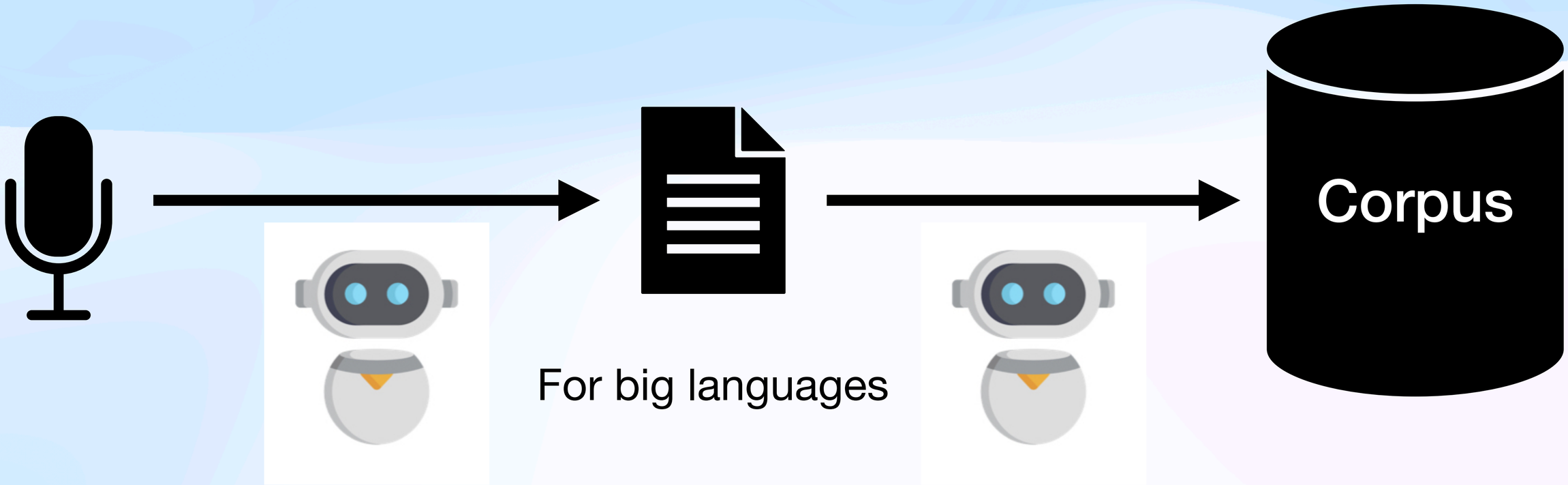
Corpora for Exploring Prominence in Natural Speech



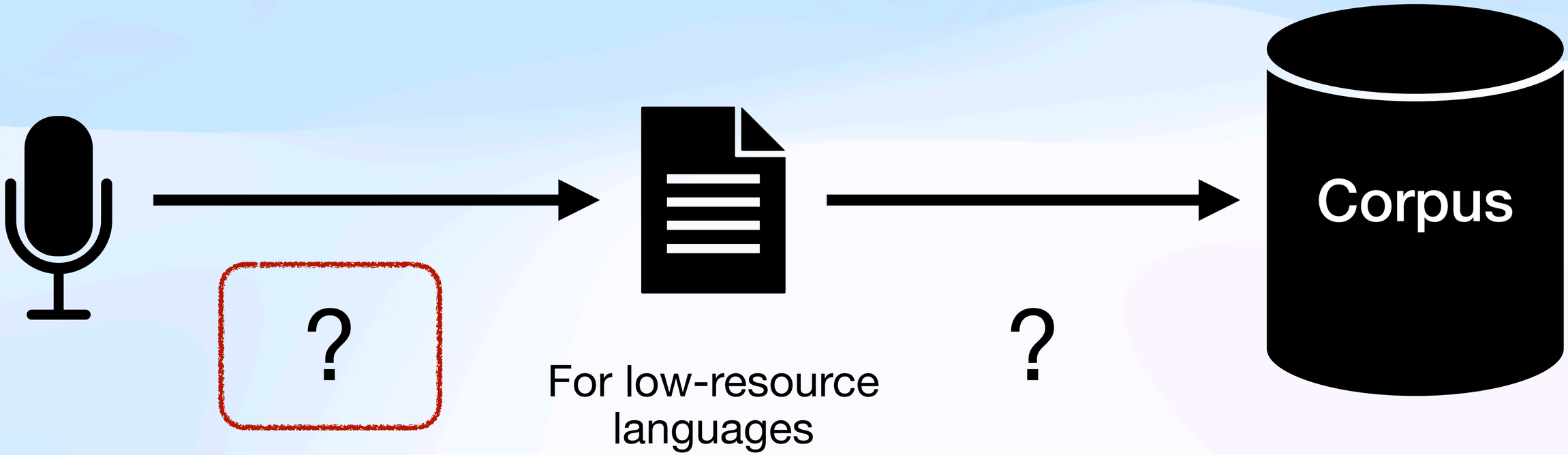
Corpora for Exploring Prominence in Natural Speech



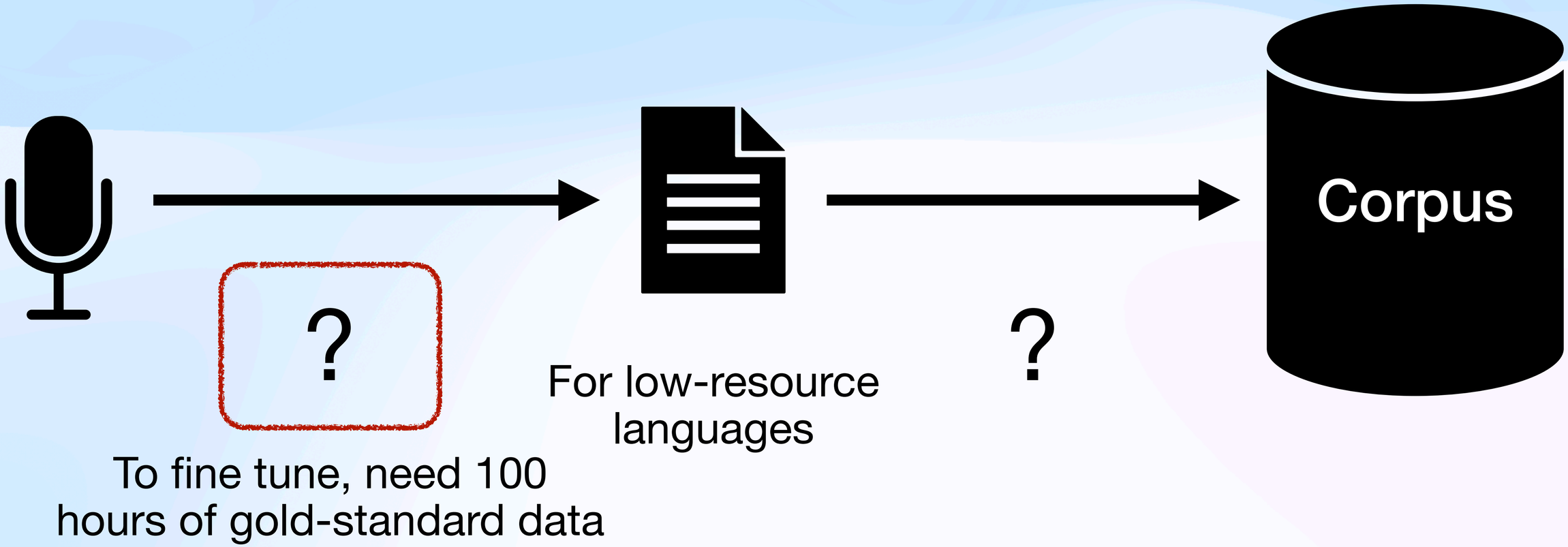
Corpora for Exploring Prominence in Natural Speech



Corpora for Exploring Prominence in Natural Speech



Corpora for Exploring Prominence in Natural Speech

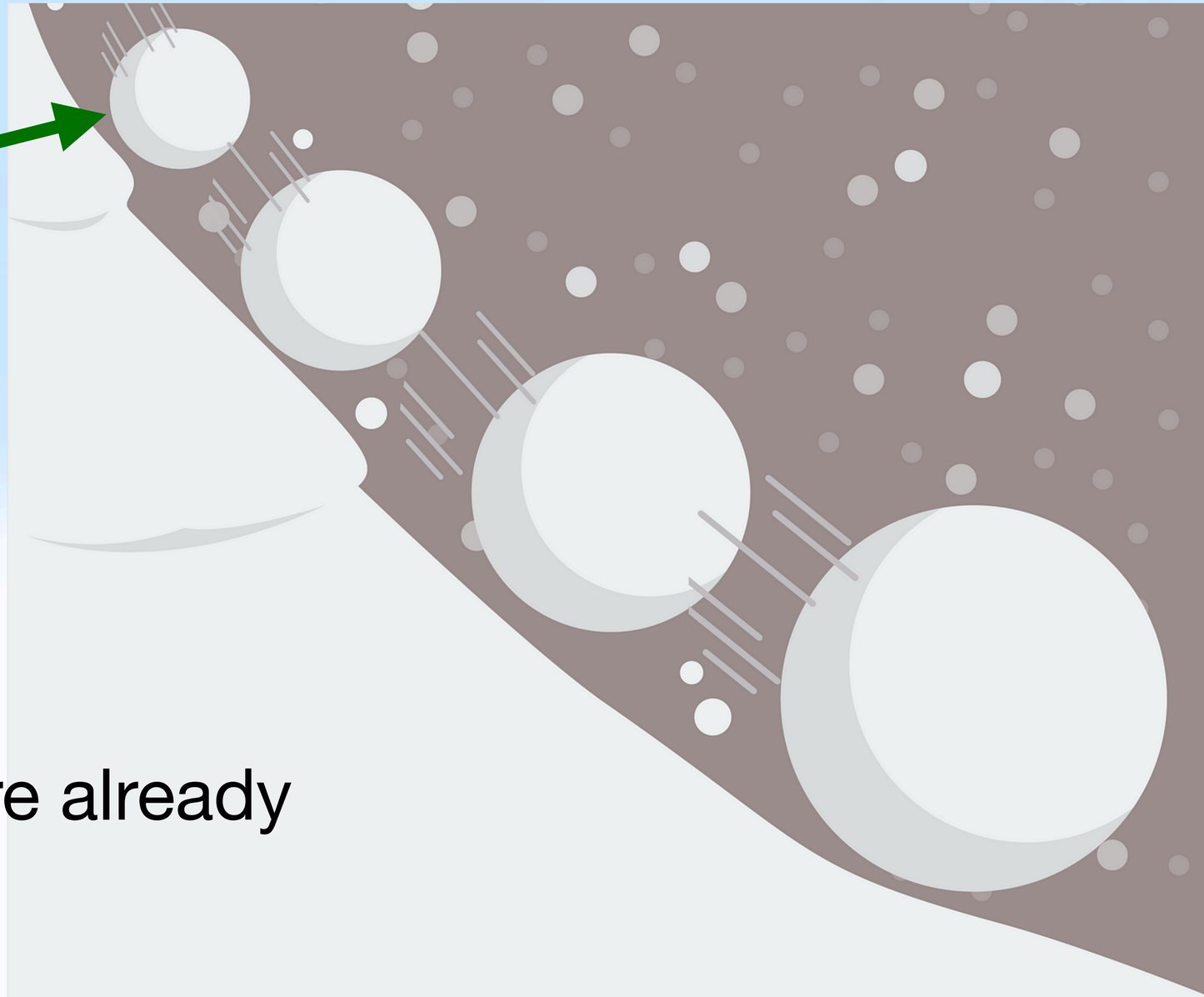


Guadeloupean Creole

ASR Models available:
Haitian Creole
French
(phonemic form)

Snowballing a Solution

- native speakers transcribe to make corpus ‘kernel’



We are up to here already

Bayesian Identification of Cognates and Correspondences

T. Mark Ellison
Linguistics, University of Western Australia,
and Analith Ltd
mark@markellison.net

T. Mark Ellison. 2007. **Bayesian Identification of Cognates and Correspondences**. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology,* pages 15–22, Prague, Czech Republic. Association for Computational Linguistics.

Abstract

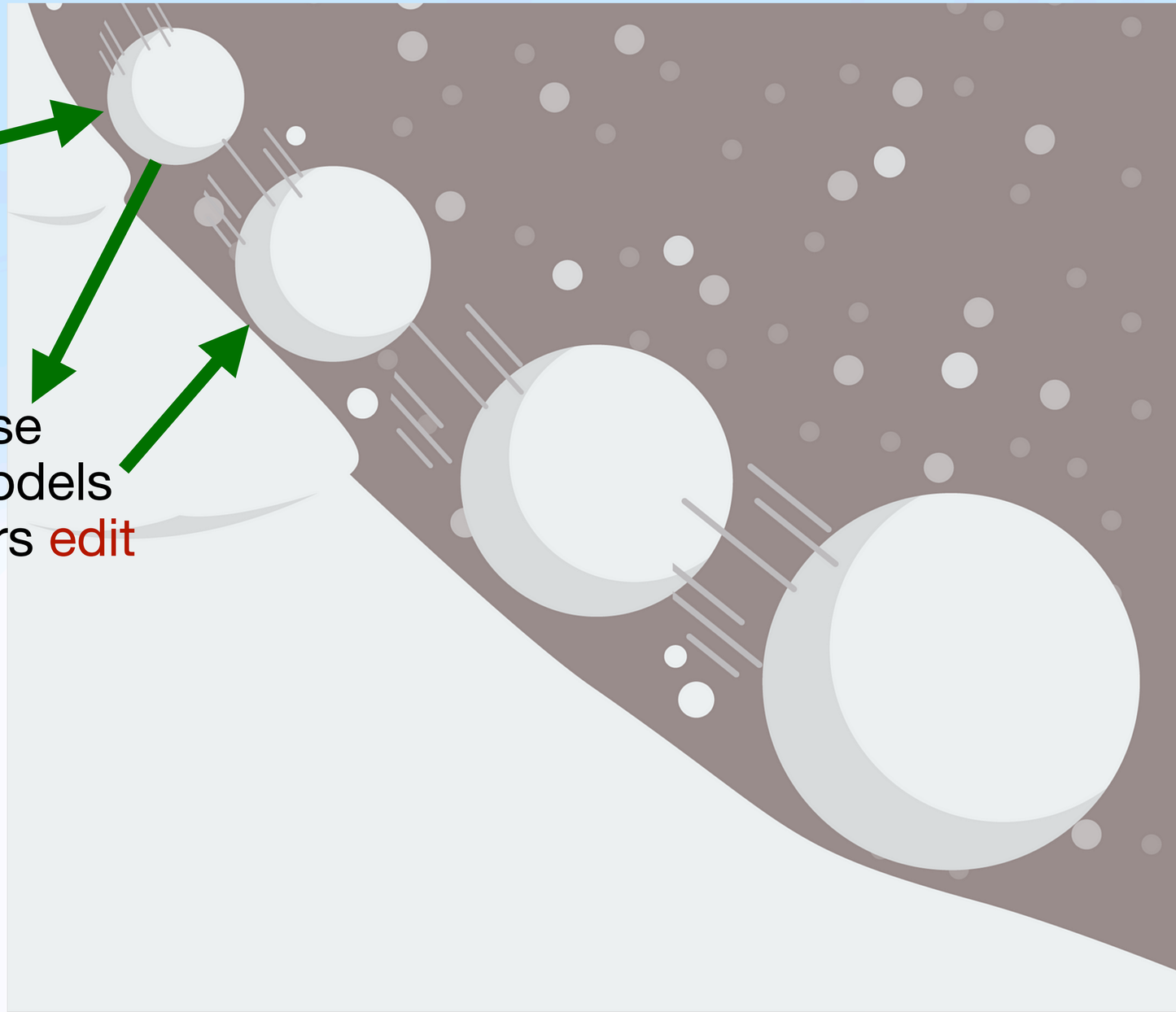
This paper presents a Bayesian approach to comparing languages: identifying cognates and the regular correspondences that compose them. A simple model of language is extended to include these notions in an account of parent languages. An expression is developed for the posterior probability of child language forms given a parent language. Bayes' Theorem offers a schema for evaluating choices

Other recent work has applied computational methods for phylogenetics to measuring linguistic distances, and/or constructing taxonomic trees from distances between languages and dialects (Dyen et al., 1992; Ringe et al., 2002; Gray and Atkinson, 2003; McMahon and McMahon, 2003; Nakleh et al., 2005; Ellison and Kirby, 2006).

A central focus of historical linguistics is the reconstruction of parent languages from the evidence of their descendents. In historical lin-

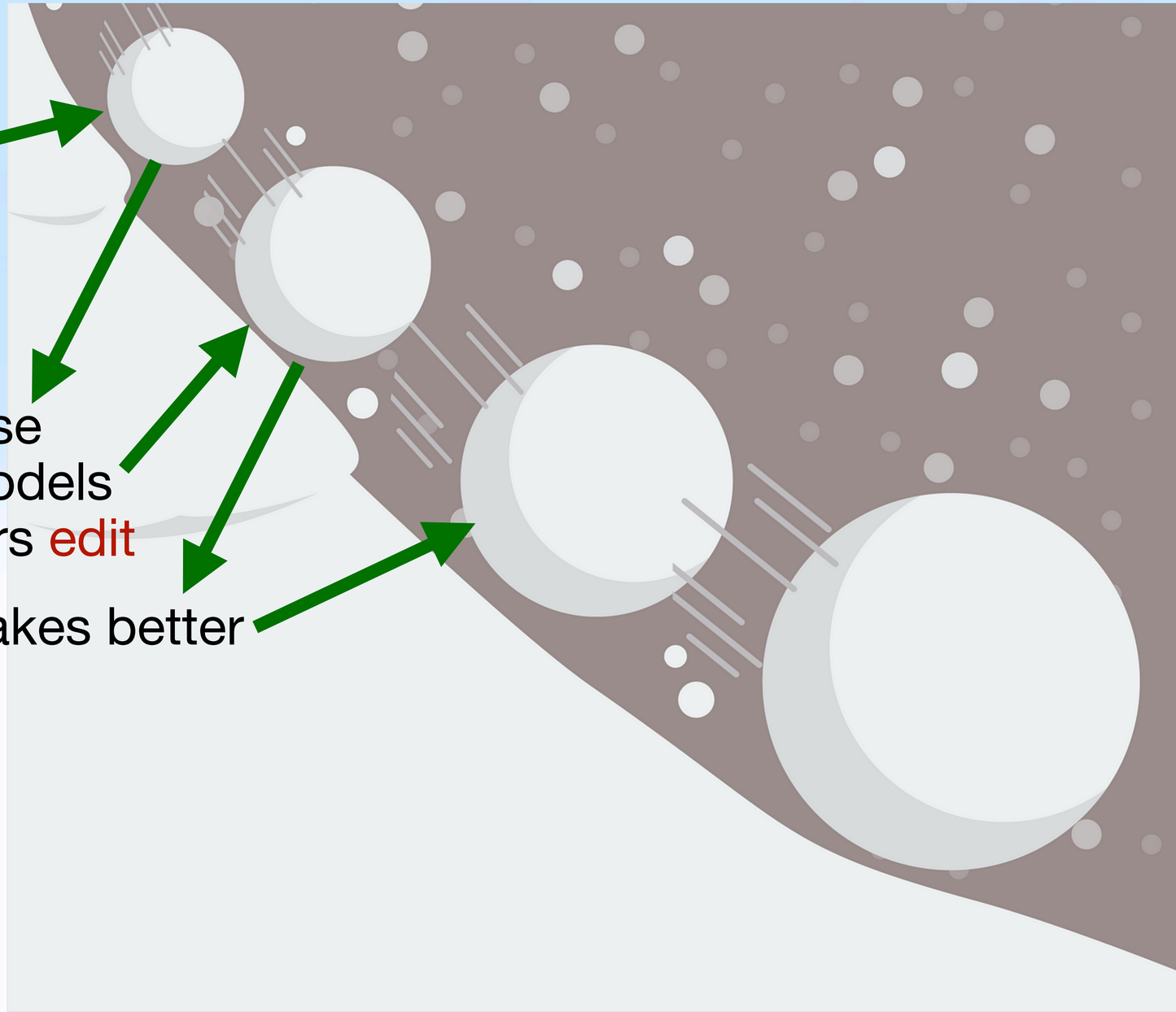
Snowballing a Solution

- native speakers transcribe to make corpus ‘kernel’
- predictive ML models hypothesise transcriptions given output of models of similar languages - transcribers edit



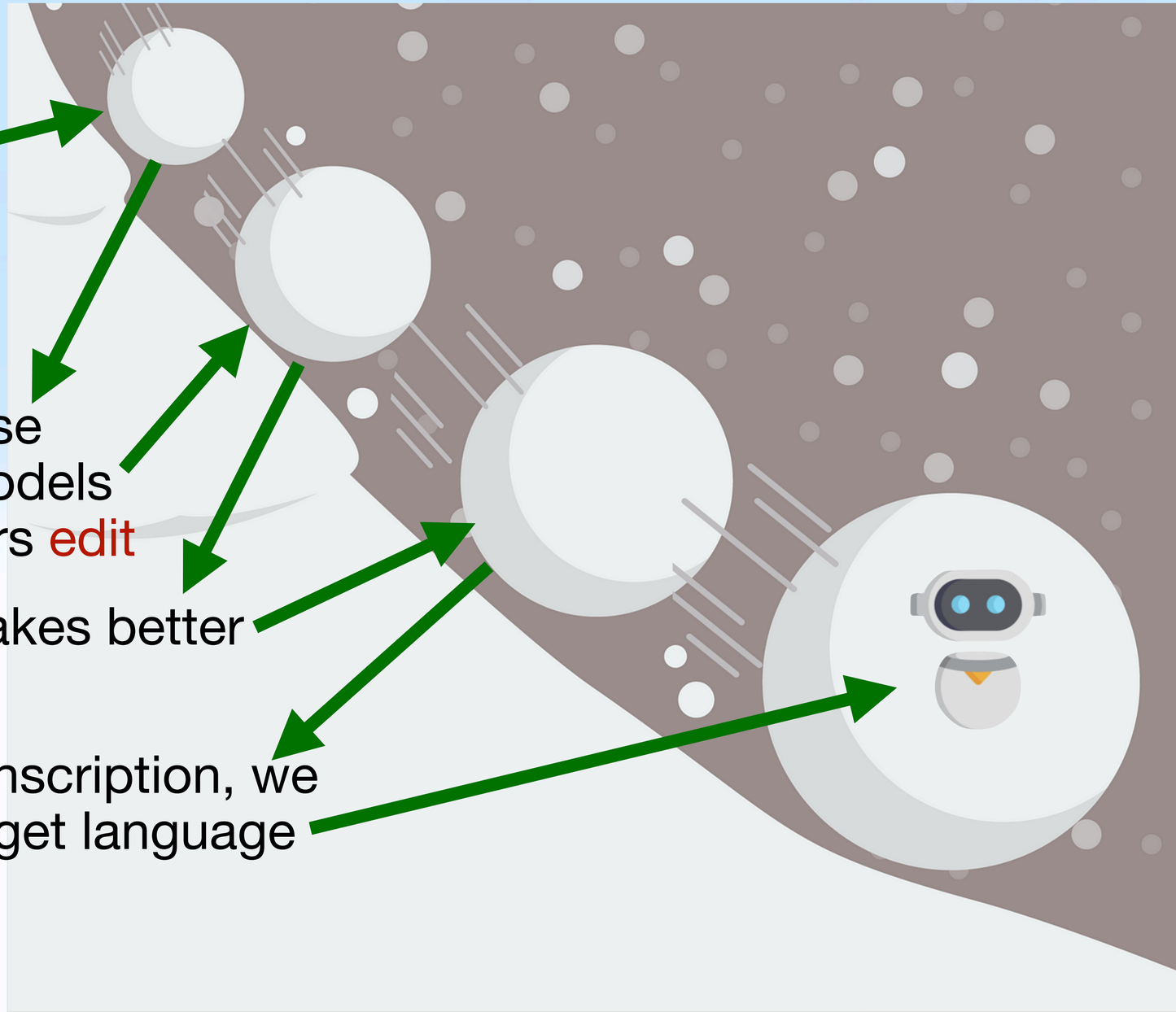
Snowballing a Solution

- native speakers transcribe to make corpus ‘kernel’
- predictive ML models hypothesise transcriptions given output of models of similar languages - transcribers edit
- more refined Bayesian model makes better predictions - transcribers check



Snowballing a Solution

- native speakers transcribe to make corpus ‘kernel’
- predictive ML models hypothesise transcriptions given output of models of similar languages - transcribers edit
- more refined Bayesian model makes better predictions - transcribers check
- with enough data from faster transcription, we can fine-tune a model for the target language



What *snowballs* look like to Australians: