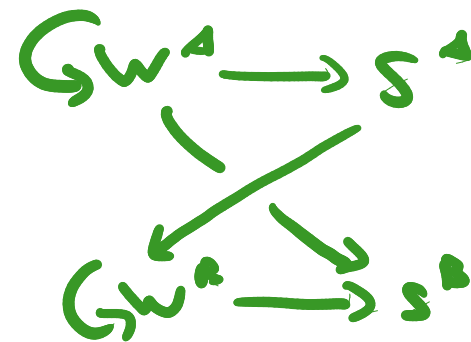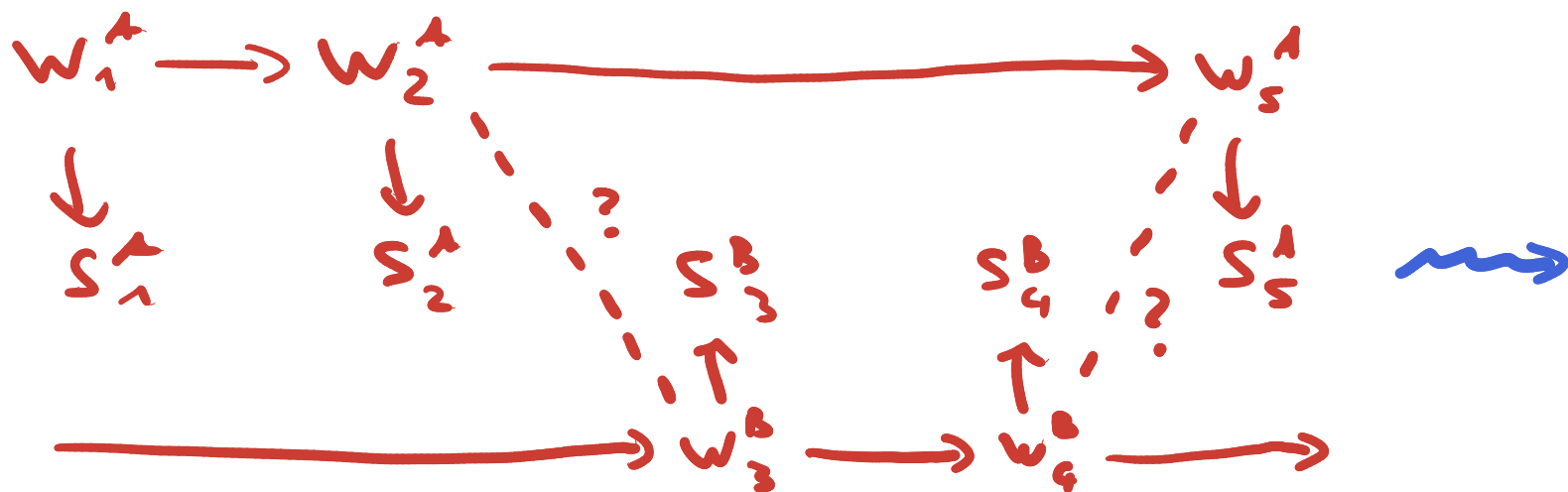# Modular Mechanisms

A Causal Perspective
on the Data Generation Process
of Conversational Speech Data

Philipp Gabler

2021-08-26

MOTIVATION I

MODELS TAKING MULTI-SPEAKER
INTERACTION SERIOUSLY?

MOTIVATION II

APPROACHES TAKING INTO ACCOUNT
DIFFERENCES IN CAUSAL STRUCTURE?

ASSOCIATION
↓
PREDICTION

⤳

MECHANISM
↓
INTERVENTION/
COUNTERFACTUAL

The Story of Speech Recognition

Read Speech

Spontaneous Speech

W ⟶ S

Anti-causal

S ⟶ W

Causal

W: Written Signal, S: Spoken Signal
(Both Observed)

The Story of Speech Recognition

Read Speech

Spontaneous Speech

Anti-Causal

Causal

W: Written Signal, S: Spoken Signal
(Both Observed)

The Story of Speech Recognition

Read Speech   Spontaneous Speech

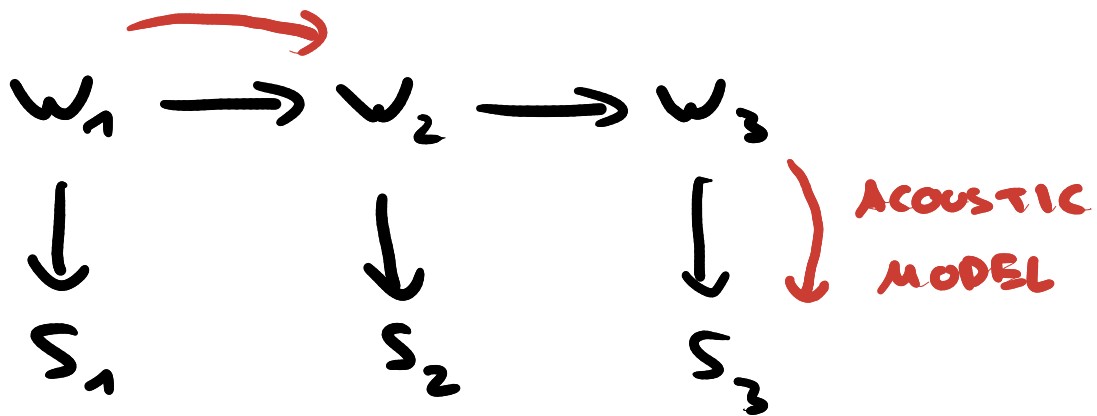Language Model

$W_1 \longrightarrow W_2 \longrightarrow W_3$

$W_1 \downarrow \quad W_2 \downarrow \quad W_3 \downarrow$ Acoustic Model

$S_1 \qquad S_2 \qquad S_3$

"Standard" HMM

??

$S_1 \longrightarrow S_2 \longrightarrow S_3$

Annotation Model ?

$S_1 \downarrow \quad S_2 \downarrow \quad S_3 \downarrow$

$W_1 \qquad W_2 \qquad W_3$

? ? ?

ABSTRACTION!

- LATENT STATE

$\Lambda \circlearrowright$

$\Lambda \longrightarrow S$

OBSERVATIONS

$\Phi \longrightarrow W$

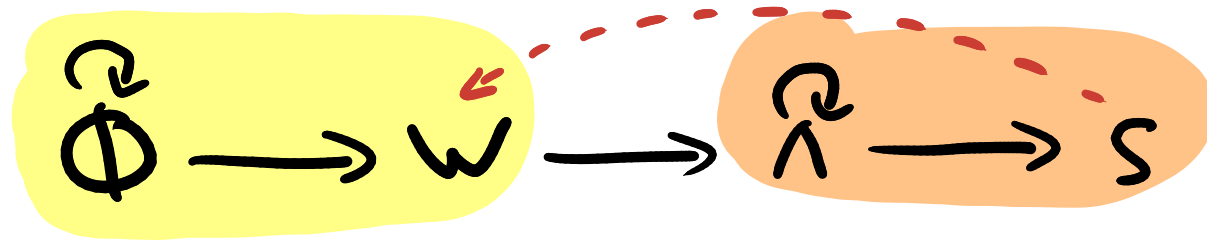$\Phi \circlearrowleft$ - INTERNALLY AUTOREGRESSIVE
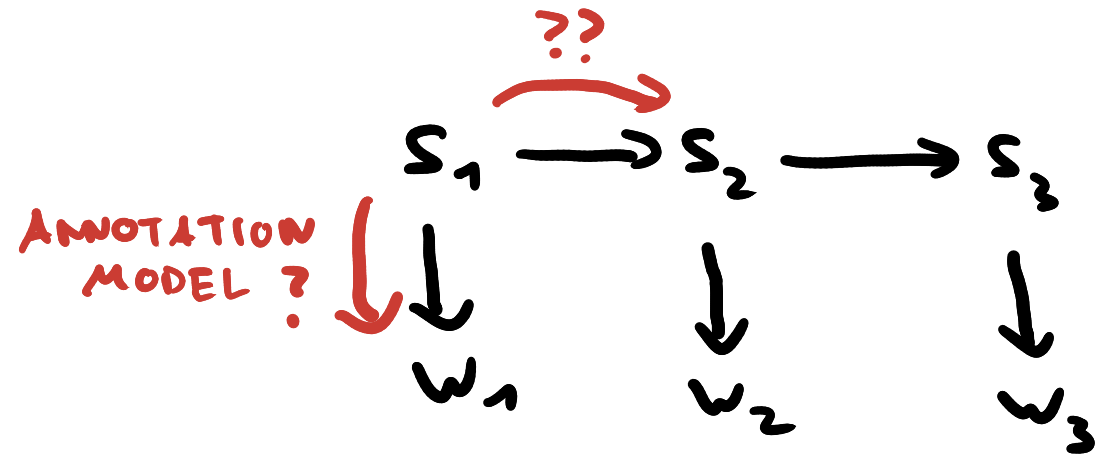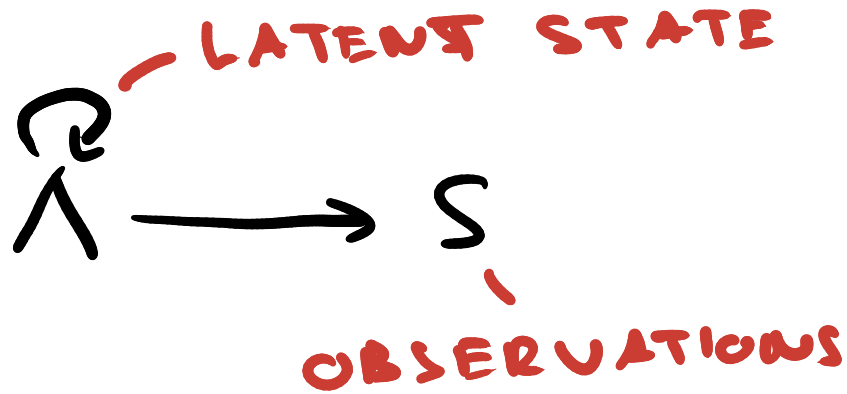
SPEAKER MODULE

WRITER MODULE
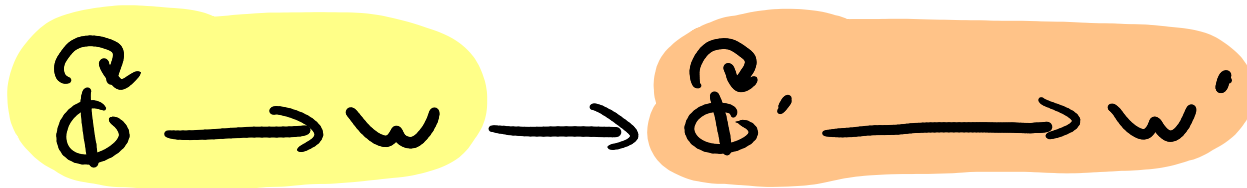
Λ, Φ : UNOBSERVED INTERNAL STATES

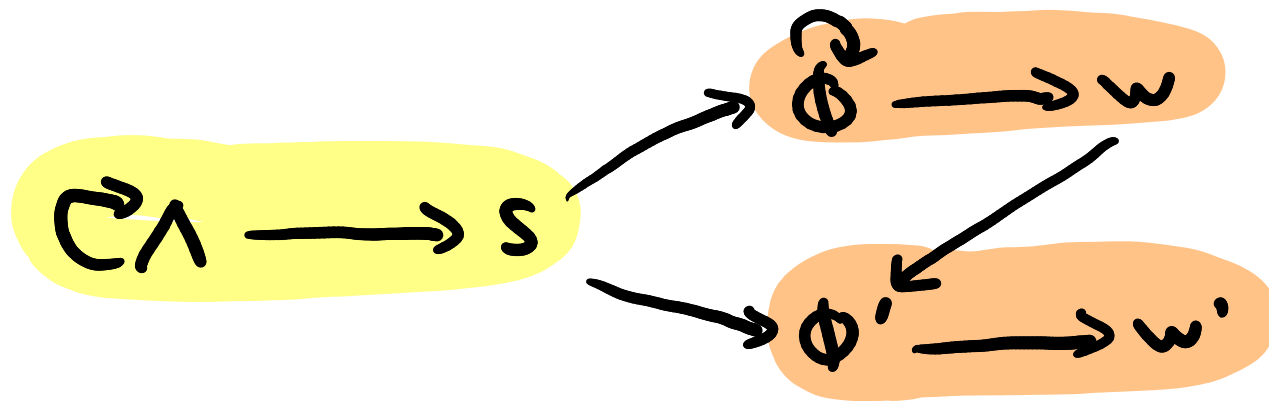# VERTICAL COUPLING

READ
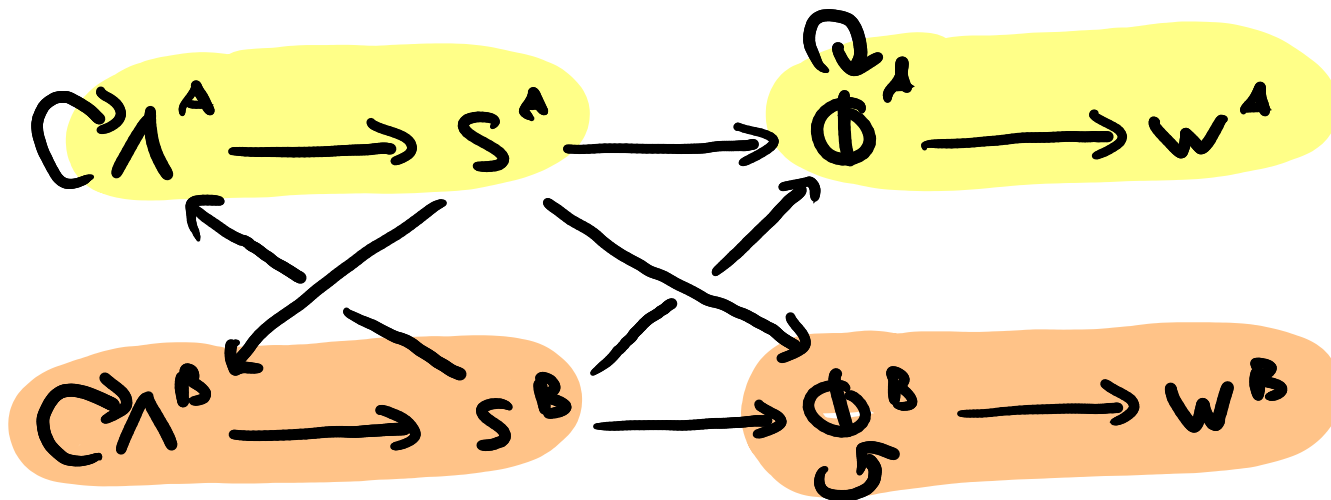SPEECH
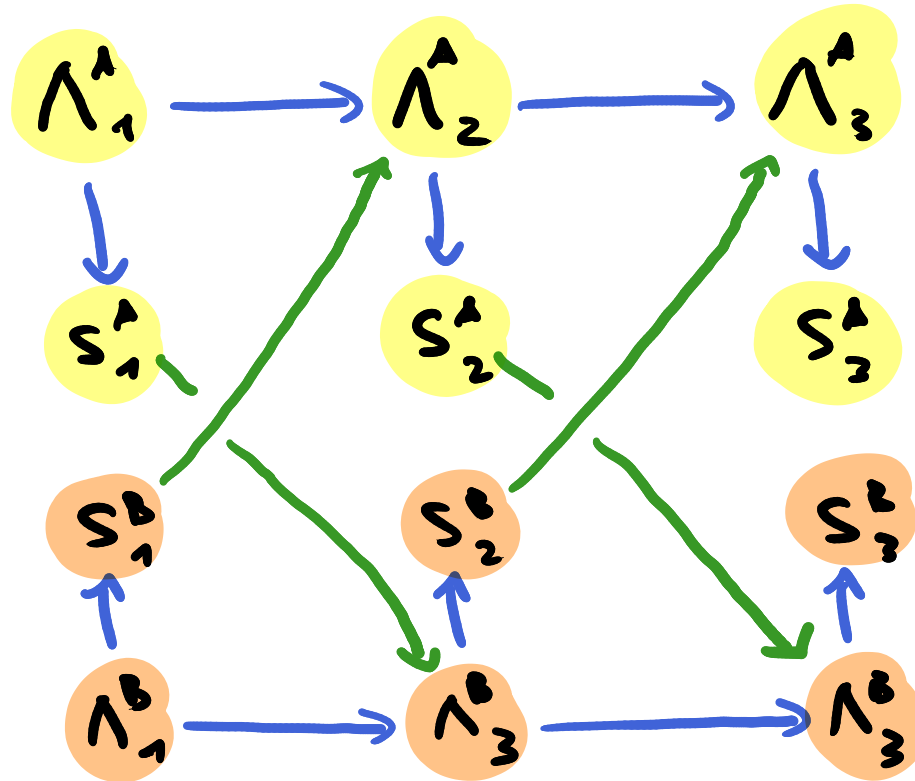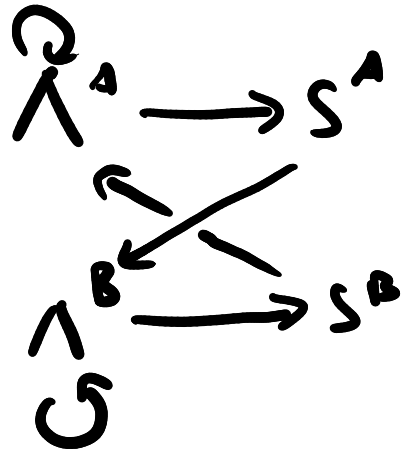
SPONTANEOUS
SPEECH

TRANSCRIPTION

Depthwise Coupling

Multiple annotation passes

Conversational speech
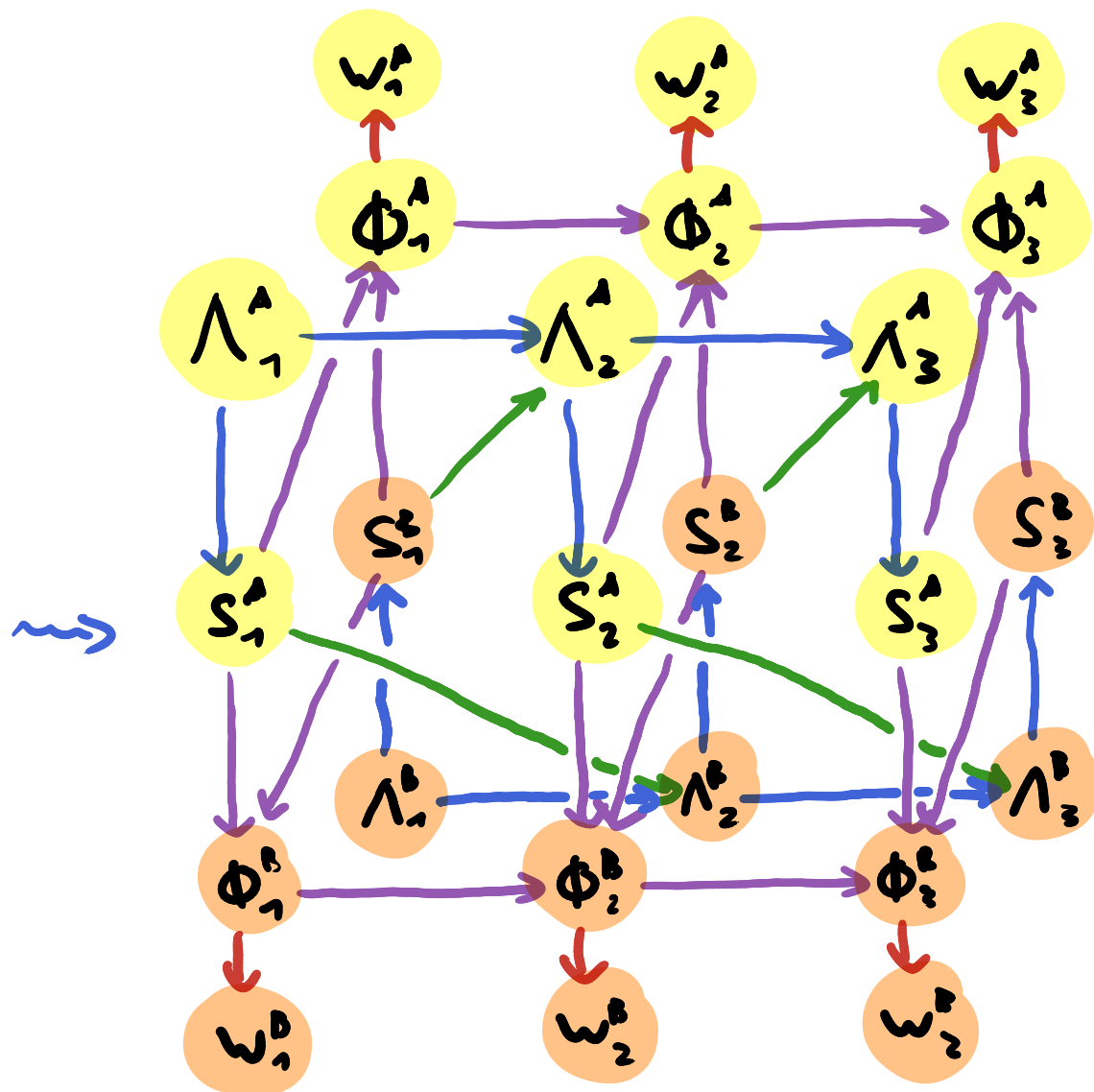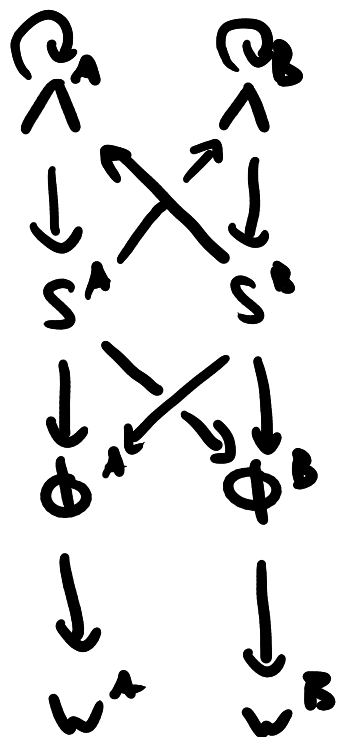
EXPANSION IN TIME

EXPANSION IN TIME

ANNOTATOR PERCEPTION

ANNOTATOR PRODUCTION

EXPANSION IN TIME

Speaker Perception

Speaker Production

Annotator Perception

Annotator Production

Contextual & Proxy Features

Context → Λ

Speaker → Λ, S

Λ → S → Φ → W

Annotator → Φ, W

Λ ⤏ IPU Discourse

S → Prosody Alignment

W → De-aligned Normalized Grammat. Features
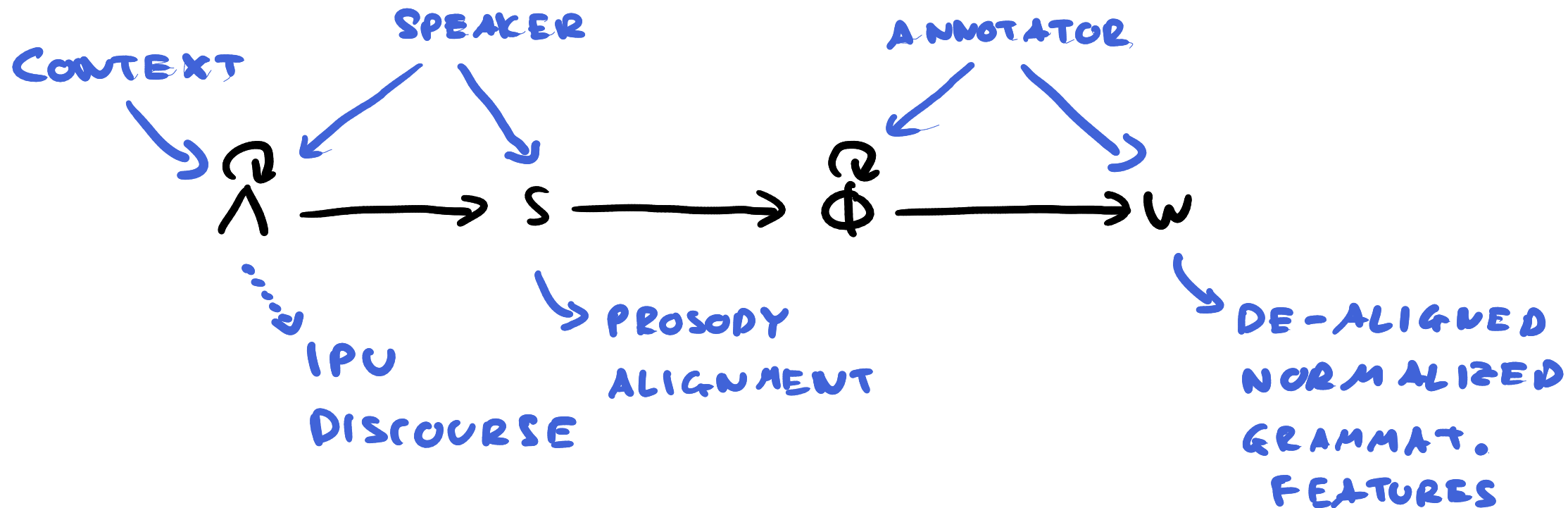
Takeaways

- Flexible tool box of conditional language models

- Clearly distinguish production and perception

- Latent & observed model for each participant

- Few assumptions made, choices left open

Question I: Structure

- HMM

- Attention

- Convolution

- Directed/Undirected

QUESTION II: REPRESENTATION

○ INTERNAL STATES:
  ↳ NEURAL NETS ($\hat{\equiv}$ VECTORS + LIN ALG)
  ↳ DISCRETE

○ ALIGNMENT
  ↳ SEQ2SEQ, CTC, PROBABILISTIC

○ SHARING OF LINGUISTIC KNOWLEDGE
  ↳ ADDITIONAL CONSTRAINTS
  ↳ MUTUAL INFORMATION / INF. BOTTLENECK

Question III: Inference

- Fully Generative
  $\hookrightarrow$ MCMC, VI, Message Passing...
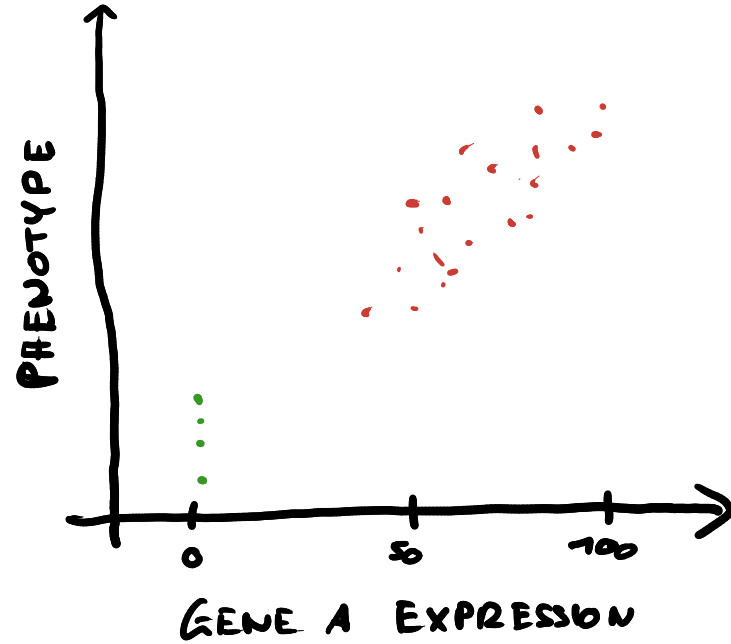
- Convert to Discriminative?

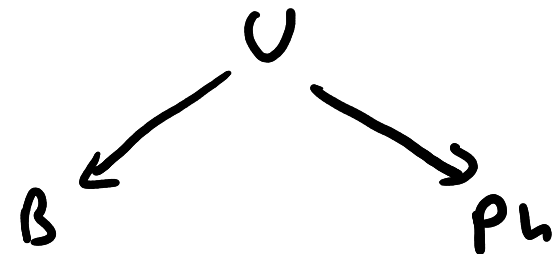- Strict-Sense LM = Marginal?

Further Considerations
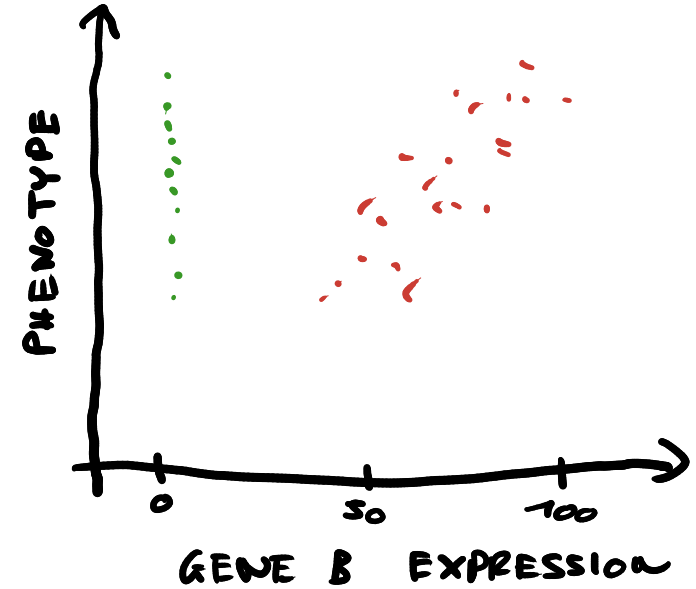
- Low Resources

- Transfer Learning

- Model / Investigate Annotator Errors?

- First Try on:
  - ↳ Movie Dialogs
  - ↳ Chat Corpora   ?
  - ↳ Auto-Subtitles

Appendix

TWO GENES

Classify What's Wrong

- Population Shift
- Acquisition Shift
- Annotation Shift
- Prevalence Shift
- Manifestation Shift

$$Z \longrightarrow X \longrightarrow Y$$

ANATOMY · IMAGE · TARGET

$$X \longleftarrow Z \longleftarrow Y$$

IMAGE · ANATOMY · CONDITION