SPEAKER

SPEAKER   ANNOTATOR
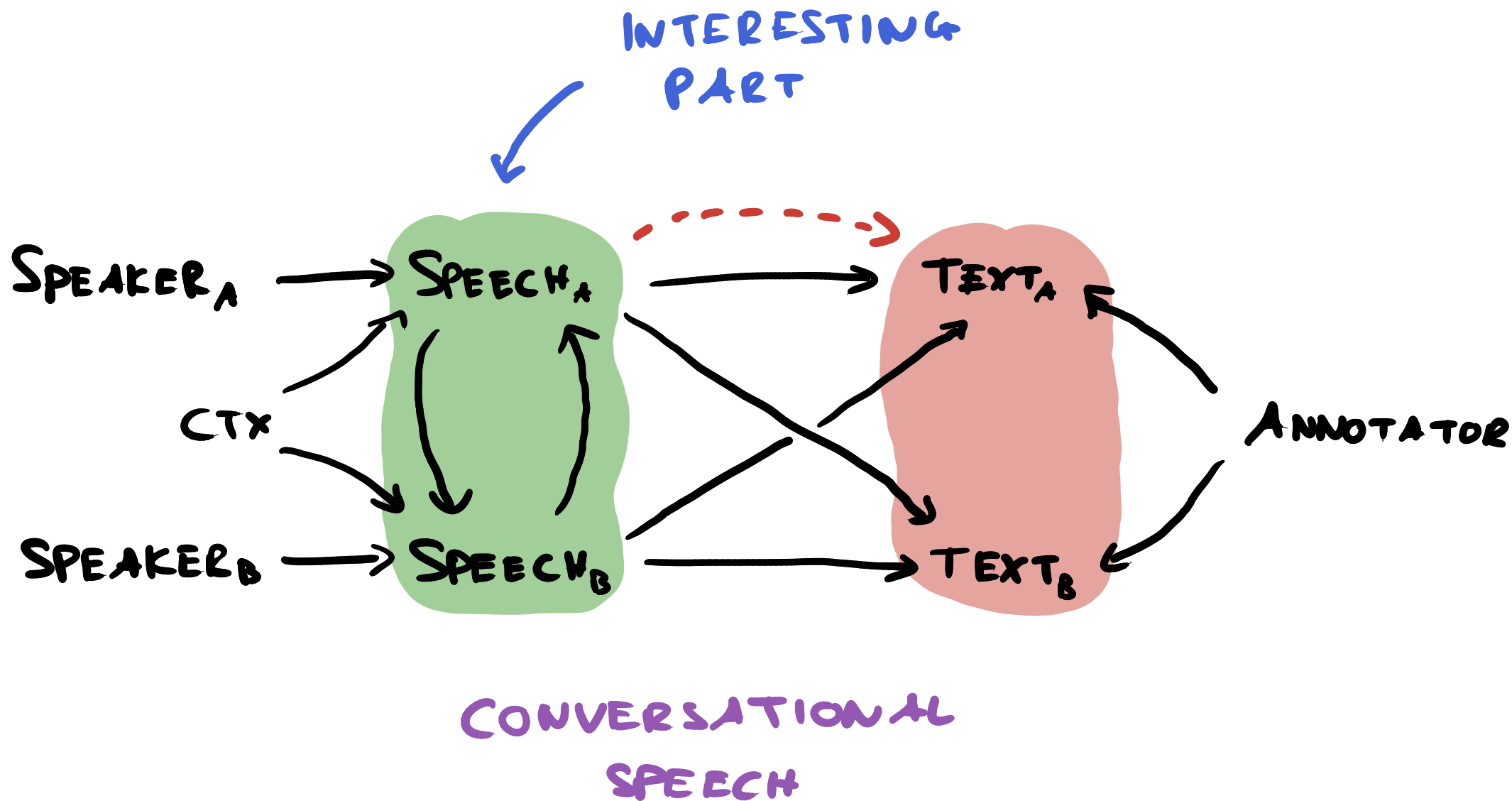
TEXT ——→ SPEECH

SPEECH ——→ TEXT

CTX

READ
SPEECH

(ANTICAUSAL)

SPONTANEOUS
SPEECH

(CAUSAL)

"COUPLED HMM" VERSION
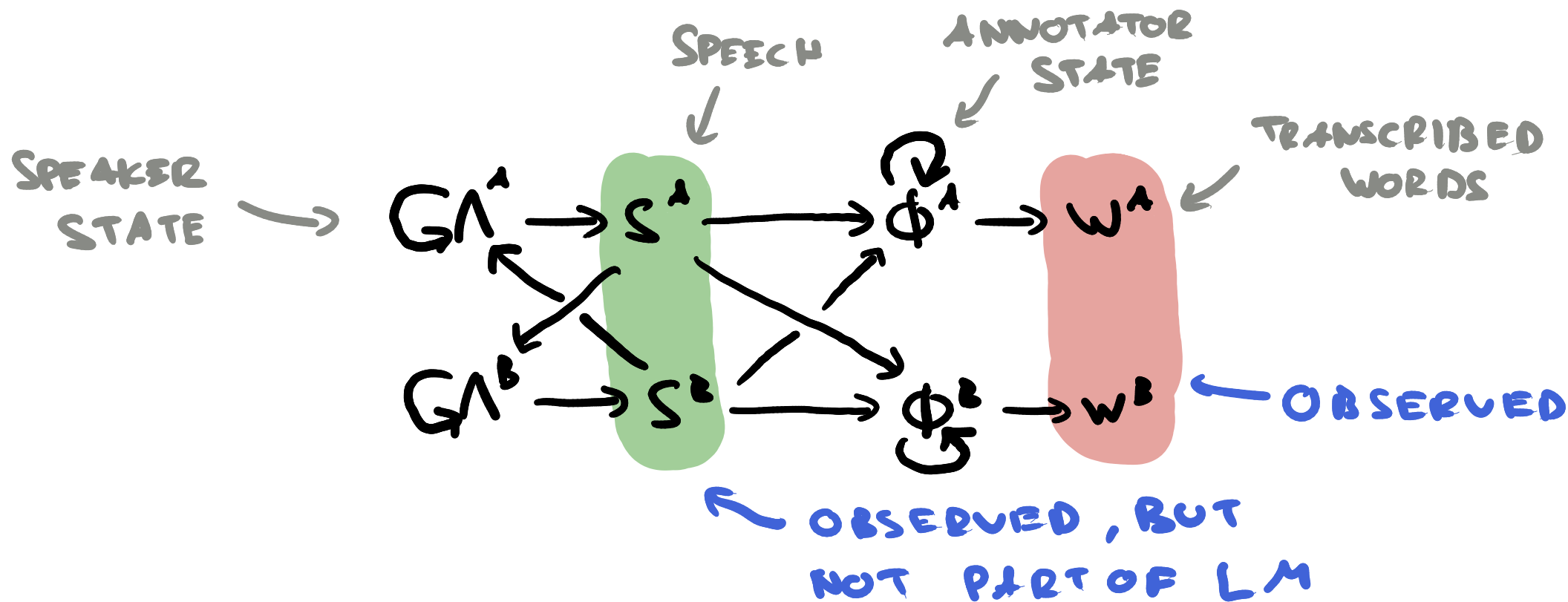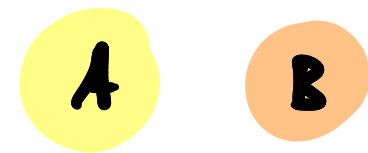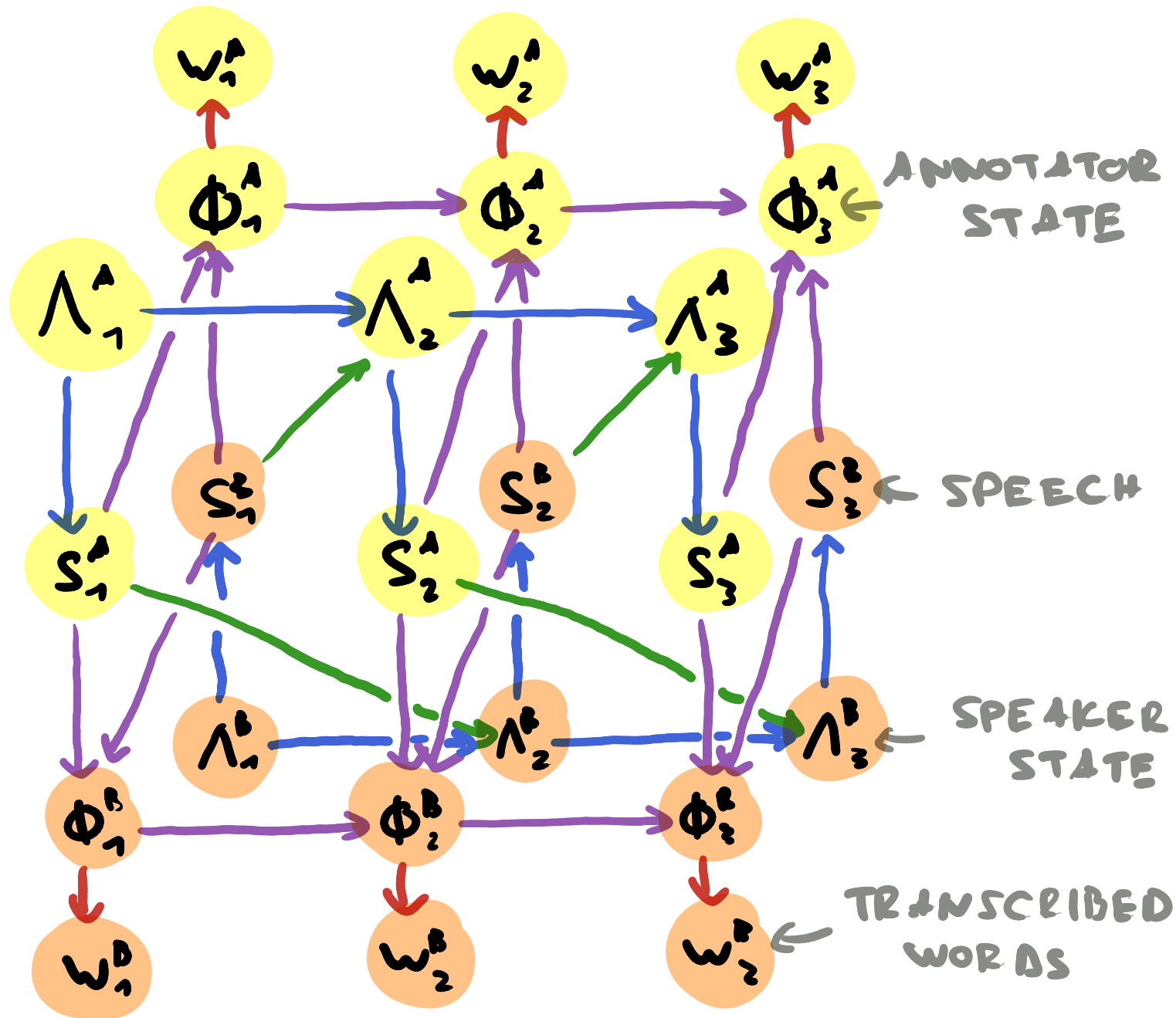
Legend:
- A (yellow)
- B (orange)
- Speaker Perception (green)
- Speaker Production (blue)
- Annotator Reception (purple)
- Annotator Production (red)

W_1^A, W_2^A, W_3^A

Φ_1^A, Φ_2^A, Φ_3^A — Annotator State

Λ_1^A, Λ_2^A, Λ_3^A

S_1^B, S_2^B, S_3^B — Speech

S_1^A, S_2^A, S_3^A

Λ_1^B, Λ_2^B, Λ_3^B — Speaker State

Φ_1^B, Φ_2^B, Φ_3^B

W_1^B, W_2^B, W_3^B — Transcribed Words

SPEECH MODEL:

OBSERVED

$$p(\Lambda_i^A \mid \Lambda_{i-1}^A, S_{i-1}^B) \; p(S_i^A \mid \Lambda_i^A)$$

ANNOTATION MODEL:

OBSERVED

$$p(\Phi_i^A \mid \Phi_{i-1}^A, S_i^A, S_i^B) \; p(W_i^A \mid \Phi_i^A)$$

POSSIBLE EXTENSIONS:

o LARGER WINDOWS: $p(\Lambda_i^A \mid \Lambda_{i-1}^A, S_{i-1}^B, S_{i-2}^B \ldots)$

o OBSERVED PROSODIC FEATURES: $p(P_i^A \mid S_i^A)$

Open Questions:

- What are the Λ_i? Chunk, sentence, discourse act, construction, ...?

- History/Context of Syntax, Semantics, Phonetics

- Markovian or attention-based?

- Correspondence Λ_i / W_i: Is a textual latent space enough? Reconstructed faithfully by the annotators?

INTERACTION OF DISCOURSE, MEANING, AND PHONOLOGY WITH SYNTAX:

↳ REDUCTION, PROSODY, & OTHER INTERACTIONS

↳ HOW IS INFORMATION TRANSMITTED?
CONSTRUCTIONS, REDUNDANCY, ON-LINE SYNTAX

↳ DISCOVERY OF PATTERNS/CONSTRUCTIONS IN SPONTANEOUS SPEECH GRAMMAR

MODELLING PERSPECTIVE:

⮡ WHAT UNITS? HOW TO INTEGRATE CONTEXTS, LONG-RANGE DEPENDENCIES?

⮡ MULTI-CHANNEL-SETTING: INTERACTING "DYNAMICAL SYSTEM"

⮡ LOW-RESOURCE PROBLEM: TRANSFER APPROACHES? EXPERTS/RULES + UNSUPERVISED?

PLAN:

① Try out "coupled model" on textual data for conceptual exploration (e.g. chat corpora, subtitles)

② Possible directions with GRASS

> Prosody integration

> Interior contexts (attention, "cache")

> Exterior contexts (speakers, world knowledge/common sense)

> Investigate low-resource problem