Improved Neural Protoform Reconstruction via Reflex Prediction

Liang Lu, Jingzhi Wang, David R. Mortensen

Language Technologies Institute, Carnegie Mellon University

{lianglu, jingzhi3, dmortens}@cs.cmu.edu

LREC-COLING 2024

Turino, Italy

Abstract

Protolanguage reconstruction is central to historical linguistics. The comparative method, one of the most influential theoretical and methodological frameworks in the history of the language sciences, allows linguists to infer protoforms (reconstructed ancestral words) from their reflexes (related modern words) based on the assumption of regular sound change. Not surprisingly, numerous computational linguists have attempted to operationalize comparative reconstruction through various computational models, the most successful of which have been supervised encoder-decoder models, which treat the problem of predicting protoforms given sets of reflexes as a sequence-to-sequence problem. We argue that this framework ignores one of the most important aspects of the comparative method: not only should protoforms be inferable from cognate sets (sets of related reflexes) but the reflexes should also be inferable from the protoforms. Leveraging another line of research—reflex prediction—we propose a system in which candidate protoforms from a reconstruction model are reranked by a reflex prediction model. We show that this more complete implementation of the comparative method allows us to surpass state-of-the-art protoform reconstruction methods on three of four Chinese and Romance datasets.

Background

Llama

Alpaca

Images: Wikipedia

capra 'goat'
/kapra/
(Italian)

cabra 'goat'
/kabra/
(Spanish)

Example words: (Campbell 2021)

capra 'goat'
/kapra/
(Italian)

cabra 'goat'
/kabra/
(Spanish)

capo 'end, chief'
/kapo/
(Italian)

cabo 'end, tip'
/kabo/
(Spanish)

⋮

⋮

Example words: (Campbell 2021)

A window into human past

Evolutionary Biology

Population Genetics

Historical Linguistics

A window into human past

Evolutionary Biology

Population Genetics

Historical Linguistics

Some Definitions

Protolanguage: a historical language
Daughter languages: descendants of a protolanguage
Protoform: a reconstructed[1] ancestral word
Reflexes: descendent words in daughter languages
Cognate set: a set of reflexes with the same ancestor

Some Definitions

capra 'goat'
(Latin)

capra
/kapra/
(Italian)

cabra
/kabra/
(Spanish)

cabra
/kabra/
(Portuguese)

chèvre
/ʃɛvr(ə)/
(French)

Example: (Campbell 2021)

Some Definitions

Protoform → ***capra*** 'goat'

(Latin) ← Protolanguage

capra
/kapra/
(Italian)

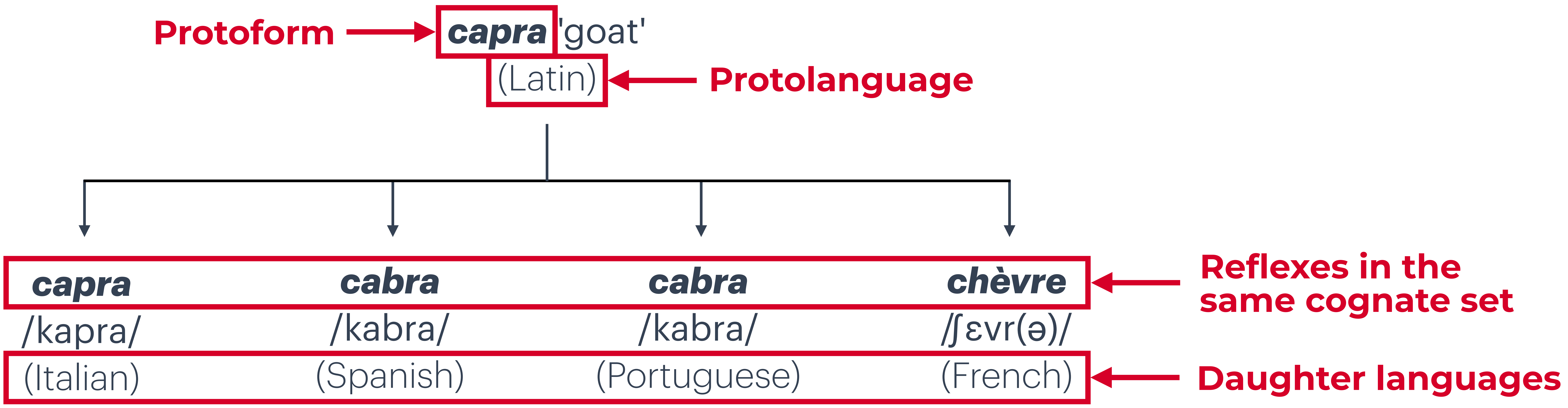
cabra
/kabra/
(Spanish)

cabra
/kabra/
(Portuguese)

chèvre
/ʃɛvr(ə)/
(French)

Example: (Campbell 2021)

Some Definitions



Example: (Campbell 2021)

The Comparative Method

The comparative method (Anttila, 1989; Campbell, 2021) uses reflexes in cognate sets to reconstruct the protoforms in a way that:

- **Maximizes the regularity of sound changes** from reconstructions to reflexes
- **Minimizes the phonetic edits** between the reconstructions and their reflexes

"Every sound change, in so far as it proceeds mechanically, is completed **in accordance with laws admitting of no exceptions**; i.e. the direction in which the change takes place is always the same for all members of a language community, apart from the case of dialect division, and all words in which the sound subject to change occurs in the same conditions are affected by the change without exception."

—H. Osthoff and K. Brugmann, *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen* i, Leipzig, 1878 (quoted in Szemerény, 1996, p. xiii)

Computational Protoform Reconstruction

- Proposed as a **computational task** (Durham and Rogers, 1969)
- Sound change **probabilistic models with a Monte Carlo inference algorithm** that operates on phylogenetic trees (Bouchard-Côté et al., 2013)
- **Sequence comparison and phonetic alignment** (List et al., 2022a)
- **Conditional random field** to label each position in the reflex with a protoform token (Ciobanu and Dinu, 2018; Ciobanu et al., 2020)
- **Sequence-to-sequence formulation:** concatenate the whole cognate into one sequence, with separators and daughter language tags (Meloni et al., 2021)

Input: *[Cantonese]:mei̯˨*[Mandarin]:mei̯˧˥*[Wu]:me̞˩˧*

(the 媚 cognate set from WikiHan)

Output: mij^{3}

Neural Protoform Reconstruction

Input: *[Cantonese]:mei̯˧*[Mandarin]:mei̯˥˩*[Wu]:me̞˧˥*
Output: mij^3

- **RNN with language embedding** (Meloni et al., 2021)
- **Transformer** (Kim et al., 2023)
- **VAE** (Variational Autoencoder) (Cui et al., 2022)

Neural Protoform Reconstruction

Input: *[Cantonese]:mei̯˧*[Mandarin]:mei̯˥˩*[Wu]:me̝˧˦*
Output: mij³

- **RNN with language embedding** (Meloni et al., 2021)
- **Transformer** (Kim et al., 2023)
- **VAE** (Variational Autoencoder) (Cui et al., 2022)

Sequence-to-sequence

Neural Protoform Reconstruction

Input: *[Cantonese]:mei̯˧*[Mandarin]:mei̯˥˩*[Wu]:me̞˧˥*
Output: mij^3

- **RNN with language embedding** (Meloni et al., 2021)
- **Transformer** (Kim et al., 2023)
- **VAE** (Variational Autoencoder) (Cui et al., 2022)

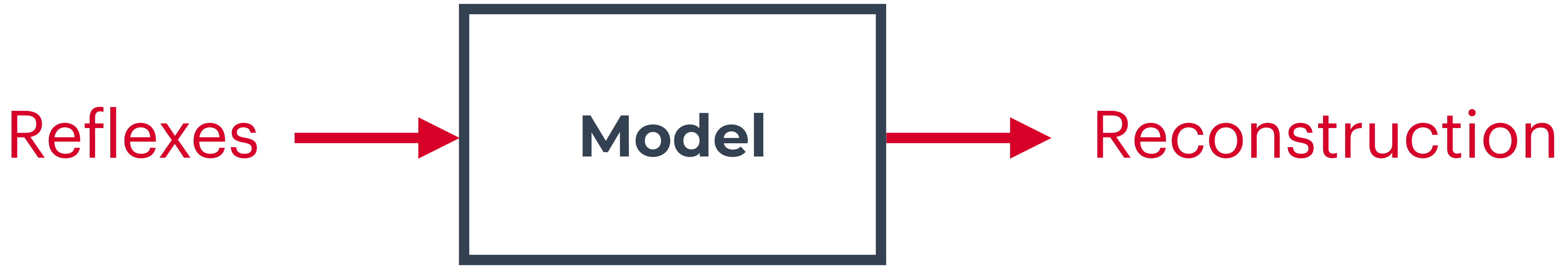
Sequence-to-sequence

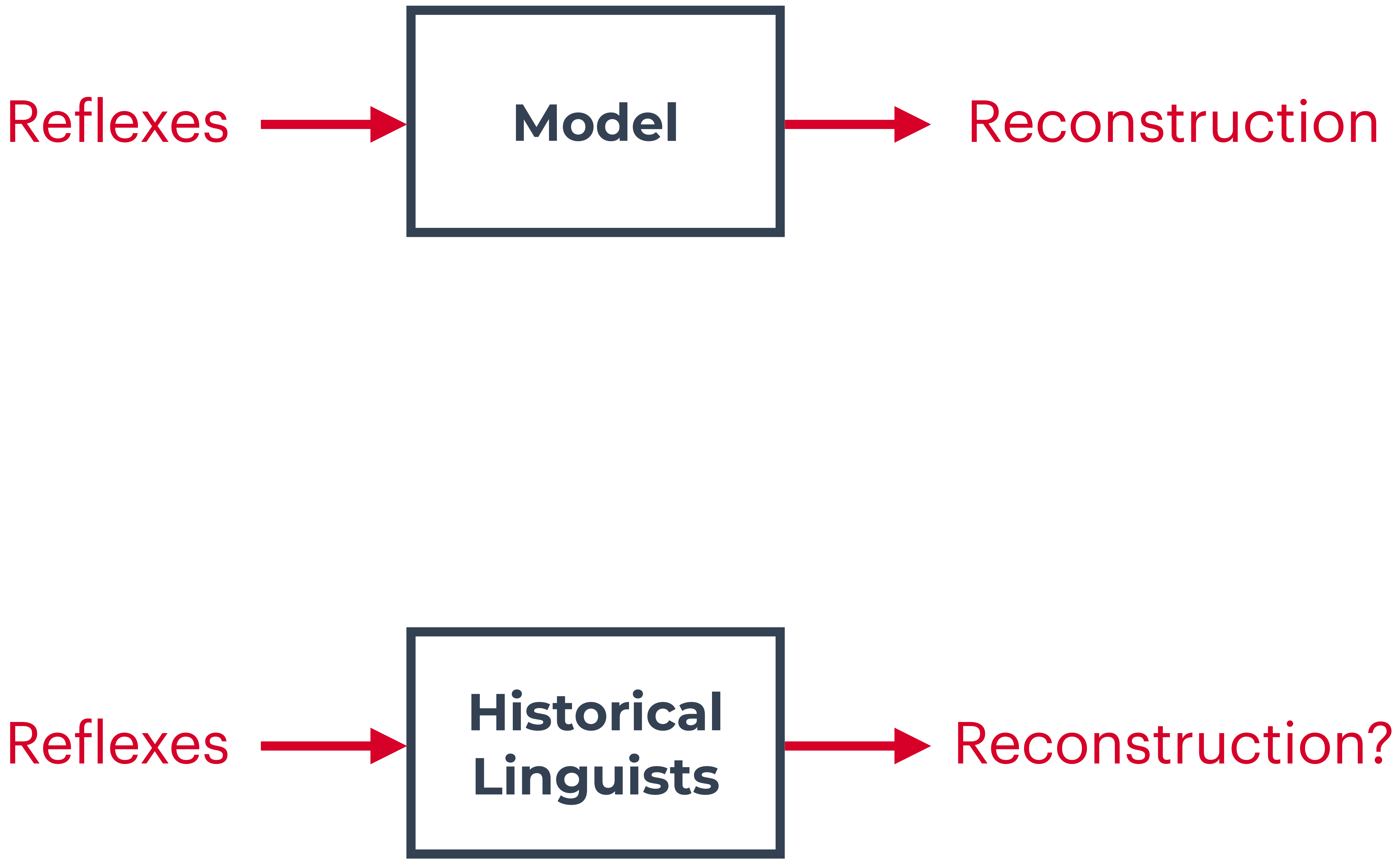
Input:	[Cantonese]	m	e	i̯	˧
	[Mandarin]	m	e	i̯	˥˩
	[Wu]	m	e̞	-	˧˥
	[Middle Chinese]	[MASK]	[MASK]	[MASK]	[MASK]
Output:	[Middle Chinese]	m	i	j	³

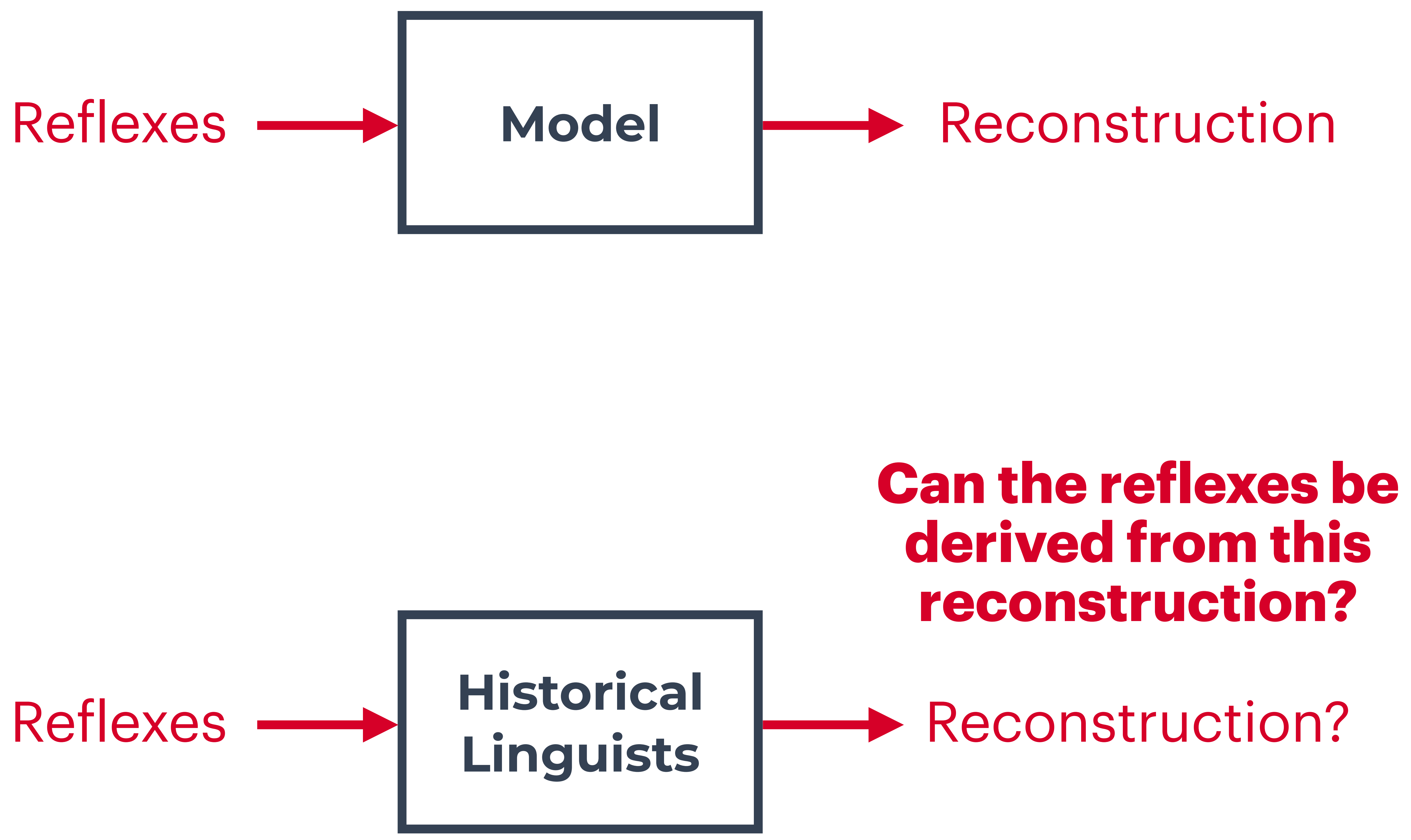
- **Cognate Transformer** (Akavarapu and Bhattacharya, 2023)

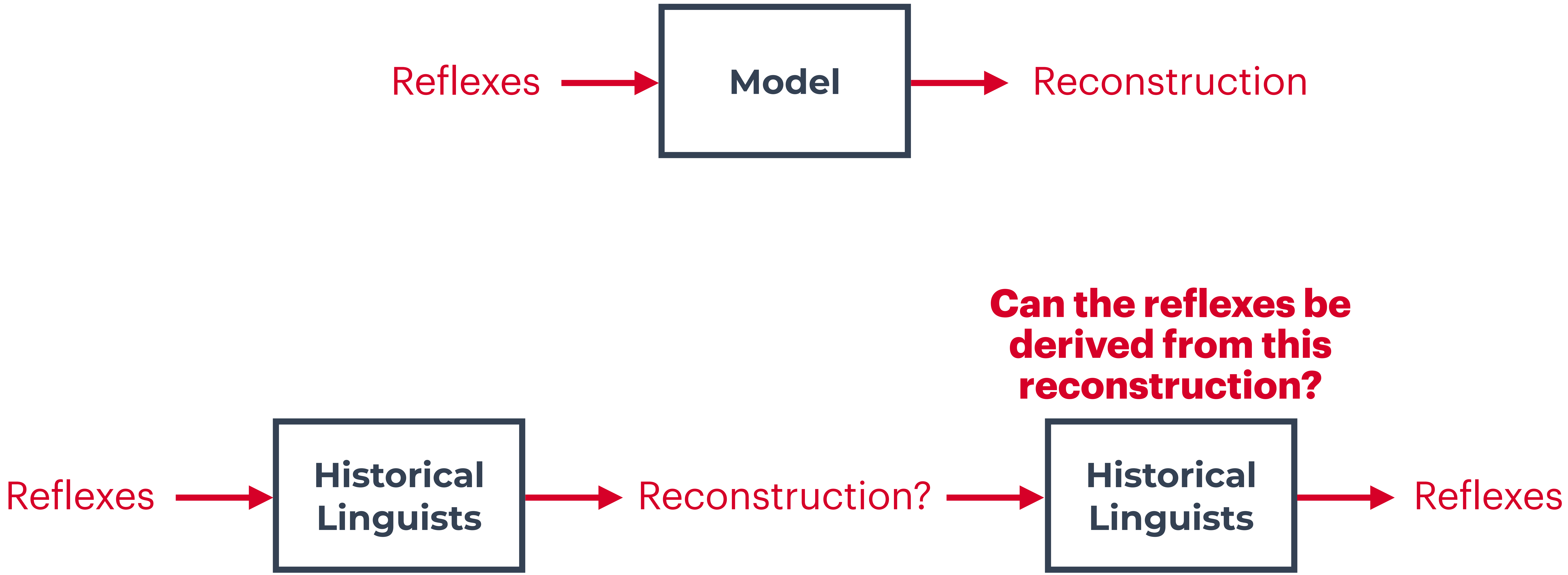
Multi-sequence encoder with classifier head

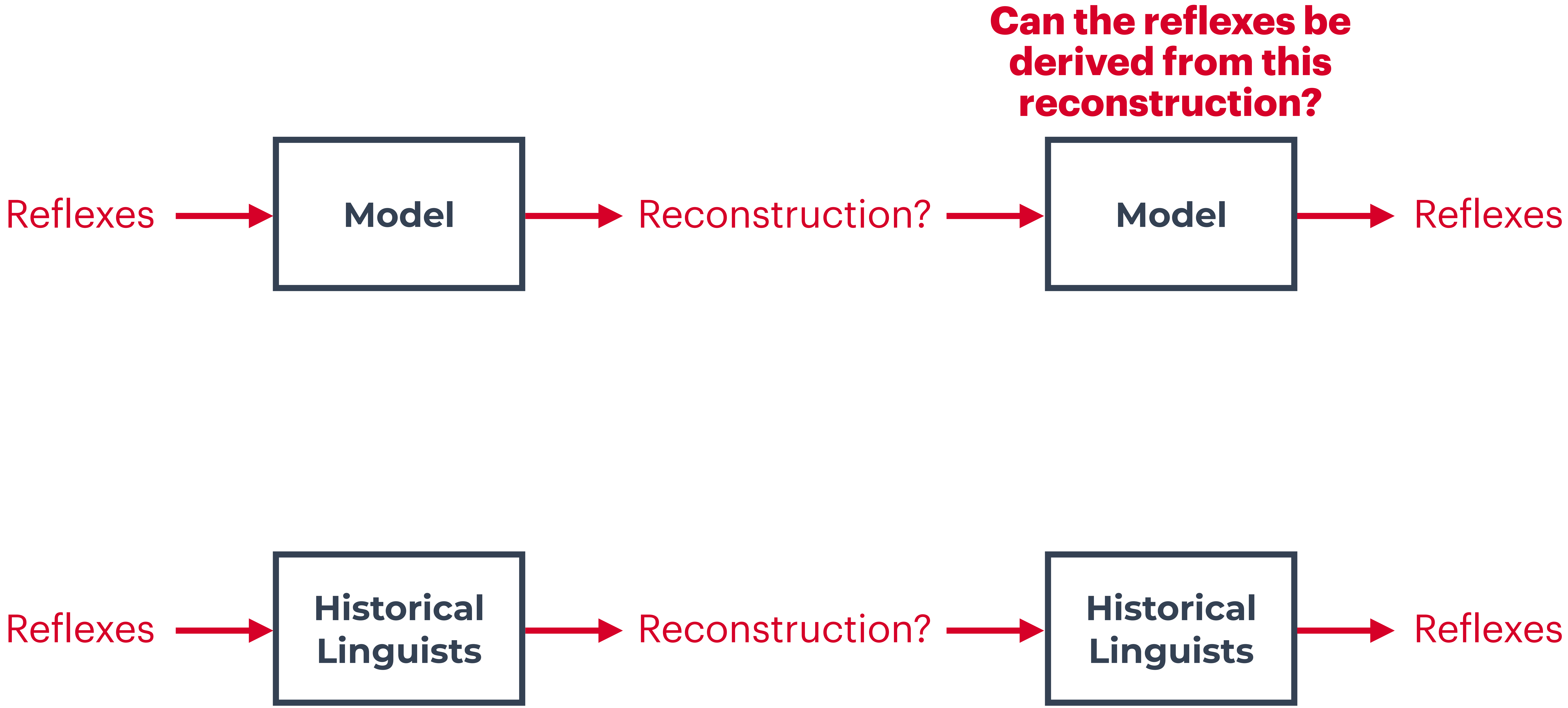
Motivation

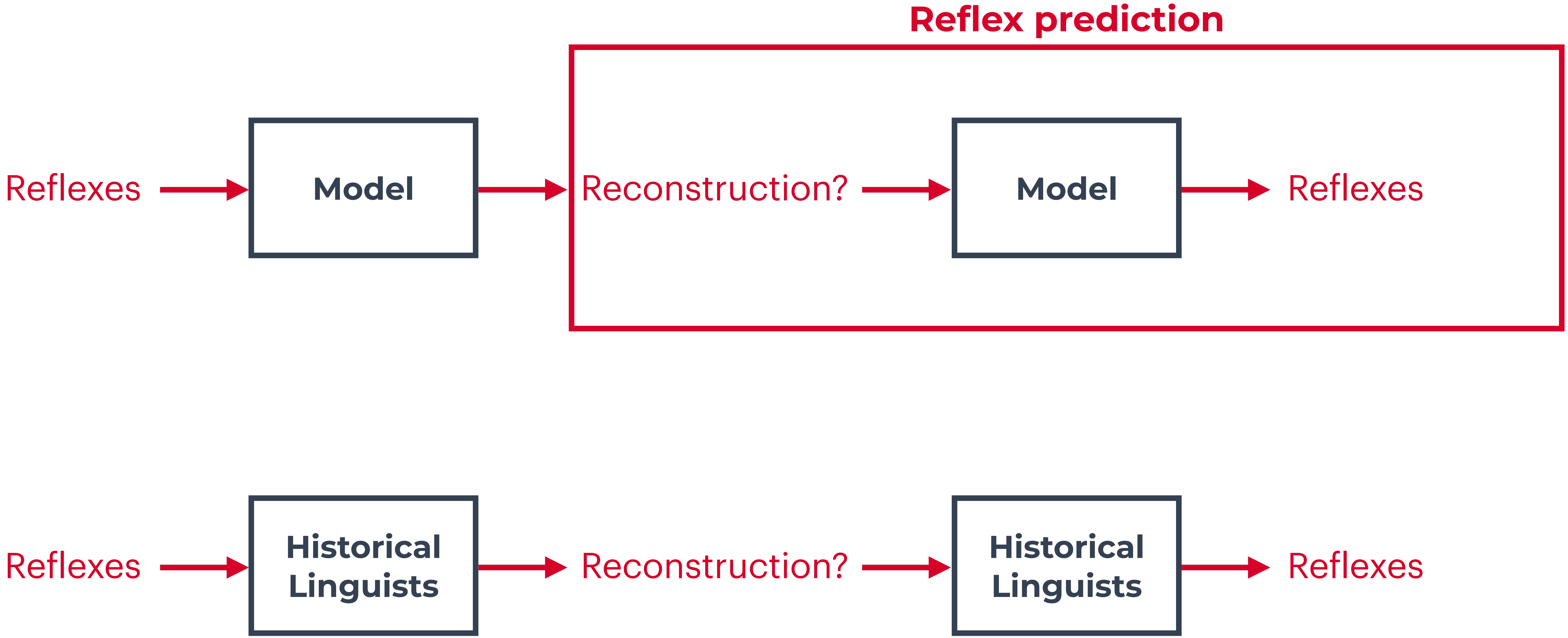

Reflexes
Model
Reconstruction


Reflexes
Model
Reconstruction
Reflexes
Historical Linguists
Reconstruction?


Reflexes
Model
Reconstruction
Can the reflexes be derived from this reconstruction?
Reflexes
Historical Linguists
Reconstruction?


Reflexes
Model
Reconstruction
Can the reflexes be derived from this reconstruction?
Reflexes
Historical Linguists
Reconstruction?
Historical Linguists
Reflexes


Can the reflexes be derived from this reconstruction?
Reflexes
Model
Reconstruction?
Model
Reflexes
Reflexes
Historical Linguists
Reconstruction?
Historical Linguists
Reflexes


Reflex prediction
Reflexes
Model
Reconstruction?
Model
Reflexes
Reflexes
Historical Linguists
Reconstruction?
Historical Linguists
Reflexes

Computational Reflex Prediction

- **Rule-based** (Marr and Mortensen 2020, 2023)
- Semi-automatic: **automatic alignment and identification of sound correspondences** on **manually annotated cognate sets** (Bodt and List, 2022)
- **LSTM encoder-decoder** augmented with part of speech and word embeddings (Cathcart and Rama, 2020)
- Replication of Cathcart and Rama (2020) with **GRU and Transformer** (Arora et al., 2023)

Computational Reflex Prediction

- **Rule-based** (Marr and Mortensen 2020, 2023)
- Semi-automatic: **automatic alignment and identification of sound correspondences** on **manually annotated cognate sets** (Bodt and List, 2022)
- **LSTM encoder-decoder** augmented with part of speech and word embeddings (Cathcart and Rama, 2020)
- Replication of Cathcart and Rama (2020) with **GRU and Transformer** (Arora et al., 2023)

Representing Reflex Prediction

Reconstruction

Input: *[Cantonese]:mei̯˨*[Mandarin]:mei̯˨˩˦*[Wu]:me̞˧˥*

Output: mij³

Reflex Prediction

Input: mij³

Output:

Input: mij³

Output:

Input: mij³

Output:

Representing Reflex Prediction

Reconstruction

Input: *[Cantonese]:mei̯˧*[Mandarin]:mei̯˥˩*[Wu]:me̝˧˥*
Output: mij^3

Reflex Prediction

Input: [Cantonese] mij^3
Output: mei̯˧

Input: [Mandarin] mij^3
Output: mei̯˥˩

Input: [Wu] mij^3
Output: me̝˧˥

Modelling the Comparative Method

Workflow

Technique

?

Predict the protoform ⟷ Sequence-to-sequence transduction

Modelling the Comparative Method

Workflow

Technique

1. Propose multiple protoform candidates

2. Verify the phonetic plausibility of the candidates

3. Adjust the likelihood of each candidate and make a prediction

Modelling the Comparative Method

Workflow		Technique
1. Propose multiple protoform candidates	⟷	Sequence-to-sequence transduction with beam search
2. Verify the phonetic plausibility of the candidates		
3. Adjust the likelihood of each candidate and make a prediction		

Modelling the Comparative Method

Workflow		Technique
1. Propose multiple protoform candidates	⟷	Sequence-to-sequence transduction with beam search
2. Verify the phonetic plausibility of the candidates	⟷	Reflex prediction
3. Adjust the likelihood of each candidate and make a prediction		

Modelling the Comparative Method

Workflow		Technique
1. Propose multiple protoform candidates	⟷	Sequence-to-sequence transduction with beam search
2. Verify the phonetic plausibility of the candidates	⟷	Reflex prediction
3. Adjust the likelihood of each candidate and make a prediction	⟷	Reranking

A Reranked Reconstruction System

Reflexes in the 必 *pit*入 ‘must’ cognate set							
Cantonese	Gan	Hakka	Jin	Mandarin	Hokkien	Wu	Xiang
piːt˥	pit˥	pit˨	pi̯əʔ˨	pi˥˩	pit˧	pi̯ɪʔ˥	pi˨˦

A Reranked Reconstruction System

Reflexes in the 必 *pit*入 'must' cognate set							
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)

A Reranked Reconstruction System

Reflexes in the 必 *pit*入 'must' cognate set							
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**

beam search reconstruction

(beam size k = 5)

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)

A Reranked Reconstruction System

Reflexes in the 必 *pit*入 'must' cognate set							
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**

beam search
reconstruction

(beam size k = 5)

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)

A Reranked Reconstruction System

Reflexes in the 必 *pit*入 'must' cognate set							
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**
piːt˥	**pit˥**	**pit˧**	**pjə̯ʔ˧**	**pi˥˩**	**pit˧**	**piɪ̯ʔ˥**	**pi˨˦**

beam search
reconstruction

(beam size k = 5)

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)

A Reranked Reconstruction System

Reflexes in the 必 *pit*入 'must' cognate set							
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**

beam search reconstruction

(beam size k = 5)

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Reflex Prediction (based on protoform candidates)								
Cantonese	Gan	Hakka	Jin	Mandarin	Hokkien	Wu	Xiang	r_i
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

Bold: correct protoform or reflexe
mᵢ: model score (sequence log probability)
rᵢ: reranker score (reflex prediction accuracy)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

reflex prediction →

Reflex Prediction (based on protoform candidates)								
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥˩	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	
pɐt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	
pei̯˧	pi˧˥	pi˥	pi˦˥	**pi˥˩**	pi˧˩	pi˧˥	pi˦˥	
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

reflex prediction

Reflex Prediction (based on protoform candidates)								
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥˩	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	
pɐt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	
pei̯˧	pi˧˥	pi˥	pi˧˥	**pi˥˩**	pi˧˩	pi˧˥	pi˧˥	
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

reflex prediction

Reflex Prediction (based on protoform candidates)								
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥˩	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	
pɐt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	
pei̯˧	pi˧˥	pi˥	pi˦˥	**pi˥˩**	pi˧˩	pi˨˧	pi˦˥	
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{bs}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

reflex prediction →

Reflex Prediction (based on protoform candidates)								
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥˩	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
pet˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	0.8750
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
pei̯˧	pi˨˦	pi˥	pi˧˥	**pi˥˩**	pi˨˩	pi˨˦	pi˧˥	0.1250
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

reflex prediction →

Reflex Prediction (based on protoform candidates)								
Cantonese	Gan	Hakka	Jin	Mandarin	Hokkien	Wu	Xiang	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥˩	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
pɐt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	0.8750
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
pei̯˧	pi˧˥	pi˥	pi˦˥	**pi˥˩**	pi˧˩	pi˧˦	pi˦˥	0.1250
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Reflex Prediction (based on protoform candidates)								
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥˩	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̝˨˦	0.2500
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̝˨˦	0.2500
pɐt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	0.8750
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̝˨˦	0.2500
pei̯˧	pi˨˦	pi˥	pi˦˥	**pi˥˩**	pi˨˩	pi˧˦	pi˦˥	0.1250
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

Reranking Result		
rank	$\hat{p}_i^{\mathbf{rk}}$	s_i

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)
s_i: adjusted score (scaled sum of model score and reranker score)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Reflex Prediction (based on protoform candidates)								
Cantonese	Gan	Hakka	Jin	Mandarin	Hokkien	Wu	Xiang	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥˩	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
pɐt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	0.8750
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
pei̯˧	pi˧˥	pi˥	pi˦˥	**pi˥˩**	pi˧˩	pi˨˧	pi˦˥	0.1250
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

Reranking Result		
rank	$\hat{p}_i^{\mathbf{rk}}$	s_i
	pjet入	0.2036
	pet入	0.0439
	pit入	0.5995
	pep入	-1.2383
	pij去	-1.4754

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)
s_i: adjusted score (scaled sum of model score and reranker score)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Reflex Prediction (based on protoform candidates)								
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥˩	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
pɐt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	0.8750
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̞˨˦	0.2500
pei̯˧	pi˧˥	pi˥	pi˦˥	**pi˥˩**	pi˨˩	pi˦	pi˦˥	0.1250
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

Reranking Result		
rank	$\hat{p}_i^{\mathbf{rk}}$	s_i
	pit入	0.5995
	pjet入	0.2036
	pet入	0.0439
	pep入	-1.2383
	pij去	-1.4754

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)
s_i: adjusted score (scaled sum of model score and reranker score)

A Reranked Reconstruction System

Beam Search

rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Reflex Prediction (based on protoform candidates)

Cantonese	Gan	Hakka	Jin	Mandarin	Hokkien	Wu	Xiang	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˨˩˦	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̝˨˦	0.2500
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̝˨˦	0.2500
pɐt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˨˩˦**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	0.8750
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̝˨˦	0.2500
pei̯˧	pi˧˥	pi˥	pi˦˥	**pi˨˩˦**	pi˧˩	pi˧˦	pi˦˥	0.1250
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˨˩˦**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

reranking

Reranking Result

rank	$\hat{p}_i^{\mathbf{rk}}$	s_i
0	**pit入**	0.5995
1	pjet入	0.2036
2	pet入	0.0439
3	pep入	-1.2383
4	pij去	-1.4754

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)
s_i: adjusted score (scaled sum of model score and reranker score)

A Reranked Reconstruction System

Beam Search		
rank	$\hat{p}_i^{\mathbf{bs}}$	m_i
0	pjet入	-0.1114
1	pet入	-0.2711
2	**pit入**	-0.5030
3	pep入	-1.5533
4	pij去	-1.6329

Reflex Prediction (based on protoform candidates)								
Cantonese	**Gan**	**Hakka**	**Jin**	**Mandarin**	**Hokkien**	**Wu**	**Xiang**	r_i
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥˩	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̝˨˦	0.2500
piːt˧	pi̯ɛt˥	pi̯et˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ɛt˧	**pi̯ɪʔ˥**	pi̯e̝˨˦	0.2500
pet˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	0.8750
piːp˧	pi̯ɛt˥	pi̯ap˨	**pi̯əʔ˨**	pi̯ɛ˥	pi̯ap˧	**pi̯ɪʔ˥**	pi̯e̝˨˦	0.2500
pei̯˧	pi˨˦	pi˥	pi˦˥	**pi˥˩**	pi˧˩	pi˧˦	pi˦˥	0.1250
piːt˥	**pit˥**	**pit˨**	**pi̯əʔ˨**	**pi˥˩**	**pit˧**	**pi̯ɪʔ˥**	**pi˨˦**	-

reranking

Reranking Result		
rank	$\hat{p}_i^{\mathbf{rk}}$	s_i
0	**pit入**	0.5995
1	pjet入	0.2036
2	pet入	0.0439
3	pep入	-1.2383
4	pij去	-1.4754

Bold: correct protoform or reflexe
m_i: model score (sequence log probability)
r_i: reranker score (reflex prediction accuracy)
s_i: adjusted score (scaled sum of model score and reranker score)

Methods

The Reranked Reconstruction Algorithm

Algorithm 1 Sequential representation of our reranked reconstruction algorithm

Require: $d_1, d_2, ..., d_n$ = reflexes in daughter languages $D_1, D_2, ..., D_n$ from a cognate set with n reflexes
Require: f_{θ_f} = a beam search-enabled reconstruction model with pre-trained parameters θ_f
Require: g_{θ_g} = a reflex prediction model with pre-trained parameters θ_g
Require: k = beam size for predicting candidate reconstructions on f_{θ_f}
Require: α = length normalization constant
Require: λ = score adjustment weight

$D \leftarrow$ "*"D_1":"d_1"*"D_2":"d_2"*" $\cdots$ "*"D_n":"d_n"*" ▷ concatenate reflex sequences into a long sequence, with language labels and separators in between

$C = [(\hat{p}_1, m_1), (\hat{p}_2, m_2), \ldots, (\hat{p}_l, m_l)] \leftarrow f_{\theta_f}(D, k, \alpha)$ ▷ beam search with beam size k to obtain a list of $l \leq k$ candidate protoform predictions $\hat{p}_i$ with their normalized log probabilities $m_i = \frac{\log P(\hat{p}_i|D)}{|\hat{p}_i|^\alpha}$ assigned by f_{θ_f} for $1 \leq i \leq l$

$C' \leftarrow [\,]$ ▷ initialize reranked candidate list
for $(\hat{p}_i, m_i)$ in C **do**
  $a \leftarrow 0$ ▷ counter for the number of correctly derived daughters
  for $j \leftarrow 1$ to n **do**
    $\hat{p}_i' \leftarrow D_j\hat{p}_i$ ▷ prepend the j-th daughter language token to the candidate protoform
    $\hat{d}_{ij} \leftarrow g_{\theta_g}(\hat{p}_i')$ ▷ predict the reflex in the j-th daughter language based on the i-th candidate
    if $\hat{d}_{ij} = d_j$ **then**
      $a \leftarrow a + 1$ ▷ increment counter if predicted reflex is correct
  $r_i \leftarrow a/n$ ▷ use the accuracy of reflex predictions as the reranker score r_i
  $s_i \leftarrow m_i + \lambda r_i$ ▷ calculate the adjusted score s_i for the i-th candidate
  $C' \leftarrow C' +\!\!+ [(\hat{p}_i, s_i)]$ ▷ append entry with adjusted score to reranked candidate list
$C' \leftarrow C'$ sorted by descending s_i
return $C'[0]$ ▷ return the candidate with the highest adjusted score

Datasets

Romance Datasets

- **Rom-phon** (Meloni et al., 2021; Ciobanu and Dinu, 2018) — IPA representation
- **Rom-orth** (Meloni et al., 2021; Ciobanu and Dinu, 2018) — Orthographic representation

Sinitic Datasets

- **WikiHan** (Chang et al., 2022)
- **WikiHan-aug** (Cui et al., 2022) — WikiHan augmented with cognate prediction (Kirov et al., 2022)
- **Hóu** (Hóu, 2004)

Dataset	Cognate sets	# varieties	Ancestor language
WikiHan (Chang et al., 2022)	5,165	8	Middle Chinese
WikiHan-aug (Cui et al., 2022)	8,780	8	Middle Chinese
Hóu (Hóu, 2004)	804	39	Middle Chinese
Rom-phon (Meloni et al., 2021; Ciobanu and Dinu, 2018)	8,703	5	Latin
Rom-orth (Meloni et al., 2021; Ciobanu and Dinu, 2018)	8,631	5	Latin

Models

Reconstruction models (sequence-to-sequence baselines)

- **Meloni et al. (2021)'s GRU** (GRU)
- **Kim et al. (2023)'s Transformer** (Trans)

Reconstruction model with beam search

- **GRU-BS** with support for beam search decoding on the same architecture as Meloni et al. (2021)'s GRU (consisting of language and token embeddings, a single-layer unidirectional encoder-decoder GRU model, and a multi-layer perceptron classifier)

Reflex Prediction models

- **Arora et al. (2023)'s Transformer** reflex prediction model
- **Kim et al. (2023)'s Transformer** reconstruction model adapted for reflex prediction
- **Arora et al. (2023)'s GRU** reflex prediction model
- **Baseline GRU** based on Meloni et al. (2021)'s reconstruction GRU, with multi-layer bidirectional encoding, target language embedding during decode, one-hot vector target language prompting, and target-language-specific connections in the decoder's classifier network

Evaluation Metrics

- **Accuracy (ACC):** The percentage of exactly correct reconstructions
- **Token edit distance (TED):** The number of token insertions, deletions, or substitutions between predictions and targets (Levenshtein et al., 1966)
- **Token error rate (TER):** A length-normalized token edit distance (Cui et al., 2022)
- **Feature error rate (FER):** A length-normalized measure of phonological edit distance by PanPhon (Mortensen et al., 2016)
- **B-Cubed F Score (BCFS):** A measure of structural similarity between predictions and targets (Amigó et al., 2009; List, 2019)

Statistics

- **Wilcoxon Rank-Sum test** (Wilcoxon, 1992) with $\alpha = 0.01$
- **Bootstrap test** (Efron and Tibshirani, 1994) with 99% confidence interval (CI) for difference in mean

Results and Analysis

Results: **Reflex Prediction**

Bold: the best-performing model for each metric

Dataset	Model	ACC%↑	TED↓	TER↓	FER↓	BCFS↑
WikiHan	GRU (baseline)	66.43%	0.5244	0.1547	0.0400	0.7394
	GRU (Arora et al., 2023)	64.45%	0.5558	0.1640	0.0428	0.7260
	Transformer (Kim et al., 2023)	66.39%	0.5302	0.1564	0.0406	0.7370
	Transformer (Arora et al., 2023)	**67.64%**	**0.5128**	**0.1513**	**0.0390**	**0.7445**
WikiHan-aug	GRU (baseline)	68.11%	0.5007	0.1477	0.0380	0.7495
	GRU (Arora et al., 2023)	66.94%	0.5159	0.1522	0.0391	0.7430
	Transformer (Kim et al., 2023)	68.96%	0.4889	0.1442	0.0371	0.7551
	Transformer (Arora et al., 2023)	**69.37%**	**0.4826**	**0.1424**	**0.0363**	**0.7572**
Hóu	GRU (baseline)	51.72%	0.7777	0.2037	0.0488	0.6783
	GRU (Arora et al., 2023)	49.26%	0.8266	0.2166	0.0528	0.6622
	Transformer (Kim et al., 2023)	55.46%	0.7576	0.1985	0.0494	0.6882
	Transformer (Arora et al., 2023)	**55.60%**	**0.7520**	**0.1970**	**0.0485**	**0.6892**
Rom-phon	GRU (baseline)	63.85%	0.7439	0.1014	**0.0426**	0.8361
	GRU (Arora et al., 2023)	48.28%	1.3257	0.1808	0.0930	0.7567
	Transformer (Kim et al., 2023)	**64.19%**	**0.7349**	**0.1002**	0.0427	**0.8380**
	Transformer (Arora et al., 2023)	63.96%	0.7442	0.1015	0.0428	0.8361
Rom-orth	GRU (baseline)	64.58%	0.7301	0.0967	-	0.8465
	GRU (Arora et al., 2023)	57.92%	0.8741	0.1158	-	0.8218
	Transformer (Kim et al., 2023)	64.80%	0.7258	0.0961	-	**0.8478**
	Transformer (Arora et al., 2023)	**65.20%**	**0.7247**	**0.0960**	-	0.8476

Average reflex prediction performance across 20 runs.

Results:
Reflex Prediction

Bold: the best-performing model for each metric

We proceed by choosing the best model for each architecture as a reranker model (highlighted).

Dataset	Model	ACC%↑	TED↓	TER↓	FER↓	BCFS↑
WikiHan	GRU (baseline)	66.43%	0.5244	0.1547	0.0400	0.7394
	GRU (Arora et al., 2023)	64.45%	0.5558	0.1640	0.0428	0.7260
	Transformer (Kim et al., 2023)	66.39%	0.5302	0.1564	0.0406	0.7370
	Transformer (Arora et al., 2023)	**67.64%**	**0.5128**	**0.1513**	**0.0390**	**0.7445**
WikiHan-aug	GRU (baseline)	68.11%	0.5007	0.1477	0.0380	0.7495
	GRU (Arora et al., 2023)	66.94%	0.5159	0.1522	0.0391	0.7430
	Transformer (Kim et al., 2023)	68.96%	0.4889	0.1442	0.0371	0.7551
	Transformer (Arora et al., 2023)	**69.37%**	**0.4826**	**0.1424**	**0.0363**	**0.7572**
Hóu	GRU (baseline)	51.72%	0.7777	0.2037	0.0488	0.6783
	GRU (Arora et al., 2023)	49.26%	0.8266	0.2166	0.0528	0.6622
	Transformer (Kim et al., 2023)	55.46%	0.7576	0.1985	0.0494	0.6882
	Transformer (Arora et al., 2023)	**55.60%**	**0.7520**	**0.1970**	**0.0485**	**0.6892**
Rom-phon	GRU (baseline)	63.85%	0.7439	0.1014	**0.0426**	0.8361
	GRU (Arora et al., 2023)	48.28%	1.3257	0.1808	0.0930	0.7567
	Transformer (Kim et al., 2023)	**64.19%**	**0.7349**	**0.1002**	0.0427	**0.8380**
	Transformer (Arora et al., 2023)	63.96%	0.7442	0.1015	0.0428	0.8361
Rom-orth	GRU (baseline)	64.58%	0.7301	0.0967	-	0.8465
	GRU (Arora et al., 2023)	57.92%	0.8741	0.1158	-	0.8218
	Transformer (Kim et al., 2023)	64.80%	0.7258	0.0961	-	**0.8478**
	Transformer (Arora et al., 2023)	**65.20%**	**0.7247**	**0.0960**	-	0.8476

Average reflex prediction performance across 20 runs.

Results:
Reranked Reconstruction

Bold: the best-performing model for each metric
Asterisk: statistically better performance than both baseline models (Meloni et al. (2021)'s GRU and Kim et al. (2023)'s Transformer)
Dagger: reranking system performs statistically better than its beam search counterpart

Dataset	Reconstruction System	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
WikiHan	GRU (Meloni et al., 2021)	55.58%	0.7360	0.1724	0.0686	0.7426
	Trans (Kim et al., 2023)	54.62%	0.7453	0.1746	0.0696	0.7393
	GRU-BS ($k = 10$)	54.88%	0.7507	0.1758	0.0701	0.7364
	GRU-BS ($k \leq 10$) + GRU Reranker	57.14%*†	0.7045*†	0.1650*†	0.0661*†	0.7515*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**57.26%***†	**0.7029***†	**0.1646***†	**0.0658***†	**0.7520***†
WikiHan-aug	GRU (Meloni et al., 2021)	54.73%	0.7574	0.1774	0.0689	0.7346
	Trans (Kim et al., 2023)	55.82%	0.7317	0.1714	0.0661	0.7416
	GRU-BS ($k = 10$)	56.64%*	0.7214	0.1690	0.0658	0.7454
	GRU-BS ($k \leq 10$) + GRU Reranker	**58.58%***†	**0.6822***†	**0.1598***†	0.0628*†	**0.7579***†
	GRU-BS ($k \leq 10$) + Trans. Reranker	58.58%*†	0.6840*†	0.1602*†	**0.0626***†	0.7575*†
Hóu	GRU (Meloni et al., 2021)	34.63%	1.0916	0.2479	0.0914	0.6697
	Trans (Kim et al., 2023)	39.01%	0.9904	0.2233	0.0875	0.6955
	GRU-BS ($k = 10$)	37.36%	1.0382	0.2328	0.0917	0.6974
	GRU-BS ($k \leq 10$) + GRU Reranker	40.50%†	0.9727†	0.2181†	0.0867†	0.7130*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**42.08%***†	**0.9503***†	**0.2131***†	**0.0850**†	**0.7170***†
Rom-phon	GRU (Meloni et al., 2021)	51.92%	0.9775	0.1244	0.0390	0.8275
	Trans (Kim et al., 2023)	53.04%	0.9050	0.1148	0.0377	0.8417
	GRU-BS ($k = 10$)	52.63%	0.9125	0.1018*	0.0353*	0.8402
	GRU-BS ($k \leq 10$) + GRU Reranker	**53.95%***†	0.8775*†	0.0979*†	0.0336*†	0.8460*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	53.85%*†	**0.8765***†	**0.0978***†	**0.0333***†	**0.8461***†
Rom-orth	GRU (Meloni et al., 2021)	69.41%	0.6004	0.0781	-	0.8906
	Trans (Kim et al., 2023)	71.05%	0.5636	0.0734	-	0.8981
	GRU-BS ($k = 10$)	71.09%	0.5531	0.0617*	-	0.8990
	GRU-BS ($k \leq 10$) + GRU Reranker	**72.60%***†	**0.5237***†	**0.0584***†	-	**0.9045***†
	GRU-BS ($k \leq 10$) + Trans. Reranker	72.50%*†	0.5246*†	0.0585*†	-	0.9044*†

Average reconstruction performance across 20 runs.

Results: **Reranked Reconstruction**

Bold: the best-performing model for each metric
Asterisk: significantly better performance than both baseline models (Meloni et al. (2021)'s GRU and Kim et al. (2023)'s Transformer)
Dagger: reranking system performs significantly better than its beam search counterpart

Our reranking system performs the best on all datasets (highlighted).

Dataset	Reconstruction System	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
WikiHan	GRU (Meloni et al., 2021)	55.58%	0.7360	0.1724	0.0686	0.7426
	Trans (Kim et al., 2023)	54.62%	0.7453	0.1746	0.0696	0.7393
	GRU-BS ($k = 10$)	54.88%	0.7507	0.1758	0.0701	0.7364
	GRU-BS ($k \leq 10$) + GRU Reranker	57.14%*†	0.7045*†	0.1650*†	0.0661*†	0.7515*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**57.26%*†**	**0.7029*†**	**0.1646*†**	**0.0658*†**	**0.7520*†**
WikiHan-aug	GRU (Meloni et al., 2021)	54.73%	0.7574	0.1774	0.0689	0.7346
	Trans (Kim et al., 2023)	55.82%	0.7317	0.1714	0.0661	0.7416
	GRU-BS ($k = 10$)	56.64%*	0.7214	0.1690	0.0658	0.7454
	GRU-BS ($k \leq 10$) + GRU Reranker	**58.58%*†**	**0.6822*†**	**0.1598*†**	0.0628*†	**0.7579*†**
	GRU-BS ($k \leq 10$) + Trans. Reranker	58.58%*†	0.6840*†	0.1602*†	**0.0626*†**	0.7575*†
Hóu	GRU (Meloni et al., 2021)	34.63%	1.0916	0.2479	0.0914	0.6697
	Trans (Kim et al., 2023)	39.01%	0.9904	0.2233	0.0875	0.6955
	GRU-BS ($k = 10$)	37.36%	1.0382	0.2328	0.0917	0.6974
	GRU-BS ($k \leq 10$) + GRU Reranker	40.50%†	0.9727†	0.2181†	0.0867†	0.7130*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**42.08%*†**	**0.9503*†**	**0.2131*†**	**0.0850†**	**0.7170*†**
Rom-phon	GRU (Meloni et al., 2021)	51.92%	0.9775	0.1244	0.0390	0.8275
	Trans (Kim et al., 2023)	53.04%	0.9050	0.1148	0.0377	0.8417
	GRU-BS ($k = 10$)	52.63%	0.9125	0.1018*	0.0353*	0.8402
	GRU-BS ($k \leq 10$) + GRU Reranker	**53.95%*†**	0.8775*†	0.0979*†	0.0336*†	0.8460*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	53.85%*†	**0.8765*†**	**0.0978*†**	**0.0333*†**	**0.8461*†**
Rom-orth	GRU (Meloni et al., 2021)	69.41%	0.6004	0.0781	-	0.8906
	Trans (Kim et al., 2023)	71.05%	0.5636	0.0734	-	0.8981
	GRU-BS ($k = 10$)	71.09%	0.5531	0.0617*	-	0.8990
	GRU-BS ($k \leq 10$) + GRU Reranker	**72.60%*†**	**0.5237*†**	**0.0584*†**	-	**0.9045*†**
	GRU-BS ($k \leq 10$) + Trans. Reranker	72.50%*†	0.5246*†	0.0585*†	-	0.9044*†

Average reconstruction performance across 20 runs.

Results: **Reranked Reconstruction**

Bold: the best-performing model for each metric
Asterisk: significantly better performance than both baseline models (Meloni et al. (2021)'s GRU and Kim et al. (2023)'s Transformer)
Dagger: reranking system performs significantly better than its beam search counterpart

Our reranking system performs the best on all datasets (highlighted).

Dataset	Reconstruction System	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
WikiHan	GRU (Meloni et al., 2021)	55.58%	0.7360	0.1724	0.0686	0.7426
	Trans (Kim et al., 2023)	54.62%	0.7453	0.1746	0.0696	0.7393
	GRU-BS ($k = 10$)	54.88%	0.7507	0.1758	0.0701	0.7364
	GRU-BS ($k \leq 10$) + GRU Reranker	57.14%*†	0.7045*†	0.1650*†	0.0661*†	0.7515*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**57.26%*†**	**0.7029*†**	**0.1646*†**	**0.0658*†**	**0.7520*†**
WikiHan-aug	GRU (Meloni et al., 2021)	54.73%	0.7574	0.1774	0.0689	0.7346
	Trans (Kim et al., 2023)	55.82%	0.7317	0.1714	0.0661	0.7416
	GRU-BS ($k = 10$)	56.64%*	0.7214	0.1690	0.0658	0.7454
	GRU-BS ($k \leq 10$) + GRU Reranker	**58.58%*†**	**0.6822*†**	**0.1598*†**	0.0628*†	**0.7579*†**
	GRU-BS ($k \leq 10$) + Trans. Reranker	58.58%*†	0.6840*†	0.1602*†	**0.0626*†**	0.7575*†
Hóu	GRU (Meloni et al., 2021)	34.63%	1.0916	0.2479	0.0914	0.6697
	Trans (Kim et al., 2023)	39.01%	0.9904	0.2233	0.0875	0.6955
	GRU-BS ($k = 10$)	37.36%	1.0382	0.2328	0.0917	0.6974
	GRU-BS ($k \leq 10$) + GRU Reranker	40.50%†	0.9727†	0.2181†	0.0867†	0.7130*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**42.08%*†**	**0.9503*†**	**0.2131*†**	**0.0850†**	**0.7170*†**
Rom-phon	GRU (Meloni et al., 2021)	51.92%	0.9775	0.1244	0.0390	0.8275
	Trans (Kim et al., 2023)	53.04%	0.9050	0.1148	0.0377	0.8417
	GRU-BS ($k = 10$)	52.63%	0.9125	0.1018*	0.0353*	0.8402
	GRU-BS ($k \leq 10$) + GRU Reranker	**53.95%*†**	0.8775*†	0.0979*†	0.0336*†	0.8460*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	53.85%*†	**0.8765*†**	**0.0978*†**	**0.0333*†**	**0.8461*†**
Rom-orth	GRU (Meloni et al., 2021)	69.41%	0.6004	0.0781	-	0.8906
	Trans (Kim et al., 2023)	71.05%	0.5636	0.0734	-	0.8981
	GRU-BS ($k = 10$)	71.09%	0.5531	0.0617*	-	0.8990
	GRU-BS ($k \leq 10$) + GRU Reranker	**72.60%*†**	**0.5237*†**	**0.0584*†**	-	**0.9045*†**
	GRU-BS ($k \leq 10$) + Trans. Reranker	72.50%*†	0.5246*†	0.0585*†	-	0.9044*†

Average reconstruction performance across 20 runs.

Results: **Reranked Reconstruction**

Bold: the best-performing model for each metric
Asterisk: significantly better performance than both baseline models (Meloni et al. (2021)'s GRU and Kim et al. (2023)'s Transformer)
Dagger: reranking system performs significantly better than its beam search counterpart

GRU-BS with reranking performs significantly better on 4 out of the 5 datasets (highlighted).

Dataset	Reconstruction System	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
WikiHan	GRU (Meloni et al., 2021)	55.58%	0.7360	0.1724	0.0686	0.7426
	Trans (Kim et al., 2023)	54.62%	0.7453	0.1746	0.0696	0.7393
	GRU-BS ($k = 10$)	54.88%	0.7507	0.1758	0.0701	0.7364
	GRU-BS ($k \leq 10$) + GRU Reranker	57.14%*†	0.7045*†	0.1650*†	0.0661*†	0.7515*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**57.26%***†	**0.7029***†	**0.1646***†	**0.0658***†	**0.7520***†
WikiHan-aug	GRU (Meloni et al., 2021)	54.73%	0.7574	0.1774	0.0689	0.7346
	Trans (Kim et al., 2023)	55.82%	0.7317	0.1714	0.0661	0.7416
	GRU-BS ($k = 10$)	56.64%*	0.7214	0.1690	0.0658	0.7454
	GRU-BS ($k \leq 10$) + GRU Reranker	**58.58%***†	**0.6822***†	**0.1598***†	0.0628*†	**0.7579***†
	GRU-BS ($k \leq 10$) + Trans. Reranker	58.58%*†	0.6840*†	0.1602*†	**0.0626***†	0.7575*†
Hóu	GRU (Meloni et al., 2021)	34.63%	1.0916	0.2479	0.0914	0.6697
	Trans (Kim et al., 2023)	39.01%	0.9904	0.2233	0.0875	0.6955
	GRU-BS ($k = 10$)	37.36%	1.0382	0.2328	0.0917	0.6974
	GRU-BS ($k \leq 10$) + GRU Reranker	40.50%†	0.9727†	0.2181†	0.0867†	0.7130*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**42.08%***†	**0.9503***†	**0.2131***†	**0.0850**†	**0.7170***†
Rom-phon	GRU (Meloni et al., 2021)	51.92%	0.9775	0.1244	0.0390	0.8275
	Trans (Kim et al., 2023)	53.04%	0.9050	0.1148	0.0377	0.8417
	GRU-BS ($k = 10$)	52.63%	0.9125	0.1018*	0.0353*	0.8402
	GRU-BS ($k \leq 10$) + GRU Reranker	**53.95%***†	0.8775*†	0.0979*†	0.0336*†	0.8460*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	53.85%*†	**0.8765***†	**0.0978***†	**0.0333***†	**0.8461***†
Rom-orth	GRU (Meloni et al., 2021)	69.41%	0.6004	0.0781	-	0.8906
	Trans (Kim et al., 2023)	71.05%	0.5636	0.0734	-	0.8981
	GRU-BS ($k = 10$)	71.09%	0.5531	0.0617*	-	0.8990
	GRU-BS ($k \leq 10$) + GRU Reranker	**72.60%***†	**0.5237***†	**0.0584***†	-	**0.9045***†
	GRU-BS ($k \leq 10$) + Trans. Reranker	72.50%*†	0.5246*†	0.0585*†	-	0.9044*†

Average reconstruction performance across 20 runs.

Results: **Reranked Reconstruction**

Bold: the best-performing model for each metric
Asterisk: significantly better performance than both baseline models (Meloni et al. (2021)'s GRU and Kim et al. (2023)'s Transformer)
Dagger: reranking system performs significantly better than its beam search counterpart

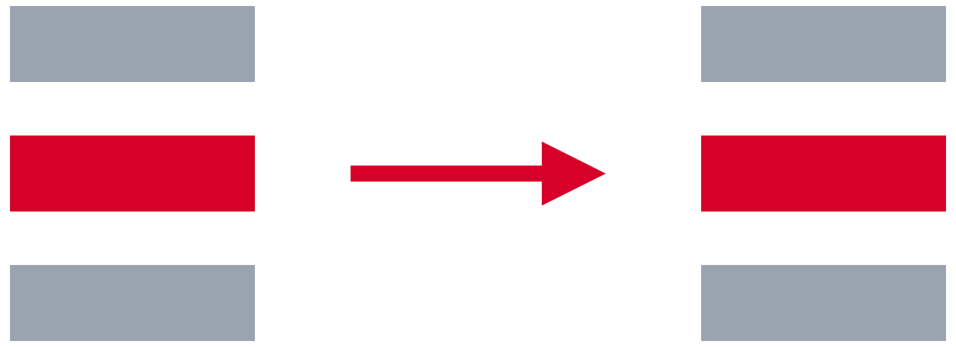
Ablation:
GRU-BS with reranking performs significantly better than GRU-BS without reranking on all 5 datasets (highlighted).

Dataset	Reconstruction System	ACC%↑	TED↓	TER↓	FER↓	BCFS↑
WikiHan	GRU (Meloni et al., 2021)	55.58%	0.7360	0.1724	0.0686	0.7426
	Trans (Kim et al., 2023)	54.62%	0.7453	0.1746	0.0696	0.7393
	GRU-BS ($k = 10$)	54.88%	0.7507	0.1758	0.0701	0.7364
	GRU-BS ($k \leq 10$) + GRU Reranker	57.14%*†	0.7045*†	0.1650*†	0.0661*†	0.7515*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**57.26%***†	**0.7029***†	**0.1646***†	**0.0658***†	**0.7520***†
WikiHan-aug	GRU (Meloni et al., 2021)	54.73%	0.7574	0.1774	0.0689	0.7346
	Trans (Kim et al., 2023)	55.82%	0.7317	0.1714	0.0661	0.7416
	GRU-BS ($k = 10$)	56.64%*	0.7214	0.1690	0.0658	0.7454
	GRU-BS ($k \leq 10$) + GRU Reranker	**58.58%***†	**0.6822***†	**0.1598***†	0.0628*†	**0.7579***†
	GRU-BS ($k \leq 10$) + Trans. Reranker	58.58%*†	0.6840*†	0.1602*†	**0.0626***†	0.7575*†
Hóu	GRU (Meloni et al., 2021)	34.63%	1.0916	0.2479	0.0914	0.6697
	Trans (Kim et al., 2023)	39.01%	0.9904	0.2233	0.0875	0.6955
	GRU-BS ($k = 10$)	37.36%	1.0382	0.2328	0.0917	0.6974
	GRU-BS ($k \leq 10$) + GRU Reranker	40.50%†	0.9727†	0.2181†	0.0867†	0.7130*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	**42.08%***†	**0.9503***†	**0.2131***†	**0.0850**†	**0.7170***†
Rom-phon	GRU (Meloni et al., 2021)	51.92%	0.9775	0.1244	0.0390	0.8275
	Trans (Kim et al., 2023)	53.04%	0.9050	0.1148	0.0377	0.8417
	GRU-BS ($k = 10$)	52.63%	0.9125	0.1018*	0.0353*	0.8402
	GRU-BS ($k \leq 10$) + GRU Reranker	**53.95%***†	0.8775*†	0.0979*†	0.0336*†	0.8460*†
	GRU-BS ($k \leq 10$) + Trans. Reranker	53.85%*†	**0.8765***†	**0.0978***†	**0.0333***†	**0.8461***†
Rom-orth	GRU (Meloni et al., 2021)	69.41%	0.6004	0.0781	-	0.8906
	Trans (Kim et al., 2023)	71.05%	0.5636	0.0734	-	0.8981
	GRU-BS ($k = 10$)	71.09%	0.5531	0.0617*	-	0.8990
	GRU-BS ($k \leq 10$) + GRU Reranker	**72.60%***†	**0.5237***†	**0.0584***†	-	**0.9045***†
	GRU-BS ($k \leq 10$) + Trans. Reranker	72.50%*†	0.5246*†	0.0585*†	-	0.9044*†

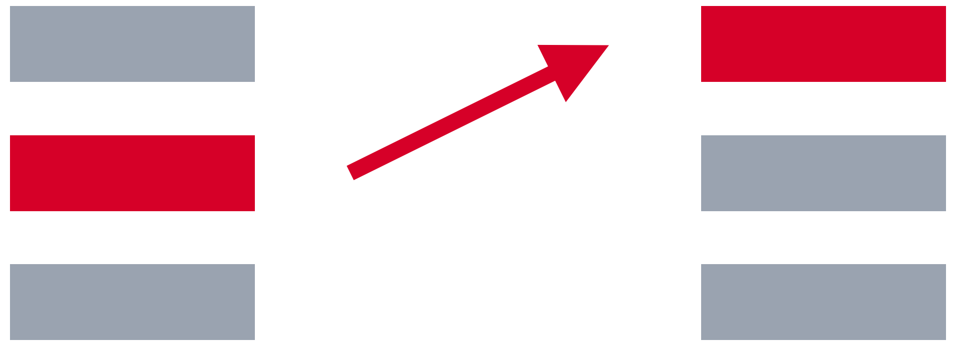
Average reconstruction performance across 20 runs.

Reranker Behavior Categories

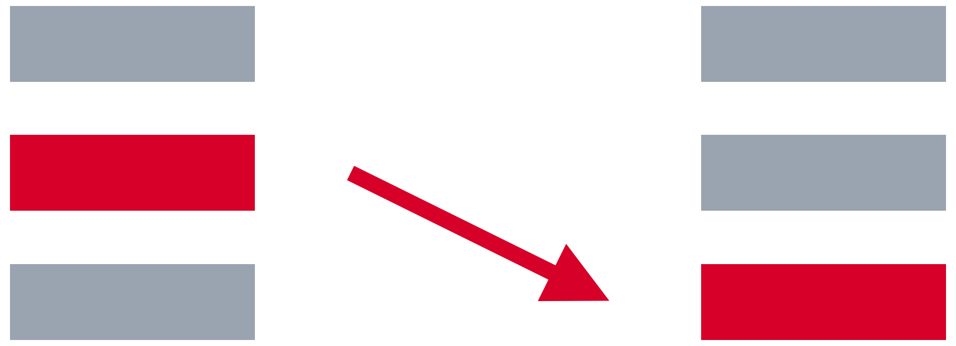
Not-in: the target protoform is not part of the beam search result



Unchanged: reranker does not change ranking of the target protoform



Improved: reranker assigns better ranking to the target protoform



Worsened: reranker assigns worse ranking to the target protoform

correct protoform

incorrect protoform

Distribution of Reranker Behavior

Dataset	Improved	Worsened	Unchanged	Not-in	Total
WikiHan	84 (8.13%)	32 (3.10%)	755 (73.09%)	162 (15.68%)	1033
WikiHan-aug	88 (8.52%)	23 (2.23%)	791 (76.57%)	131 (12.68%)	1033
Hóu	26 (16.15%)	15 (9.32%)	88 (54.66%)	32 (19.88%)	161
Rom-phon	109 (6.21%)	61 (3.48%)	1198 (68.30%)	386 (22.01%)	1754
Rom-orth	75 (4.29%)	23 (1.32%)	1367 (78.16%)	284 (16.24%)	1749

Improved/Changed (%)
72.41%
79.28%
63.41%
64.12%
76.53%

The distribution of reranker behavior categorization on the test set (left), and the corresponding rate of ranking improvement among instances with changed (i.e. improved or worsened) ranking (right).

Distribution of Reranker Behavior

Dataset	Improved	Worsened	Unchanged	Not-in	Total
WikiHan	84 (8.13%)	32 (3.10%)	755 (73.09%)	162 (15.68%)	1033
WikiHan-aug	88 (8.52%)	23 (2.23%)	791 (76.57%)	131 (12.68%)	1033
Hóu	26 (16.15%)	15 (9.32%)	88 (54.66%)	32 (19.88%)	161
Rom-phon	109 (6.21%)	61 (3.48%)	1198 (68.30%)	386 (22.01%)	1754
Rom-orth	75 (4.29%)	23 (1.32%)	1367 (78.16%)	284 (16.24%)	1749

Improved/Changed (%)
72.41%
79.28%
63.41%
64.12%
76.53%

The distribution of reranker behavior categorization on the test set (left), and the corresponding rate of ranking improvement among instances with changed (i.e. improved or worsened) ranking (right).

Distribution of Reranker Behavior

Dataset	Improved	Worsened	Unchanged	Not-in	Total
WikiHan	84 (8.13%)	32 (3.10%)	755 (73.09%)	162 (15.68%)	1033
WikiHan-aug	88 (8.52%)	23 (2.23%)	791 (76.57%)	131 (12.68%)	1033
Hóu	26 (16.15%)	15 (9.32%)	88 (54.66%)	32 (19.88%)	161
Rom-phon	109 (6.21%)	61 (3.48%)	1198 (68.30%)	386 (22.01%)	1754
Rom-orth	75 (4.29%)	23 (1.32%)	1367 (78.16%)	284 (16.24%)	1749

Improved/Changed (%)
72.41%
79.28%
63.41%
64.12%
76.53%

The distribution of reranker behavior categorization on the test set (left), and the corresponding rate of ranking improvement among instances with changed (i.e. improved or worsened) ranking (right).

Errors and Phonetic Distances

D_T: normalized token edit distance

D_F: normalized feature edit distance

Dataset	Category	$D_T(\hat{p}, R) < D_T(p, R)$ (R more similar to $\hat{p}$ than p by D_T)	$D_F(\hat{p}, R) < D_F(p, R)$ (R more similar to $\hat{p}$ than p by D_F)
WikiHan	Worsened	**37.50%**	**53.12%**
	Unchanged	35.56%	43.33%
	Improved	28.30%	32.08%
Rom-phon	Worsened	**60.66%**	**62.30%**
	Unchanged	47.21%	51.48%
	Improved	47.17%	49.06%

R: reflexes **p̂: predicted protoform** **p: target protoform**

Comparison between the phonetic similarity between the reflexes R and the predicted protoform p̂ versus the target protoform p for each category of the reranker's behavior among reconstruction errors

Errors and Phonetic Distances

D_T: normalized token edit distance

D_F: normalized feature edit distance

Dataset	Category	$D_T(\hat{p}, R) < D_T(p, R)$ (R more similar to $\hat{p}$ than p by D_T)	$D_F(\hat{p}, R) < D_F(p, R)$ (R more similar to $\hat{p}$ than p by D_F)
WikiHan	Worsened	**37.50%**	**53.12%**
	Unchanged	35.56%	43.33%
	Improved	28.30%	32.08%
Rom-phon	Worsened	**60.66%**	**62.30%**
	Unchanged	47.21%	51.48%
	Improved	47.17%	49.06%

R: reflexes **p̂: predicted protoform** **p: target protoform**

Comparison between the phonetic similarity between the reflexes R and the predicted protoform p̂ versus the target protoform p for each category of the reranker's behavior among reconstruction errors

Error Instances

Dataset	Category	Worsened Proto	Worsened Prôto	Unchanged Proto	Unchanged Prôto	Improved Proto	Improved Prôto
WikiHan	Middle Chinese	**mjukʷ**	**mukʷ**	**tʰ ja ŋ**	**t͡ɕʰa ŋ**	**t͡sʰje k**	**t͡s e k**
	Cantonese	m␣ʊk	mʊk	t͡sʰ␣ɔːŋ	t͡sʰɔːŋ	t͡s ␣ɪ k	t͡s ɪ k
	Hakka	m␣uk	muk			t͡s ␣a k	t͡s a k
	Mandarin	m␣u␣	mu␣	t͡ʂʰ␣a ŋ	t͡ʂʰa ŋ	t͡ɕʰ␣i ␣	t͡ɕʰi ␣
	Hokkien	b␣ɔk	bɔk	t͡ɕʰ␣i̯ɔŋ	t͡ɕʰi̯ɔŋ	t͡ɕʰ␣i̯ək	t͡ɕʰi̯ək
Rom-phon	Latin	**astʰma**	**as␣ma**	**f␣ɛrɪtatɛm**	**f␣ɛritam**	**tɛksɛrɛ**	**tɪsserɛ**
	Romanian	ast mə	astmə			t␣␣sese	t␣␣sese
	French	as␣ m␣	as␣m␣	fjɛʁ␣te␣␣␣	fjɛʁ␣te␣	ti␣se␣␣	ti␣se␣␣
	Italian	az␣ ma	az␣ma	f␣erita␣␣␣	f␣erita␣	tɛssere	tɛssere
	Spanish	as␣ ma	as␣ma			tex␣eɾ␣	tex␣eɾ␣
	Portuguese	aʒ␣ mɐ	aʒ␣mɐ			tɨ␣seɪ␣	tɨ␣seɪ␣

Color key: substitution insertion (␣) deletion

Error Instances

睦 *mjuk^{w}* 'friendly'

Proto	Prôto
mjuk^{w}	muk^{w}
m␣ʊk	mʊk
m␣uk	muk
m␣u␣	mu␣
b␣ɔk	bɔk

Proto: target protoform
Prôto: predicted protoform

Error Instances

昶 *tʰjaŋ* 'long daytime'

Proto	P̂roto
tʰ ja ŋ	t͡ɕʰa ŋ
t͡sʰ⎵ɔːŋ	t͡sʰɔːŋ
t͡ʂʰ⎵a ŋ	t͡ʂʰa ŋ
t͡ɕʰ⎵i̯ɔŋ	t͡ɕʰi̯ɔŋ

磧 *t͡sʰjek* 'gravel'

Proto	P̂roto
t͡sʰje k	t͡s e k
t͡s ⎵ɪ k	t͡s ɪ k
t͡s ⎵a k	t͡s a k
t͡ɕʰ⎵i ⎵	t͡ɕʰi ⎵
t͡ɕʰ⎵i̯ək	t͡ɕʰi̯ək

Error Instances

asthma **'asthma'**

Proto	Prôto
ast^{h}ma	**as␣ma**
ast mə	astmə
as␣ m␣	as␣m␣
az␣ ma	az␣ma
as␣ ma	as␣ma
aʒ␣ mɐ	aʒ␣mɐ

feritatem **'ferocity'**

Proto	Prôto
f␣ɛrɪtatɛm	**f␣ɛritam**
fjɛʁ␣te␣␣␣	fjɛʁ␣te␣
f␣erita␣␣␣	f␣erita␣

Distribution of Reranker Behavior

Dataset	Improved	Worsened	Unchanged	Not-in	Total
WikiHan	84 (8.13%)	32 (3.10%)	755 (73.09%)	162 (15.68%)	1033
WikiHan-aug	88 (8.52%)	23 (2.23%)	791 (76.57%)	131 (12.68%)	1033
Hóu	26 (16.15%)	15 (9.32%)	88 (54.66%)	32 (19.88%)	161
Rom-phon	109 (6.21%)	61 (3.48%)	1198 (68.30%)	386 (22.01%)	1754
Rom-orth	75 (4.29%)	23 (1.32%)	1367 (78.16%)	284 (16.24%)	1749

Improved/Changed (%)
72.41%
79.28%
63.41%
64.12%
76.53%

Conclusion

Conclusion

Our reranked reconstruction system provides an elegant to replicate the synergy between reconstruction and reflex prediction in the comparative method. Our results serve as a vindication of the idea that designing reconstruction systems with the comparative method in mind can be more powerful than relying solely on sequence-to-sequence techniques

- **Keep linguists' methods in mind** in computational linguistics!
- **Synergizing related tasks** (reflex prediction, reconstruction, cognate prediction) in historical linguistics can lead to better results
- **Using reflex prediction** in neural reconstruction is possibly a new framework for future reconstruction research

Links

Links

Paper: https://arxiv.org/abs/2403.18769 (or conference site)

Code: https://github.com/cmu-llab/reranked-reconstruction

References

References

Bibliographical References

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2023. Cognate transformer for automated phonological reconstruction and cognate reflex prediction. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 6852–6862, Singapore. Association for Computational Linguistics.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information retrieval, 12:461–486.

Raimo Anttila. 1989. Historical and Comparative Linguistics. John Benjamins Publishing.

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2023. Jambu: A historical linguistic database for South Asian languages. In Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 68–77, Toronto, Canada. Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases.

Timotheus A. Bodt and Johann-Mattis List. 2022.

Reflex prediction: A case study of Western KhoBwa. Diachronica, 39(1):1–38.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. Proceedings of the National Academy of Sciences, 110(11):4224–4229.

L. Campbell. 2021. Historical Linguistics: An Introduction. Edinburgh University Press.

Chundra Cathcart and Taraka Rama. 2020. Disentangling dialects: A neural approach to IndoAryan historical phonology and subgrouping. In Proceedings of the 24th Conference on Computational Natural Language Learning, pages 620– 630, Online. Association for Computational Linguistics.

Kalvin Chang, Chenxuan Cui, Youngmin Kim, and David R. Mortensen. 2022. WikiHan: A New Comparative Dataset for Chinese Languages. In Proceedings of the 29th International Conference on Computational Linguistics, pages 3563–3569, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alina Maria Ciobanu and Liviu P. Dinu. 2018. Ab Initio: Automatic Latin Proto-word Reconstruction. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1604–1614, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alina Maria Ciobanu, Liviu P. Dinu, and Laurentiu Zoicas. 2020. Automatic Reconstruction of Missing Romanian Cognates and Unattested Latin Words. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 3226–3231, Marseille, France. European Language Resources Association.

Chenxuan Cui, Ying Chen, Qinxin Wang, and David R Mortensen. 2022. Neural ProtoLanguage Reconstruction. Technical report, Carnegie Mellon University, Pittsburgh, PA.

Stanton P. Durham and David Ellis Rogers. 1969. An Application of Computer Programming to the Reconstruction of a Proto-Language. In International Conference on Computational Linguistics COLING 1969: Preprint No. 5, Sånga Säby, Sweden.

Bradley Efron and Robert J Tibshirani. 1994. An introduction to the bootstrap. CRC press.

Clémentine Fourrier. 2022. Neural Approaches to Historical Word Reconstruction. Ph.D. thesis, Université PSL (Paris Sciences & Lettres).

Andre He, Nicholas Tomlin, and Dan Klein. 2023.

Neural Unsupervised Reconstruction of Protolanguage Word Forms. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1636–1649, Toronto, Canada. Association for Computational Linguistics.

Bernhard Karlgren. 1974. Analytic Dictionary of Chinese and Sino-Japanese. Courier Corporation.

Young Min Kim, Kalvin Chang, Chenxuan Cui, and David R. Mortensen. 2023. Transformed Protoform Reconstruction. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 24–38, Toronto, Canada. Association for Computational Linguistics.

Christo Kirov, Richard Sproat, and Alexander Gutkin. 2022. Mockingbird at the SIGTYP 2022 Shared Task: Two Types of Models for the Prediction of Cognate Reflexes. In Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 70–79, Seattle, Washington. Association for Computational Linguistics.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, volume 10, pages 707–710. Soviet Union.

Johann-Mattis List. 2019. Beyond edit distances: Comparing linguistic reconstruction systems. Theoretical Linguistics, 45(3-4):247–258.

Johann-Mattis List, Robert Forkel, and Nathan Hill. 2022a. A New Framework for Fast Automated Phonological Reconstruction Using Trimmed Alignments and Sound Correspondence Patterns. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 89–96, Dublin, Ireland. Association for Computational Linguistics.

Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan Hill, and Ryan Cotterell. 2022b. The SIGTYP 2022 Shared Task on the Prediction of Cognate Reflexes. In Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 52–62, Seattle, Washington. Association for Computational Linguistics.

Clayton Marr and David Mortensen. 2023. Largescale computerized forward reconstruction yields new perspectives in French diachronic phonology. Diachronica, 40(2):238–285.

Clayton Marr and David R. Mortensen. 2020. Computerized Forward Reconstruction for Analysis in Diachronic Phonology, and Latin to French Reflex Prediction. In Proceedings of LT4HALA 2020 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 28–36, Marseille, France. European Language Resources Association (ELRA).

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab Antiquo: Neural Proto-language Reconstruction. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4460– 4473, Online. Association for Computational Linguistics.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Remo Nitschke. 2021. Restoring the Sister: Reconstructing a Lexicon from Sister Languages using Neural Machine Translation. In Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 122–130, Online. Association for Computational Linguistics.

Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. 2021. MSA Transformer.

Szemerényi, O. J. L. (1996). Introduction to Indo-European Linguistics. Oxford University Press UK.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In Breakthroughs in Statistics: Methodology and Distribution, pages 196–202. Springer.

Language Resource References

Baxter, William H. 2014. Baxter-Sagart Old Chinese Reconstruction, Version 1.1 (20 September 2014).

Chang, Kalvin and Cui, Chenxuan and Kim, Youngmin and Mortensen, David R. 2022. WikiHan: A New Comparative Dataset for Chinese Languages. International Committee on Computational Linguistics.

Ciobanu, Alina Maria and Dinu, Liviu P. 2018. Ab Initio: Automatic Latin Proto-word Reconstruction. Association for Computational Linguistics.

Cui, Chenxuan and Chen, Ying and Wang, Qinxin and Mortensen, David R. 2022. Neural ProtoLanguage Reconstruction.

Hóu, Jīngyī. 2004. Xiàndài Hànyǔ Fāngyán Yīnkù 现代汉语方言音库 [Phonological Database of Chinese Dialects].

Meloni, Carlo and Ravfogel, Shauli and Goldberg, Yoav. 2021. Ab Antiquo: Neural Proto-language Reconstruction. Association for Computational Linguistics.