Carnegie Mellon University
Language
Technologies
Institute

LLab
Language Technologies Institute
Carnegie Mellon University

Semisupervised Neural Proto-Language Reconstruction

Liang Lu[1], Peirong Xie[2], David R. Mortensen[1]

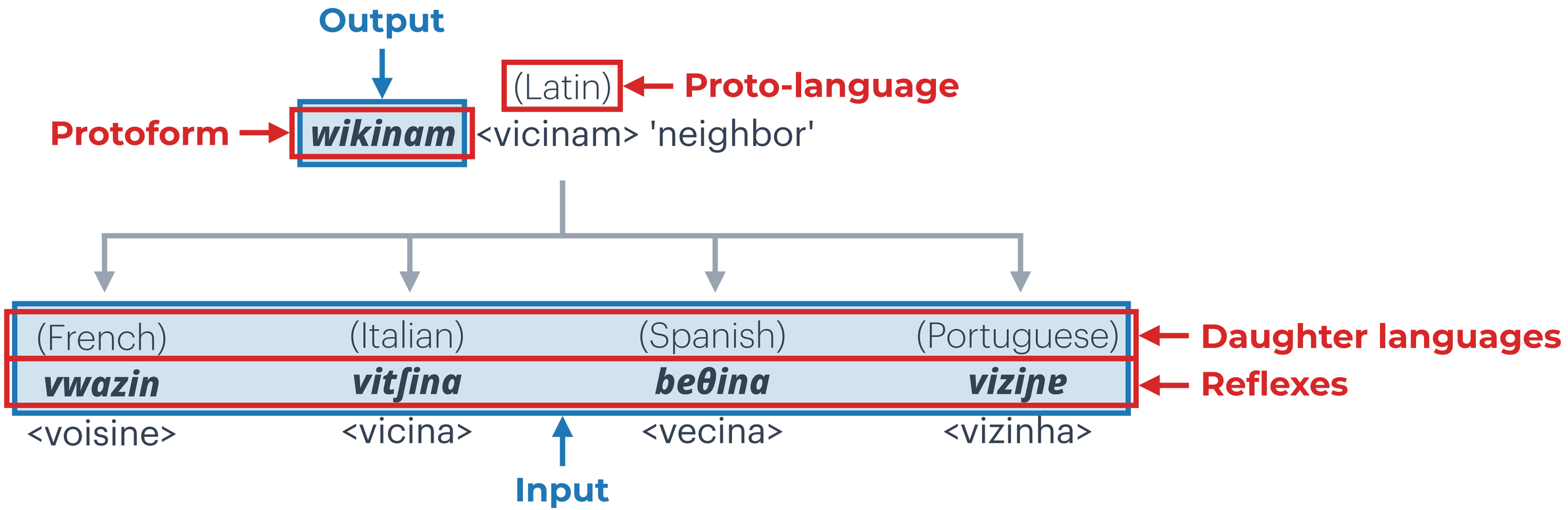
[1]Carnegie Mellon University, [2]University of Southern California

lianglu@cs.cmu.edu, louisxie@usc.edu, dmortens@cs.cmu.edu

ACL 2024

Bangkok, Thailand

Protoform Reconstruction



Shown: an example from the Romance dataset (Meloni et al., 2021; Ciobanu and Dinu, 2018)

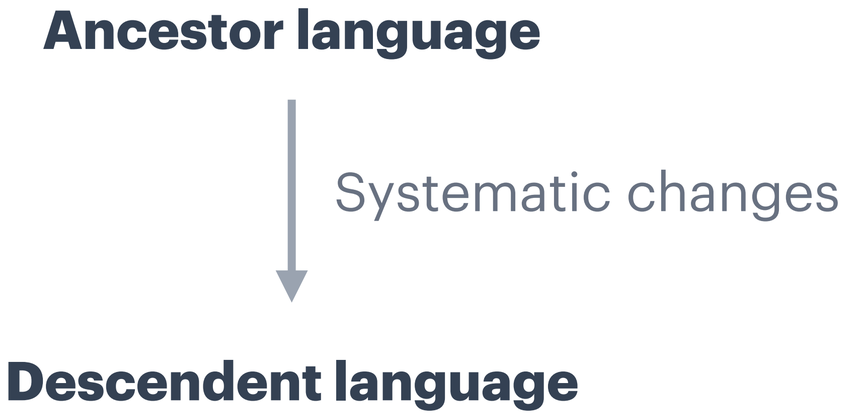
A 19th-Century Discovery

Languages change in systematic ways, and **it is possible to reproducibly reconstruct proto-languages** using these systematic patterns, even when no record of the proto-language survives.

Historical linguists use the **comparative method** to reconstruct proto-languages.

The comparative method's **regularity principle**:

- Sound changes are regular
- Reflexes should be derivable deterministically from reconstructions using a single set of sound change rules

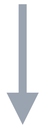


Supervised Neural Reconstruction

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

Input sequence (concatenated reflexes)

[Cantonese]:meiˌɬ[Mandarin]:meiˌɴ*[Wu]:mɐˌɬ*



Supervised sequence-to-sequence model

Target prediction (Middle Chinese)

mij³

Shown: the 煝 cognate set from the Chinese dataset (Chang et al., 2022)

A Hypothetical Example: Tangkhulic Languages

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r u k	r u k	n u k	n u k
Ukhrul	r u	r u	n u	n u
Protoform Label	d u	r u	n u	n ɨ

A Hypothetical Example: Tangkhulic Languages

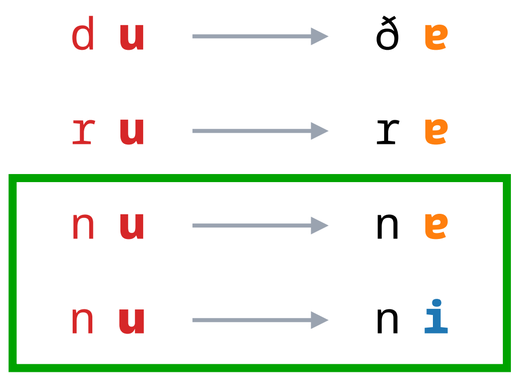
Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r u k	r u k	n u k	n u k
Ukhrul	r u	r u	n u	n u
Protoform Label	d u	r u	(hidden)	(hidden)

A Hypothetical Example: Tangkhulic Languages

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r **u** k	r **u** k	n **u** k	n **u** k
Ukhrul	r **u**	r **u**	n **u**	n **u**
Protoform Label	d **u**	r **u**	(hidden)	(hidden)
Supervised Only	d **u**	r **u**	n **u**	n **u**

A Hypothetical Example: Tangkhulic Languages

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r u k	r u k	n u k	n u k
Ukhrul	r u	r u	n u	n u
Protoform Label	d u	r u	(hidden)	(hidden)
Supervised Only	d u	r u	n u	n u



Trouble: Cannot deterministically derive the reflexes!

A Hypothetical Example: Tangkhulic Languages

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r u k	r u k	n u k	n u k
Ukhrul	r u	r u	n u	n u
Protoform Label	d u	r u	(hidden)	(hidden)
Supervised Only	d u	r u	n u	n ~~u~~
Our Model	d u	r u	n u	n *

Something other than **u**

d u → ð ɐ
r u → r ɐ
n u → n ɐ
n ~~u~~ → n i

d u → ð ɐ
r u → r ɐ
n u → n ɐ
n * → n i

A Hypothetical Example: Tangkhulic Languages

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r **u** k	r **u** k	n **u** k	n **u** k
Ukhrul	r **u**	r **u**	n **u**	n **u**
Protoform Label	d **u**	r **u**	n **u**	n **ɨ**
Supervised Only	d **u**	r **u**	n **u**	n ~~**u**~~
Our Model	d **u**	r **u**	n **u**	n *

← Indeed not **u**

↑ Something other than **u**

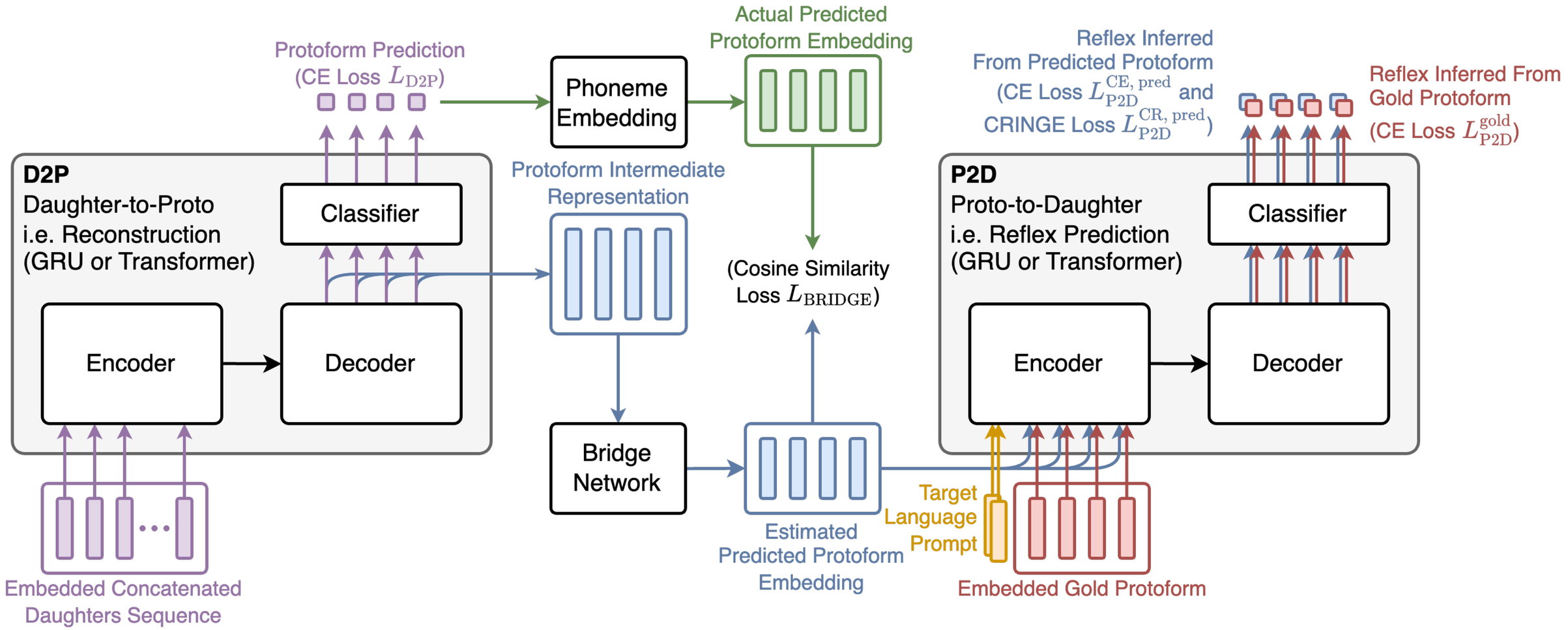
A Hypothetical Example: Tangkhulic Languages

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r u k	r u k	n u k	n u k
Ukhrul	r u	r u	n u	n u
Protoform Label	d u	r u	n u	n ɨ
Our Model: DPD	d u	r u	n u	n *

Reflexes → **Reconstruction** → Protoform? → **Reflex Prediction** → Reflexes?

Daughter-to-Proto-to-Daughter (DPD)

The DPD (Dughter-to-Proto-to-Daughter) Architecture



Training Strategies

1. Supervised only (**SUPV**)
2. Bootstrapping (**BST**)
3. Π-model (**ΠM**)
4. Π-model with Bootstrapping (**ΠM-BST**)
5. **DPD**
6. DPD with Bootstrapping (**DPD-BST**)
7. DPD merged with Π-model (**DPD-ΠM**)
8. DPD-ΠM with Bootstrapping (**DPD-ΠM-BST**)

Naïve baseline

Semisupervised techniques

Combination of semisupervised techniques

DPD-based strategies

Semisupervised Datasets

We randomly remove labels from existing fully-labeled datasets to create semisupervised train sets.

	Chinese (Sinitic)	Romance
5%	181	304
10%	362	607
20%	723	1,214
30%	1,084	1,821
100%	3,615	6,071

Number of labeled training examples (i.e. cognate sets with an associated gold protoform) in the train set for each labeling setting and dataset, as well as the total number of cognate sets for reference (100%).

Results

For all data settings, sub-model architectures, and language families, **the winning model** (by best average accuracy) **is always a DPD model**.

The **DPD-ΠM-BST** model that combines **DPD**, **Π-Model**, and **Bootstrapping** is usually the best.

Language Family	Sub-Model Architecture	5% Labeled	10% Labeled	20% Labeled	30% Labeled
Chinese	Transformer	DPD	DPD-ΠM-BST	DPD-ΠM-BST	DPD-ΠM
	GRU	DPD-ΠM-BST	DPD-ΠM-BST	DPD-ΠM-BST	DPD-ΠM
Romance	Transformer	DPD-ΠM-BST	DPD-ΠM-BST	DPD-ΠM-BST	DPD-ΠM-BST
	GRU	DPD-BST	DPD-BST	DPD-ΠM-BST	DPD-BST

Conclusion

DPD marks a step forward toward **building practical computational reconstruction systems** that can assist early-stage proto-language reconstruction projects.

Even in the age of Transformers, **linguistic ideas that predate NLP**, such as the comparative method, can still **provide valuable insights for approaching NLP challenges**.