Digging for old loanwords

Etymologists need tools

- Discovering new etymologies is an ongoing task
  - Subtask: Searching for old contact layers
- Problem: Intransparent & slow workflow
- Solution: standardise data, formalise steps
- Goal:
  - Make the field more accessible,
  - speed up etymologisations

Find loans between 2 word lists

The task:

- Pick two unrelated languages
- Take list of word-meaning pairs of both
- Compare the two lists and flag potential loans

Concept lists and phonetic similarity

Common solution:

- Map meanings to concepts
- Match forms of same concepts
- Calculate phonetic similarity of forms
- Most similar ones are the best candidates

Limitations of the common solution

Concept lists:

- don't capture semantic change
- are only available for core vocabulary, while loans appear in the periphery

Similarity measures:

- ignore knowledge about sound correspondences

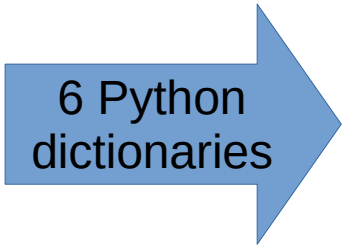
Solution: word vectors & etymological dictionaries

Word vectors: Input: word1, word2, Output: semantic similarity score

Etymological dictionaries: Align entries, extract horizontal and vertical sound correspondences, apply to new examples

Extracting etymological information

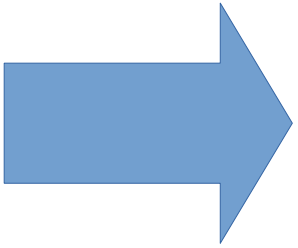
ID	Lg 1	Lg 2
1	hehe	kiki
2	buba	pupo




```
[
{'h': ['k'], 'e': ['i'], 'b': [p], 'u': ['u'], 'a':
['o']},
{'h<*k': 2, 'e<*i': 2, 'b<*p': 2, 'u<*u':
1, 'a<*o': 1},
{'h<*k': [1], 'e<*i': [1], 'b<*p': [2],
'u<*u': [2], 'a<*o': [2]},
{'CVCV': ['CVCV']},
{'CVCV<*CVCV': 2},
{'CVCV<*CVCV': [1, 2]}
]
```

Apply etymological information

```
[
{'h': ['k'], 'e': ['i'], 'b': [p], 'u': ['u'], 'a':
['o']},
{'h<*k': 2, 'e<*i': 2, 'b<*p': 2, 'u<*u':
1, 'a<*o': 1},
{'h<*k': [1], 'e<*i': [1], 'b<*p': [2],
'u<*u': [2], 'a<*o': [2]},
{'CVCV': ['CVCV']},
{'CVCV<*CVCV': 2},
{'CVCV<*CVCV': [1, 2]}
]
```



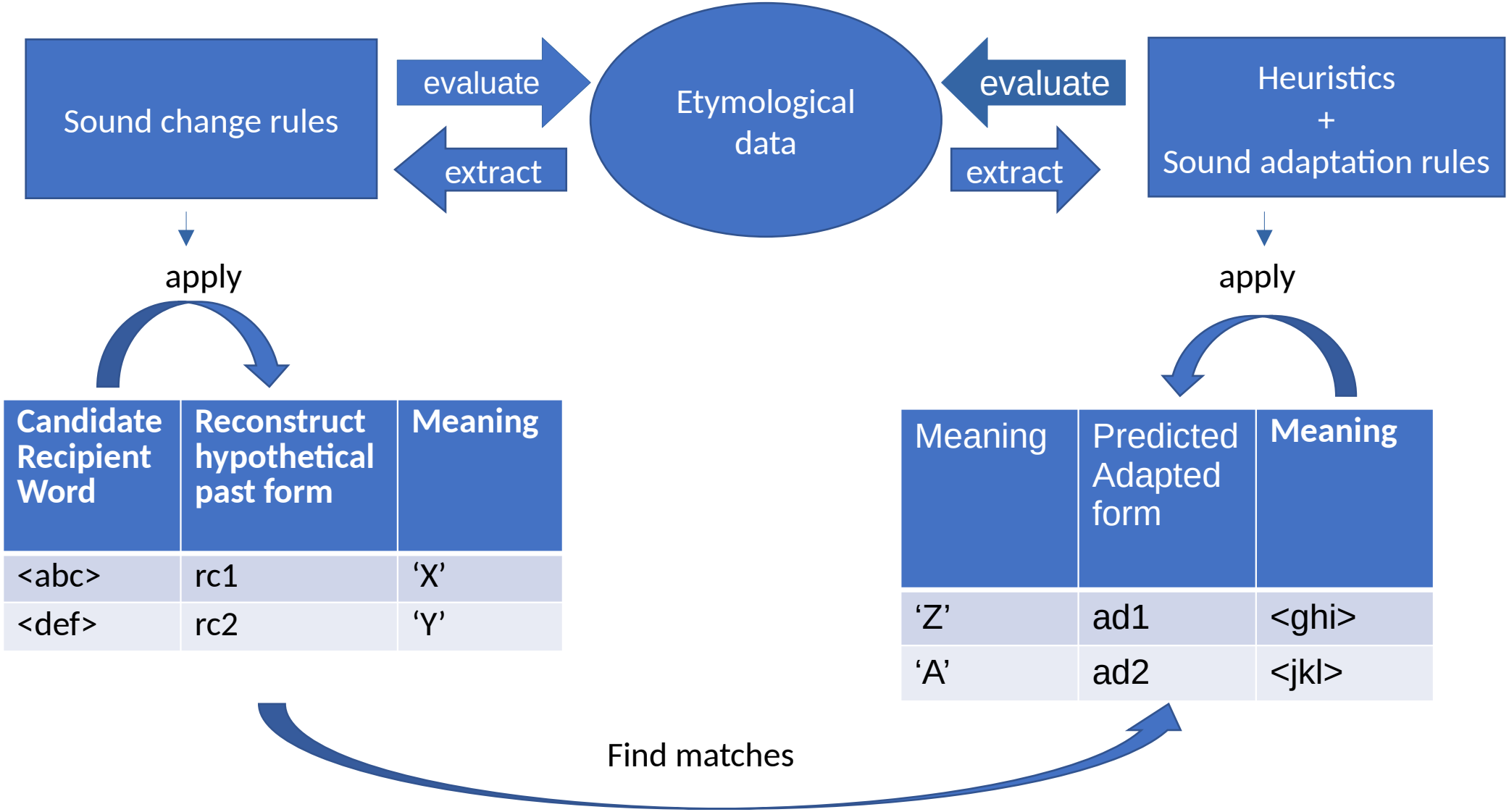
Example:

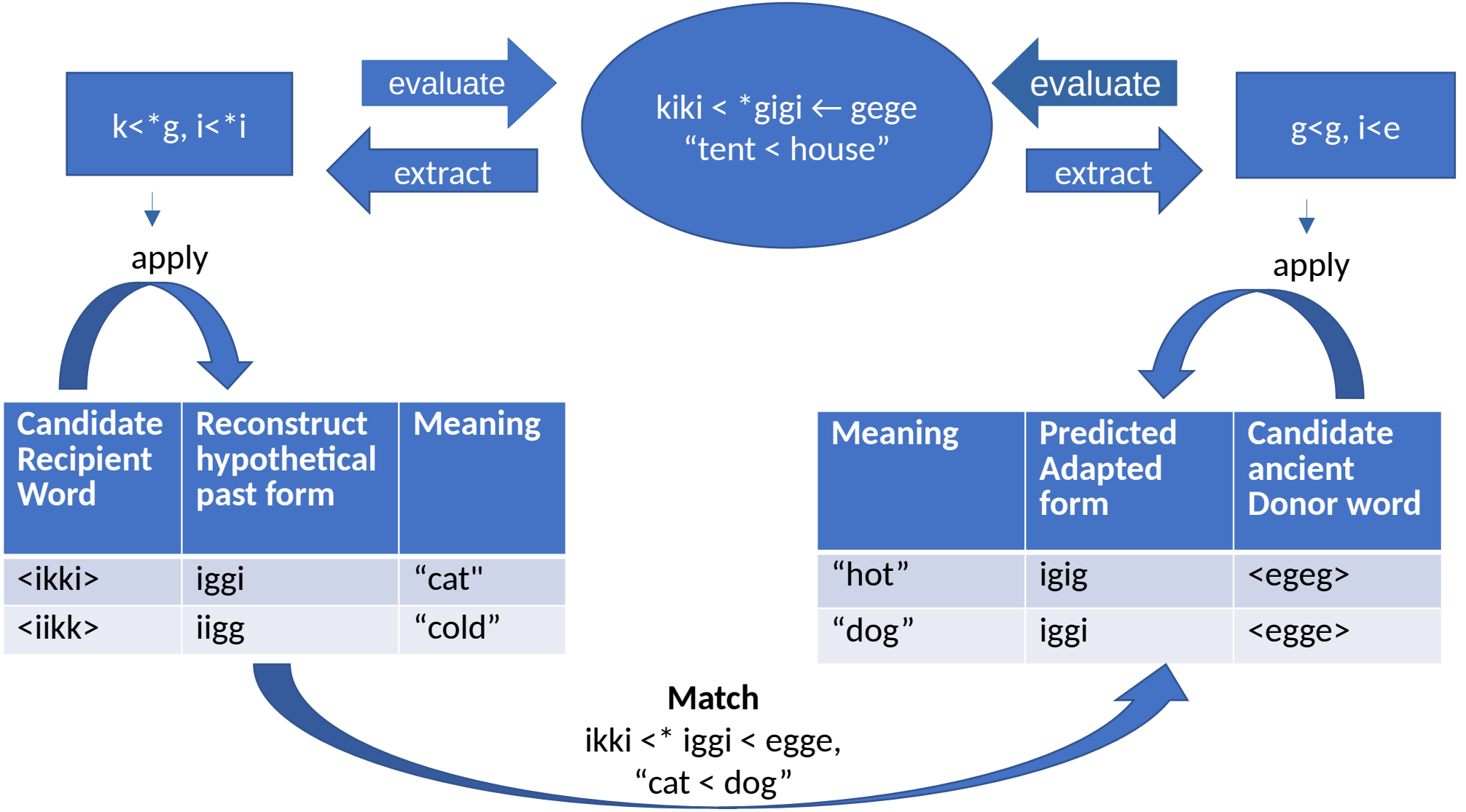
huhu → kuku
bebe → pipi
habe → kapi

Applying word vectors

After phonetic matching of forms:

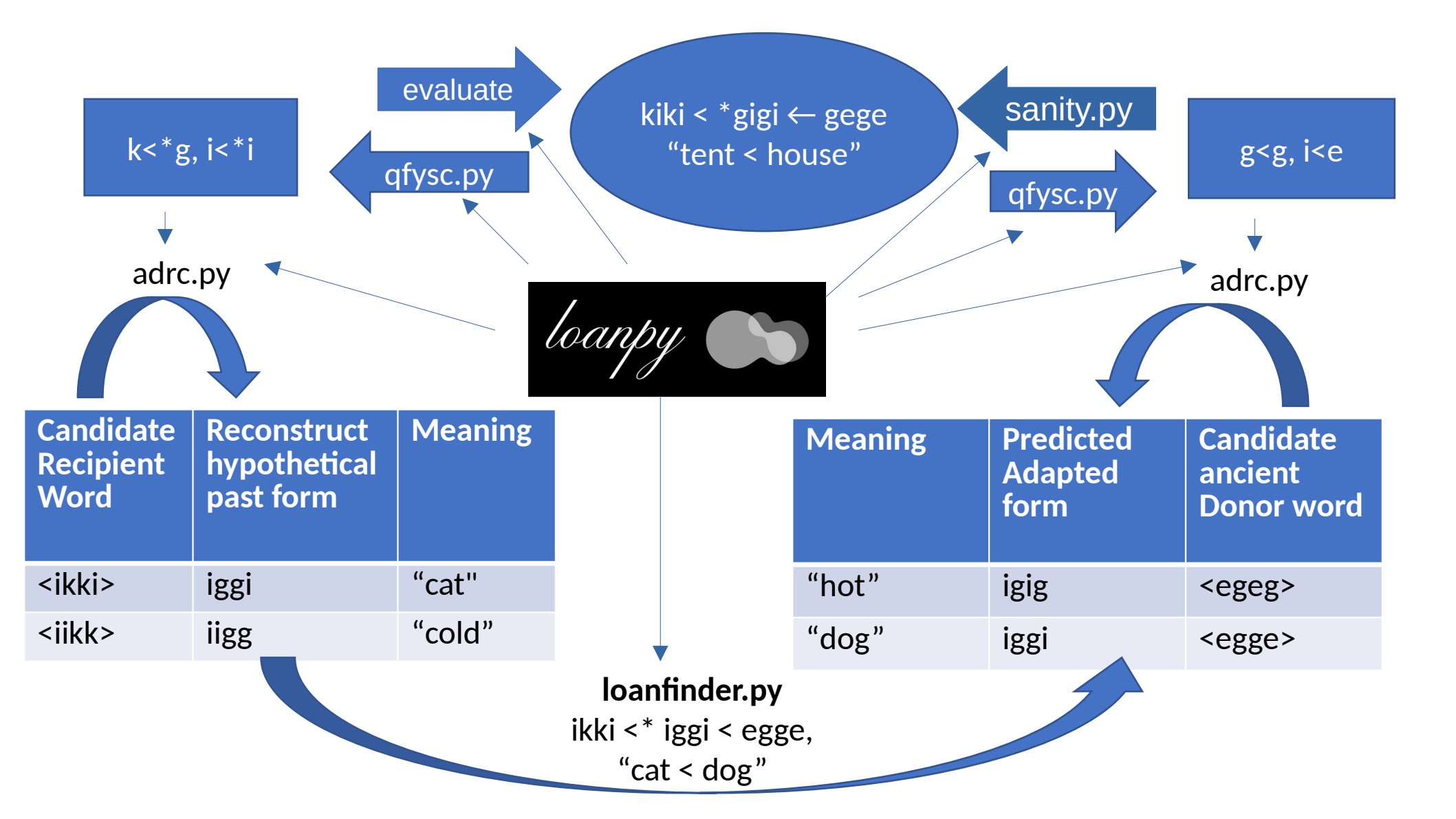
Form lg1	Form lg2	Meaning lg1	Meaning lg2	semantic similarity
kuku	kuku	‘bad’	‘sad’	0.9
pipi	pipi	‘nice’	‘fun’	0.9
kapi	kapi	‘fire’	‘burn’	0.85


Sound change rules
evaluate
extract
Etymological data
evaluate
extract
Heuristics + Sound adaptation rules
apply
apply
Candidate Recipient Word
Reconstruct hypothetical past form
Meaning
<abc>
rc1
'X'
<def>
rc2
'Y'
Meaning
Predicted Adapted form
Meaning
'Z'
ad1
<ghi>
'A'
ad2
<jkl>
Find matches



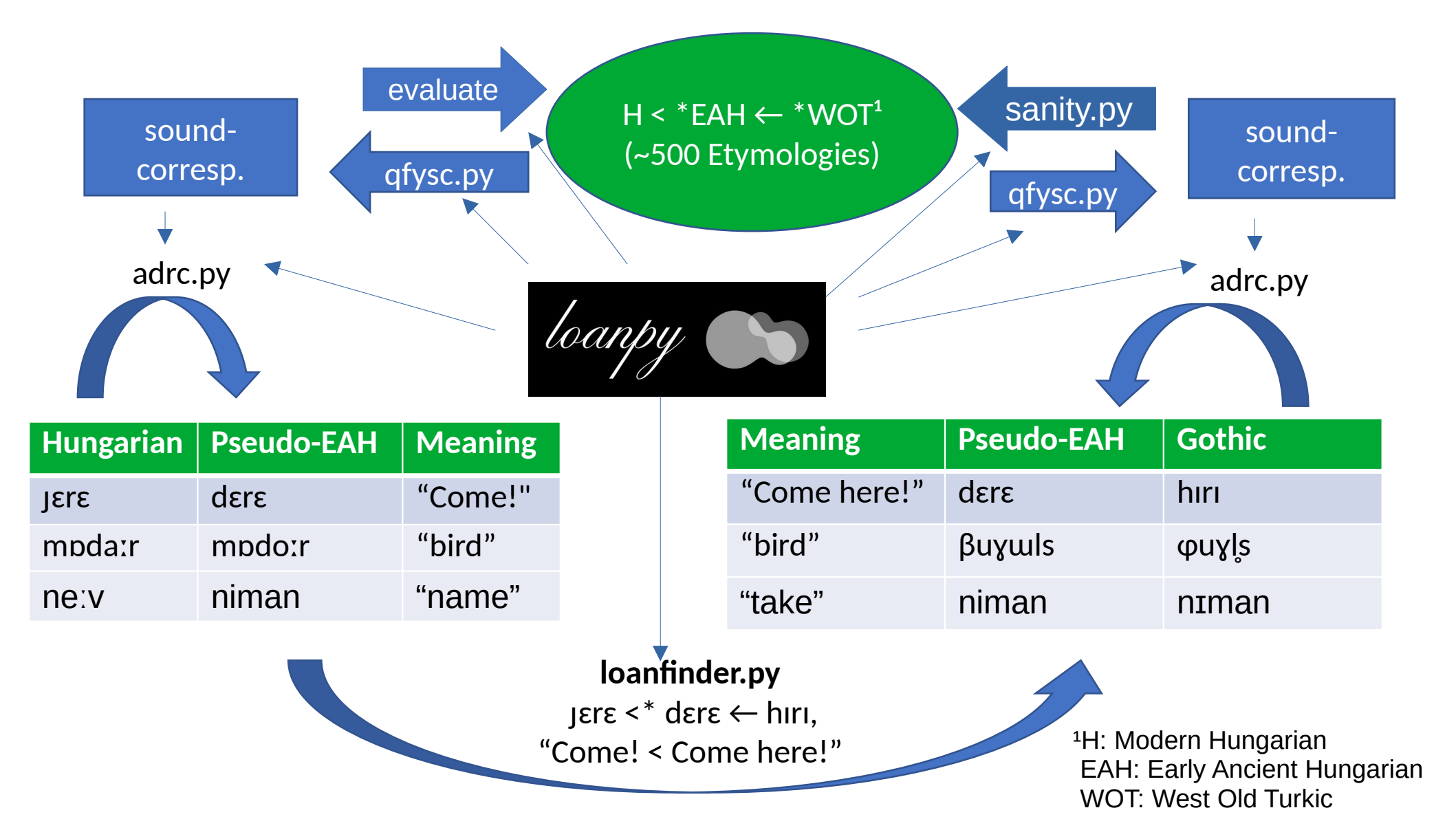
Candidate Recipient Word	Reconstruct hypothetical past form	Meaning
<ikki>	iggi	“cat"
<iikk>	iigg	“cold”

Meaning	Predicted Adapted form	Candidate ancient Donor word
“hot”	igig	<egeg>
“dog”	iggi	<egge>



Candidate Recipient Word	Reconstruct hypothetical past form	Meaning
<ikki>	iggi	“cat"
<iikk>	iigg	“cold”

Meaning	Predicted Adapted form	Candidate ancient Donor word
“hot”	igig	<egeg>
“dog”	iggi	<egge>



Hungarian	Pseudo-EAH	Meaning
ɟɛrɛ	dɛrɛ	“Come!"
mɒdaːr	mɒdoːr	“bird”
neːv	niman	“name”

Meaning	Pseudo-EAH	Gothic
“Come here!”	dɛrɛ	hɪrɪ
“bird”	βuɣɯls	ɸuɣl̥s
“take”	niman	nɪman

loanfinder.py

ɟɛrɛ <* dɛrɛ ← hɪrɪ,

“Come! < Come here!”

¹H: Modern Hungarian
EAH: Early Ancient Hungarian
WOT: West Old Turkic

Ways ahead

- error analyses of predictions
- align sounds with EDICTOR
- add front-end
- get better word-vectors
- make predictions with AI
- create gold-standard
- chance-similarity baseline?