



A Morphological Analyzer for St. Lawrence Island / Central Siberian Yupik

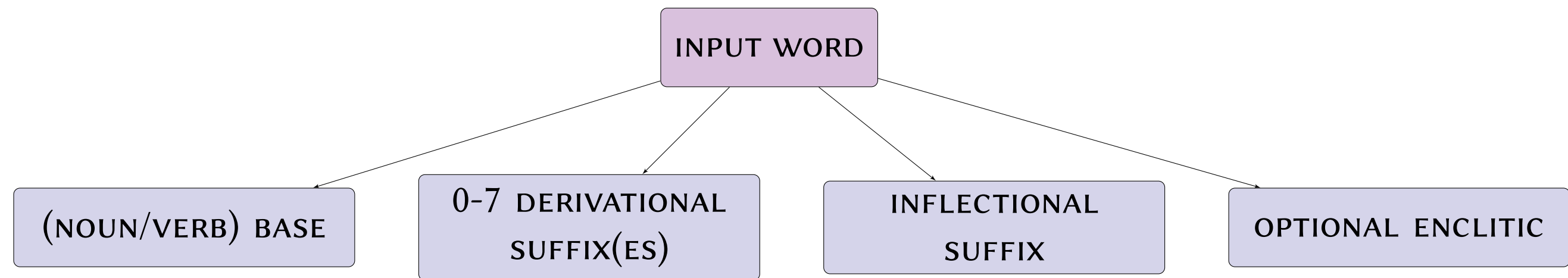
Emily Chen

Lane Schwartz

University of Illinois Urbana-Champaign
echen41@illinois.edu lanes@illinois.edu

1. Introduction

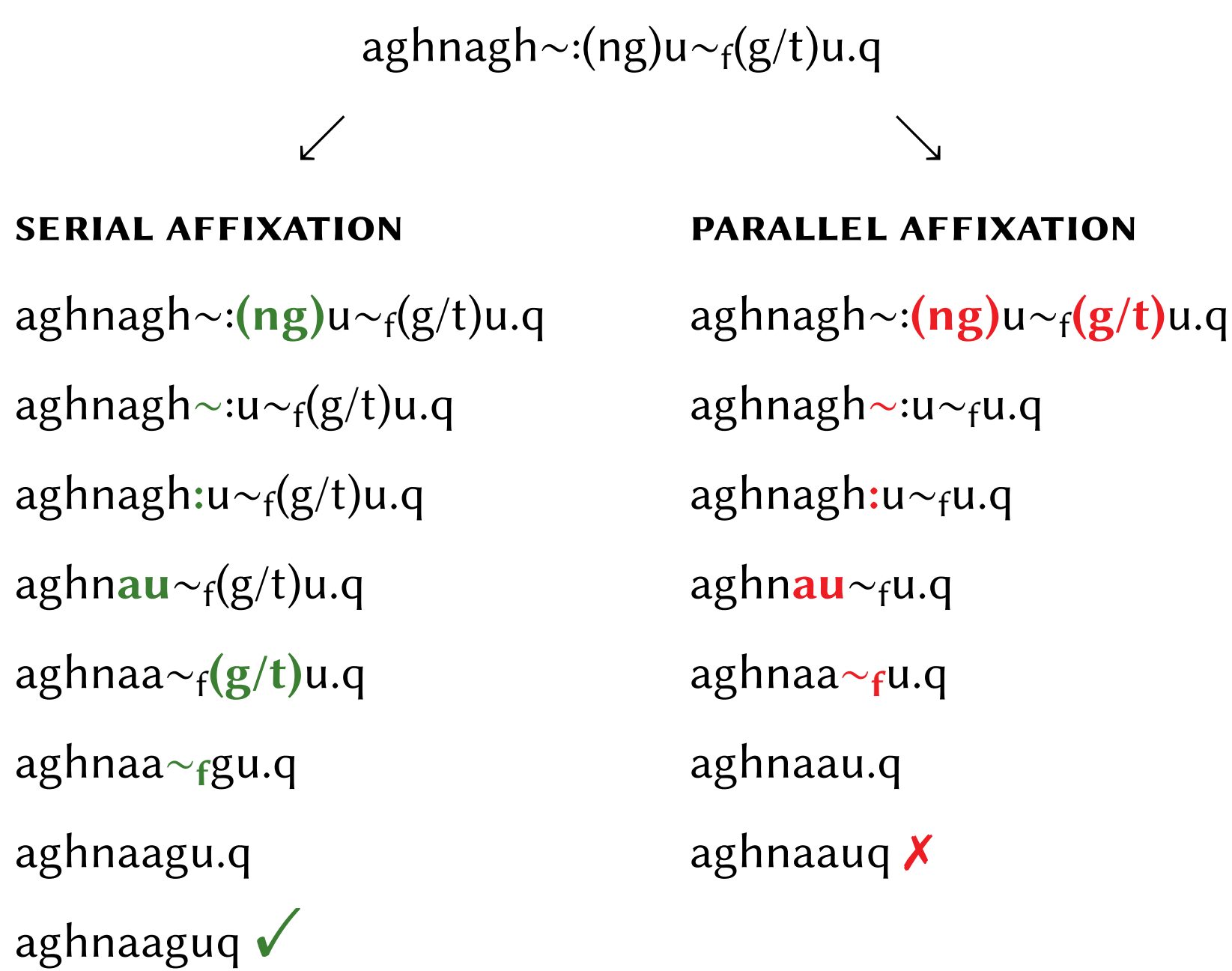
- PROJECT OBJECTIVES: To develop and evaluate a morphological analyzer for Central Siberian Yupik, a highly **agglutinative** and **polysynthetic** language of the Inuit-Yupik language family
- The analyzer parses a given word into its component morphemes, in accordance with the established template of most Yupik words



2. Yupik Morphophonology

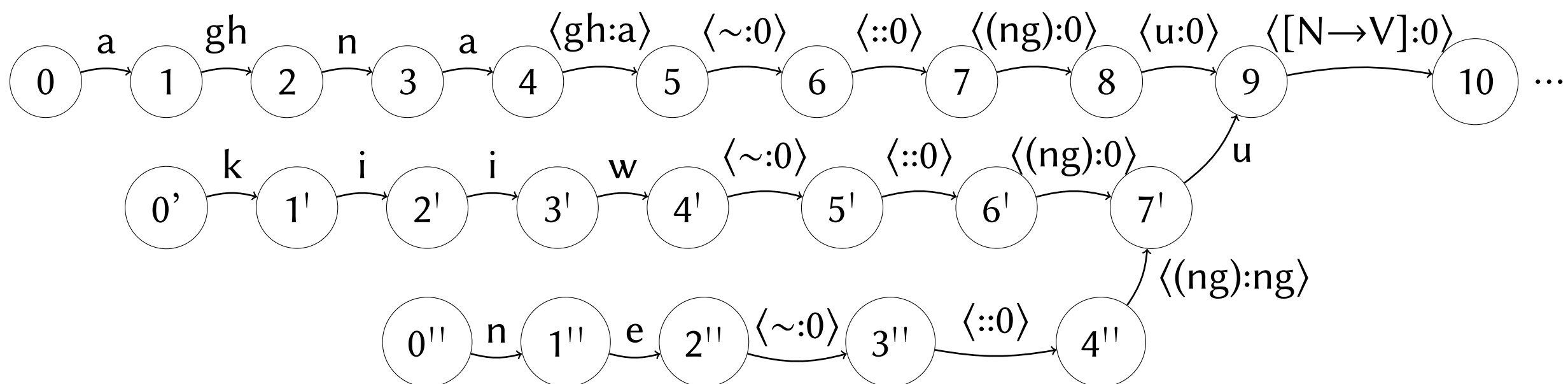
- Each suffix or **postbase** is associated with morphophonological processes that occur at the immediate left base-postbase boundary, where affixation of postbases proceeds **serially**
- (1) aghnaaguq
aghnaagh- ~:(ng)u- ~r(g/t)u.q
woman- -to.be.N- -INTR_IND.3SG
'She is a woman'

Symbol	Morphophonological Process
(ng)	Affixes -ng- to V-final bases
~	Drops penultimate and base-final -e
:	Drops uvulars, -gh-, between vowels
(g/t)	Affixes -g- to VV-final bases and -t- to C-final bases
~f	Drops base-final -e



3. Foma Toolkit

- The analyzer is implemented in a specialized finite-state toolkit called **Foma**



3.1 Lexc File

```
LEXICON NBase
aghnaagh NPostbase;

LEXICON NPostbase
~%:(ng)u[N→V]:~%:(ng)u VInfl;

LEXICON VInfl
[V] [Intr] [Ind] [3Sg] :~f(g/t)uq #;
```

- The *lexc* file constructs underlying forms according to Yupik morphotactics, which are mapped to corresponding intermediate forms

- aghnaagh~:(ng)u[N→V][V][Intr][Ind][3Sg]
↓
aghnaagh~:(ng)u~r(g/t)uq

3.2 Foma File

- The *foma* file transforms underlying forms into surface forms via rules defined by the programmer
- To model serial affixation of morphemes
 1. Insert a ^ (*MBndry*) at each morpheme boundary
 2. Specify a symbol / its morphonological process applies *only if* it is preceded by a single caret
- aghnaagh~:(ng)u~r(g/t)uq
↓
aghnaagh^ ~:(ng)u^ ~r(g/t)uq
↓
aghnaa^ ~r(g/t)uq
↓
aghnaaguq
- The morphophonological rules are iterated eight times to account for seven potential derivational postbases and one inflectional postbase

```
define ResolveAllomorphy
  "(ng)" -> ng || V MBndry MPSym* _ .o.
  "(g/t)" -> g || V V MBndry MPSym* _ .o.
  "(g/t)" -> t || C MBndry MPSym* _ .o.
  "(ng)" -> 0,
  "(g/t)" -> 0 || WBndry Alph+ MBndry _ ;

define UvularDropping
  "gh" -> 0 || V _ MBndry MPSym* "':" V .o.
  "':" -> 0 || WBndry Alph+ MBndry _ ;

define FinalE
  e -> 0 || _ MBndry MPSym* "~f" .o.
  "~f" -> 0 || WBndry Alph+ MBndry _ ;

define VowelDominance
  a u -> a a ;

define Grammar [
  InsertMBndry .o.
  !! ITERATION 1 !!
  MP Rules Cascade .o.
  CleanupMBndry .o.
  ;
  !! ITERATION 8 !!
  MP Rules Cascade .o.
  CleanupMBndry .o.
  ;
];
```

4. Sample Output

- ✓ Sample Output

Piyukuvek qergesek qelpeghtikek
'Open the window(s) if you (take a) walk'
piyukuvek piyug[V][Intr][Cond][2Sg]
qergesek qergese[N][Abs][Unpd][Du]
qelpeghtikek qelpeghte[V][Trns][Opt][PRS][2Sg][3Du]

Tukuqa neghsameng gaaghaquq
'My host is cooking seal'
tukuqa tukugh[N][Abs][1SgPoss][SgPosd]
neghsameng neghsagh[N][Abl_Mod][Unpd][Sg]
gaaghaquq gaagh~(g_i)aqe[V→V][V][Intr][Ind][3Sg]

- ✗ Sample Output

nuyaqatakestaaghaaguq +?
sivuqaghmun +?
nome-emun +?
aelqat +?

5. Evaluation

- The evaluation corpus consisted of 8 Yupik texts:
 - 1 collection of translation exercises from a reference grammar (~ 800 words)
 - 4 anthologies of Yupik folk stories (~7000-16,000 words each)
 - 3 elementary readers (~5000 words each)
- Coverage percentage of the analyzer for each of the eight evaluation texts at each stage of development was calculated, where

$$\text{Coverage \%} = \frac{\text{Number of Words Analyzed}}{\text{Number of Words in Text}}$$

Stage	Evaluation Texts							
	Ref	SLI1	SLI2	SLI3	Ungi	Lvl1	Lvl2	Lvl3
1	92.07	32.51	35.54	37.64	47.29	38.65	39.53	38.25
2	95.35	47.99	47.78	50.74	59.44	50.01	51.03	50.81
3	95.35	51.45	51.43	54.25	63.95	52.74	53.96	53.71
4	95.85	64.38	62.76	63.36	74.05	64.49	65.71	64.77
5	96.60	73.41	73.57	73.29	82.66	73.86	72.45	74.66