



**HAL**  
open science

# Advancing the study of endangered languages with computational tools for morphology

Dimitri Lévêque, Thomas Pellard

► **To cite this version:**

Dimitri Lévêque, Thomas Pellard. Advancing the study of endangered languages with computational tools for morphology: The case of Asama verb paradigms. LIFT 2019: Scientific meeting of the “Computational, formal & field linguistics” research group, Nov 2019, Orléans, France. , 2019. hal-02428825

**HAL Id: hal-02428825**

**<https://hal.science/hal-02428825>**

Submitted on 13 Mar 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# ADVANCING THE STUDY OF ENDANGERED LANGUAGES WITH COMPUTATIONAL TOOLS FOR MORPHOLOGY

## The case of Asama verb paradigms

Dimitri Lévêque  
Inalco-CNRS-EHESS, CRLAO  
leveque.dimitri@gmail.com

Thomas Pellard  
CNRS-EHESS-Inalco, CRLAO  
thomas.pellard@cns.fr

### BENEFITS OF USING COMPUTATIONAL TOOLS

1. Test the accuracy and coverage of the analysis of the morphology.
  2. Partially automate the glossing of texts.
  3. Enable quantitative measures of morphological complexity.
- ⇒ Advance the study of endangered languages.

### ASAMA

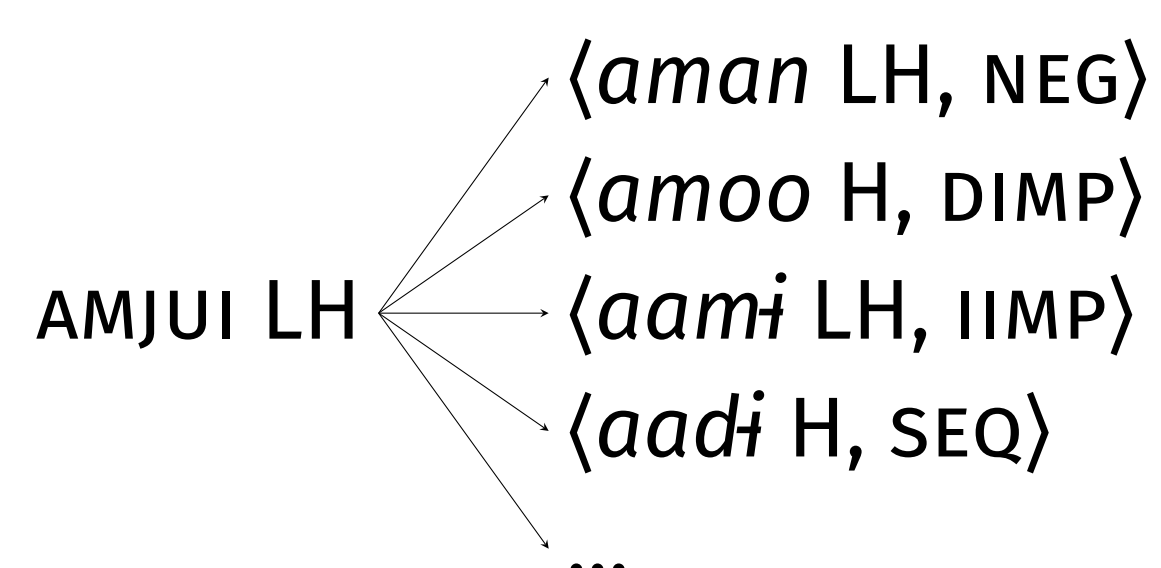


- Endangered Ryukuan language (Japonic family).
- Reference grammar project by D. Lévêque.
- 100 full + 400 near-full inflectional verb paradigms collected.
- Non-canonical phenomena (e.g. alternation of tones and vowel length) in verb inflection.

	'to sell'	'to knit'	'to go out'
NPST	ujui H	amjui H	izjjujui LH
CVB2	urug99si H	amjug99si H	izjirug99si LH
NEG	uran H	aman H	izjiran LH
DIMP	uroo H	amoo H	izjiroo LH
IIMP	ur#i H	aami HL	izjiri LH
CVB	ui H	amii H	izji LH
DES	uicjaahai H	amicjaahai H	izjicjaahai LH
PST	utan H	adan H	izjitan LH
SEQ	ut#i H	aadi HL	izjiti LH
PROG	utui LHL	aadui HL	izjitiui LHL

### 1 LINGUISTIC DOCUMENTATION

- GOALS** To **speed up and ease** the glossing of texts.
- ISSUES** Small amount of data, limited time, usual tools (Toolbox) not well-fitted to non-concatenative phenomena.
- SOLUTION** The transducer produces **complete paradigms** for all existing lexemes.



```
tourner -IPFV -PART -PROH máwaj -u -N -na
tourner -IPFV -PST -PART máwaj -ut -a -N
```

- RESULTS** A simple Python script is used to format the output, which can then be directly imported into Toolbox and used in the process of interlinearisation.

```
\lx máwajunna
\mo máwaj -u -n -na
\ge tourner -IPFV -PART -PROH

\lx máwajutna
\mo máwaj -ut -a -N
\ge tourner -IPFV -PST -PART
```

### WHY COMPUTATIONAL TOOLS?

- "The insufficiency of paper-and-pencil linguistics" (Karttunen 2006).
- Finite-state transducers (Beesley & Karttunen 2003).
- Bidirectional:
  - PRODUCTION** lexeme/root → inflected form;
  - RECOGNITION** inflected form → lexeme/root.
- Syntax similar to that of phonological rewrite rules familiar to many linguists.

```
regex V -> 0 | | V V _ ;
regex i -> i | | r _ ;
regex 0 -> j | | s _ a ;
```

- FSTs have sufficient power to handle (almost) any morphological phenomena.
- Provide a maximally precise and explicit description.
- Free software implementations (Foma).

### 2 LINGUISTIC ANALYSIS

- GOALS** **Check** the accuracy of the description and test hypotheses on new lexemes.
- ISSUES** Non-canonical verb morphology, multiple classes.
- DESCRIPTION** Two orthogonal sets of rules associated with two sets of classes.

	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	T <sub>4</sub>	T <sub>5</sub>
Xbj	-	Xb	Xb	Xd	
Xmj	-	Xm	Xm	Xd	
Xkj	-	Xk	Xk	Xcj	
Xcj	-	Xc	Xt	Xccj	
Xj	X	Xr	Xr	Xcj	
Xj	X	Xr	Xr	Xccj	
Xj	X	X	Xr	Xtt	

Class	NPST	IMP.DIR	SEQ
I1	RT <sub>1</sub> ui (H)	R:T <sub>4</sub> i (H)	RT <sub>5</sub> ui (H)
I2	RT <sub>1</sub> ui (H)	R:T <sub>4</sub> i (H)	RT <sub>5</sub> ui (H)
II1	RT <sub>1</sub> ui (H)	RT <sub>4</sub> #i (H)	RT <sub>5</sub> ui (LHL)
II2	RT <sub>1</sub> ui (H)	R:T <sub>4</sub> i (H)	RT <sub>5</sub> ui (LHL)
III	RT <sub>1</sub> ui (LH)	R:T <sub>4</sub> i (HL)	R:T <sub>5</sub> ui (HL)
IV	RT <sub>1</sub> ui (LH)	R:T <sub>4</sub> i (LH)	R:T <sub>5</sub> ui (LHL)
V	RT <sub>1</sub> ui (HL)	RT <sub>4</sub> i (HL)	RT <sub>5</sub> ui (HL)

### SOLUTIONS

- FSTs allow to:
- Check the formal description on lexemes and highlight the problems (comparison between data and output);
  - Test the results with the speaker (computation of the whole paradigm for lexemes with incomplete data and check with the speaker).

- RESULTS** Robustness of formal analysis is confirmed by the computational study, and lacks of data in the description are highlighted.

### IMPLEMENTATION

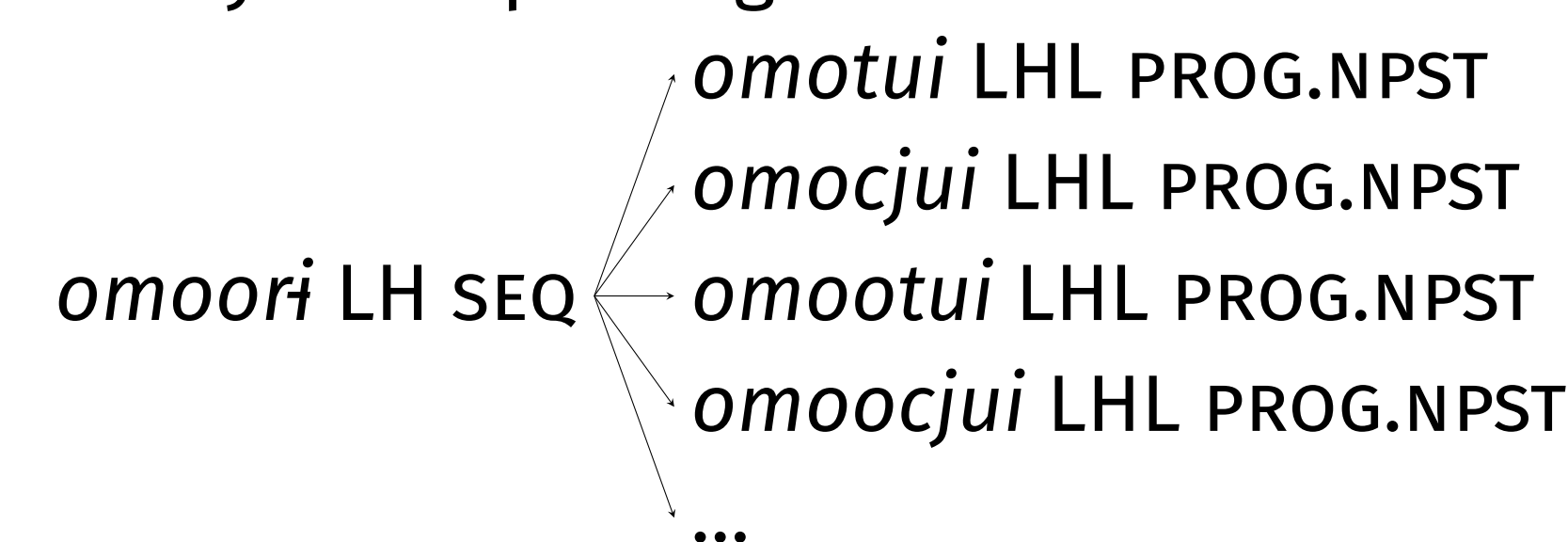
- Content paradigm cell  $\langle L, \sigma \rangle \rightarrow$  realised cell  $w$  ( $\langle \text{AMJUI LH, SEQ} \rangle \rightarrow \text{aadi HL}$ )
- Unlabeled form  $w \rightarrow$  all possible analyses  $\langle L, \sigma \rangle$ 
  - $\langle \text{AMJUI LH, SEQ} \rangle$
  - $\langle \text{AAMJUI H, SEQ} \rangle$
  - $\langle \text{ABJUI LH, SEQ} \rangle$
  - $\langle \text{AABJUI H, SEQ} \rangle$
- $L \rightarrow$  full realised paradigm, i.e. a set of cells  $\{\langle w, \sigma \rangle, \langle w', \sigma' \rangle, \dots\}$ .
- Any realised form  $\langle L, \sigma \rangle \rightarrow$  list of all possible realised forms of any other paradigm cell.

### 3 THEORETICAL MORPHOLOGY

- GOALS** Exploring the **implicative structure** with complexity measures and Shannon **entropy** and solve the *Paradigm Cell Filling Problem* ("What licenses reliable inferences about the inflected (and derived) surface forms of a lexical item?", Ackerman et al. 2009: 54).

- ISSUES** Too few full paradigms, no pre-existing interactive database.

- SOLUTION** Composition of two FSTs in order to obtain for any given realized cell  $\langle w, \sigma \rangle$  the list of all possible realized forms of any other paradigm cell.



- RESULTS** Entropy measures help identifying the **sources of uncertainty** (unpredictable segmental alternations, neutralisation of vowel length and neutralisation of tone) and the **principal parts** of the system (Finkel & Stump 2007).

H(C L)	NPST	CONV	IIMP	SEQ	PROG.NPST
NPST		0.000	0.000	0.222	0.244
CONV	0.923		0.082	0.296	0.550
IIMP	0.950	0.211		0.222	0.525
SEQ	1.362	1.012	0.425		0.317
PROG.NPST	1.204	0.910	0.423	0.000	

### REFERENCES

Ackerman, Farrell & Blevins, James P. & Malouf, Robert. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In Blevins, James P. & Blevins, Juliette (eds.), *Analogy in grammar: Form and acquisition*, 54–81. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199547548.003.0003>.

Ackerman, Farrell & Malouf, Robert. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89(3), 429–464. <https://doi.org/10.1353/lan.2013.0054>.

Beesley, Kenneth & Karttunen, Lauri. 2003. *Finite state morphology*. Stanford: Center for the Study of Language and Information Publications.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

Bonami, Olivier & Luís, Ana R. 2015. Sur la morphologie implicative dans la conjugaison du portugais: Une étude quantitative. In Léonard, Jean-Léo (ed.), *Morphologie flexionnelle et dialectologie romane*, 111–151. Leuven: Peeters.

Finkel, Raphael & Stump, Gregory. 2007. Principal parts and morphological typology. *Morphology* 17, 39–75. <https://doi.org/10.1007/s11525-007-9115-9>.

Hulden, Mans. 2009. Foma: A finite-state compiler and library. In *Proceedings of the Demonstrations Session at EAFL 2009*, 29–32. <https://www.aclweb.org/anthology/E09-2008>.

Jacques, Guillaume & Lahaussois, Aimée & Michailovsky, Boyd & Rai, Dhan Bahadur. 2012. An overview of Khaling verbal morphology. *Language and Linguistics* 13(6), 1095–1170. [http://www.ling.sinica.edu.tw/Files/LL/Documents/Journals/13\\_6/j2012\\_6\\_03\\_7314.pdf](http://www.ling.sinica.edu.tw/Files/LL/Documents/Journals/13_6/j2012_6_03_7314.pdf).

Karttunen, Lauri. 2003. Computing with realizational morphology. In Gelbukh, Alexander (ed.), *Computational linguistics and intelligent text processing*, 203–214. Berlin: Springer. [https://doi.org/10.1007/3-540-36456-0\\_20](https://doi.org/10.1007/3-540-36456-0_20).

Karttunen, Lauri. 2006. The insufficiency of paper-and-pencil linguistics: The case of Finnish prosody. In Butt, Miriam & Dalrymple, Mary & Holloway King, Tracy (eds.), *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*, 287–300. Stanford: Center for the Study of Language and Information Publications. <http://roa.rutgers.edu/article/view/828>.

Matthews, Peter H. 1972. *Morphology*. 1st edn. Cambridge: Cambridge University Press.

Pellard, Thomas & Yamada, Masahiro. 2017. Verb morphology and conjugation classes in Dunan (Yonaguni). In Kiefer, Ferenc & Blevins, James P. & Bartos, Huba (eds.), *Perspectives on morphological organization*, 31–49. Leiden: Brill. [https://doi.org/10.1163/9789004342934\\_004](https://doi.org/10.1163/9789004342934_004). <https://hal.archives-ouvertes.fr/hal-01493096>.

Snoek, Conor & Thunder, Dorothy & Loo, Kaidi & Arppe, Antti & Lachler, Jordan & Moshagen, Sjur & Trosterud, Trond. 2014. Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 34–42. <https://doi.org/10.3115/v1/W14-2205>.

Stump, Gregory T. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CB09780511486333>.

Stump, Gregory & Finkel, Raphael. 2013. *Morphological typology: From word to paradigm*. Cambridge: Cambridge University Press.