Towards a Systematic and Quantitative Identification of Sound Change Rules – Algonquian Historical Linguistics Meets Optical Character Recognition Evaluation

Antti Arppe[1], Katherine Schmirler[1], Eddie Antonio Santos[1,2] & Arok Wolvengrey[3]
[1] Alberta Language Technology Lab, University of Alberta
[2] Canadian Indigenous Languages Technology Project, National Research Council of Canada
[3] First Nations University of Canada

Within the historical-comparative methodological tradition in linguistics, rules concerning sound changes have conventionally been devised painstakingly manually by individual linguists, attempting to consider all the relevant data as to the regularities through which individual sounds at some earlier of a language could be seen to have changed. These sound change rules concern both the morpho-phonemic context in which a particular phonemic change occurred, as well as the temporal order in which these changes took place, since some such rules require some other changes to have applied earlier in order to be triggered. Typically, these rules are based on a relatively restricted set of attested or posited forms in the antecedent proto-language (at most ranging up several thousand forms). A reverse instantiation of such rules is working through them backwards in time to reconstruct historically posited antecedent forms for contemporaneous vocabulary. But to what extent can one be confident that historical linguists have systematically considered the entire available data, without accidentally missing some individual forms, or some more subtle patterns or rules? And might one be able to quantify how broadly or narrowly the posited rules have applied across the available data, and how much of the phonemic make-up of the proto-stage language has remained immutable over the passage of time?

(1) Historical morphonological rules from Proto-Algonquian to modern Cree
- a. t → c / _ $ [is | isis | si] $ (nominal and verbal diminutives)
- b. ê → â / $ _ $n $
- c. Palatalization rules
  - i. T → t | s / _ $ i | î
  - ii. t → t | c / _ $ i | î
  - iii. T | t → t / _ $ *e
- d. *e → i
- e. Ø → i / C $ _ C
- f. Vowel-glide-*i* rules
  - i. [a | â] [w | y | ý] $ i → â
  - ii. [ê [w | y | ý] $ i → ê
  - iii. [i | î] [w | y | ý] $ [i | î] → î
  - iv. [o | ô] [w | y | ý] $ i → ô
- g. V-V rules
  - i. [a | â] $ [a | â] → â
  - ii. [ê] $ [ê] → ê
  - iii. [i | î] $ [i | î] → î
  - iv. [o | ô] $ [o | ô] → ô
- h. w $ o → [o | ô] / C _
- i. w $ w → [w | ow] / C _ V
- j. [m | n] $ → h / _ [p | t | k | c]

Sound change rules applying for the evolution from proto-Algonquian to Plains Cree, a contemporary Algonquian language spoken primarily in Western Canada, have been worked out

quite comprehensively over the last century, resulting in a set of 18 primary rules under 10 broader groupings (for overview, see example 1 above, based on example 10 in Arppe et al. 2018) that have been gleaned from Cook and Muehlbauer (2010), Wolfart (1996), Wolvengrey (2001), and other sources. These historical morphophonological rules have been previously turned into an ordered series of formal rewrite-rules to create a probabilistic, weighted finite-state-based computational model of the derivation of modern Plains Cree stems from their proto-Algonquian antecedents (Anonymous 2018), which was trained and tested with an extensive collection of pairings of modern Cree stems and their proto-Algonquian morphological decompositions provided in the lexical database underlying Wolvengrey (2001), the most comprehensive current lexical resource for Plains Cree. They concluded that approximately four-fifths of the contemporary forms could be fully derived and explained in terms of the aforementioned general historical rules, whereas the remainder represented idiosyncrecies and contextually much more restricted rules. But could this extensive historical linguistic dataset be used to systematically infer and verify these same rules?

Evaluation toolkits for Optical Character Recognition include functionalities for creating *confusion matrices*, which show how often various single or multiple character sequences from the *ground truth* (i.e. a validated transcription of the original scanned source) are correctly or incorrectly interpreted by the Optical Character Recognition process (e.g. the `accuracy` command in Santos, forthcoming 2019). We decided to treat the proto-Algonquian forms as the verified "correct" original source and the matching contemporary Cree stems as the output, resulting from overall historical sound change processes (comparable at a meta-level to what takes place in OCR) and reflecting whatever phonological "mismatches" exist between current Cree and proto-Algonquian. Of course, this is dependent on the extent that the orthograhical standard for the modern and proto-Algonquian forms in Wolvengrey (2001) essentially faithfully reflects their actual phonemic forms now and in the past. Since `accuracy` prints statistics for how often a confusion between the (verified) input and output is found, this function inadvertently becomes a tool for reporting systematic changes, along with how often the effect is attested, of which the most frequent are shown in (2) below.

(2) Frequencies of sound sequence mismatches between Proto-Algonquian (PA) and modern Cree

N	Confusion: PA-Cree[1]	Rule(s) in (1)	Rewrite rule(s) in the historical computational model
2503	{/}-{i}	(e)	CiCEpenthesisRule
1912	{w/}-{o}	(e), (f)(iv)	CiCEpenthesisRule, CwiC2CoCRule
1628	{aw/i}-{â}	(f)(i)	aGiRule
1186	{t/}-{c}	(c)(ii), (a)	tcRule | dimRule
1140	{w/i}-{o}	(f)(iv)	CwiC2CoCRule
1008	{t/}-{ci}	(c)(ii), (e)	tcRule, CiCEpenthesisRule
916	{T/}-{t}	(c)(ii,iii)	TtRule
406	{/a}-{}	?	?
393	{/ê/}-{â}	(b)	ên2ânRule
322	{w/}-{}	?	?
320	{iy/i}-{î}	(f)(iii)	iGiRule
…	…	…	…

[1] The forward-slash character {/} designates a morpheme boundary in the reconstructed Proto-Algonquian form; {} denotes an empty string, i.e. deletion.

What we found was that the aforementioned established, general historical sound change rules from proto-Algonquian to Cree were the most frequent pairings in the confusion matrix, validating that linguists had previously quite correctly and comprehensively identified the most prevalent regular changes. What we could also establish are the relative scopes at which the various rules had impacted Cree vocabulary, as well as that a majority of sounds (58.7%) had in fact remained the same over the passage of time (at least relative to the phonological systems at the Proto-Algonquian stage and contemporay Cree). In addition, we were able to identify a number regular changes that appear to have been overlooked in the current compilation of the historical rules, but this could be explained due to their relative rarity in the data.

Importantly, this simple technique allowed us to both verify the existing body of knowledge concerning the phonological evolution from proto-Algonquian to contemporary Cree, empirically rank the sound change rules in terms of their scope of application in the known vocabulary, as well as provide some new though seemingly more secondary candidates for such sound change rules. Besides historical sound change, this same approach could be applied for identifying and quantifying regular transformations between dialects or text genres, or more generally any input-output pairings resulting from some regular linguistic process.

REFERENCES

Arppe, Antti, Katherine Schmirler, Miikka Silfverberg, Mans Hulden & Arok Wolvengrey (2018). Insights from computational modelling of the derivational structure of Plains Cree stems. *Papers of the 48th Algonquian Conference* (PAC48), 48, 1-18.

Cook, Clare and Jeffrey Muehlbauer (2010). A Morpheme Index of Plains Cree. Unpublished manuscript, University of Manitoba, Winnipeg. http://www.academia.edu/304874/A_morpheme_index_of_Plains_Cree.pdf

Santos, Eddie Antonio (forthcoming 2019). OCR evaluation tools for the 21st century. Proceedings of the *3rd Workshop on the Use of Computational Methods for Endangered Languages* (ComputEL-3), Honolulu, Hawai'i, 26-27 February 2019.

Wolfart, H. C. (1996). *Sketch of Cree, an Algonquian Language. Handbook of North American Indians. vol. 17, Languages, ed. by Ives Goddard, pp. 390–439. Washington, DC: Smithsonian Institution.*

Wolvengrey, Arok (2001). *nêhiyawêwin: itwêwina / Cree: Words*. Regina: University of Regina Press.