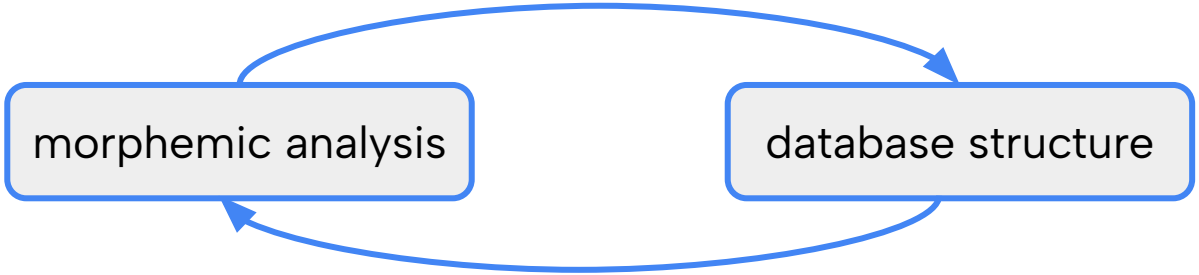
Blackfoot Words: A Lexical Database of a Polysynthetic Language

Natalie Weber, Jem Burch, Leander He, Corine Huang,
Katie Hur, & Alara O'Bryan

SSILA 2025 • Yale University
January 25, 2025

Overview

- Lexical database for a polysynthetic language
- Feedback loop:



- This talk: our methods for analysis with respect to
  - Stem **recursivity**
  - Multiple **templatic positions**
  - Lexical **categories**

Blackfoot Words database

- **~91,500** word tokens have been digitized by the Blackfoot Lab
- From **52** sources (wordlists, dictionaries, grammars)
  - All four major dialects
  - Timespan: 1743–2017

- Published:
  1. v1.0: 4,553 word tokens from 9 sources
  2. v1.1: words partially analyzed into stems and morphemes

- Forthcoming:
  3. v1.2: complete analysis

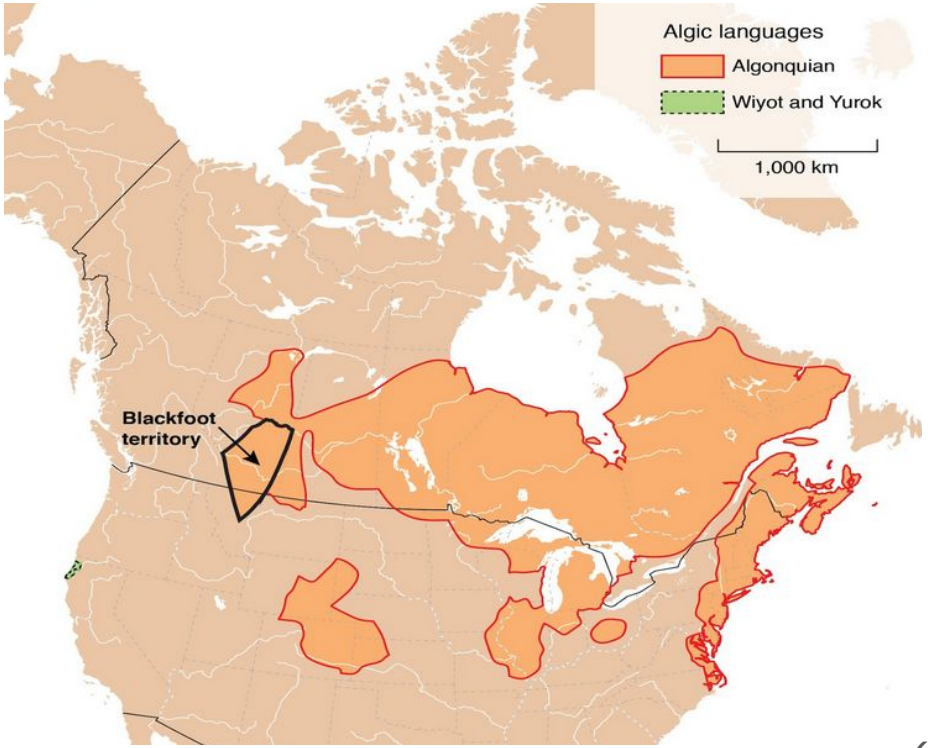
(Weber et al. 2023; https://www.blackfootwords.com/)

Roadmap

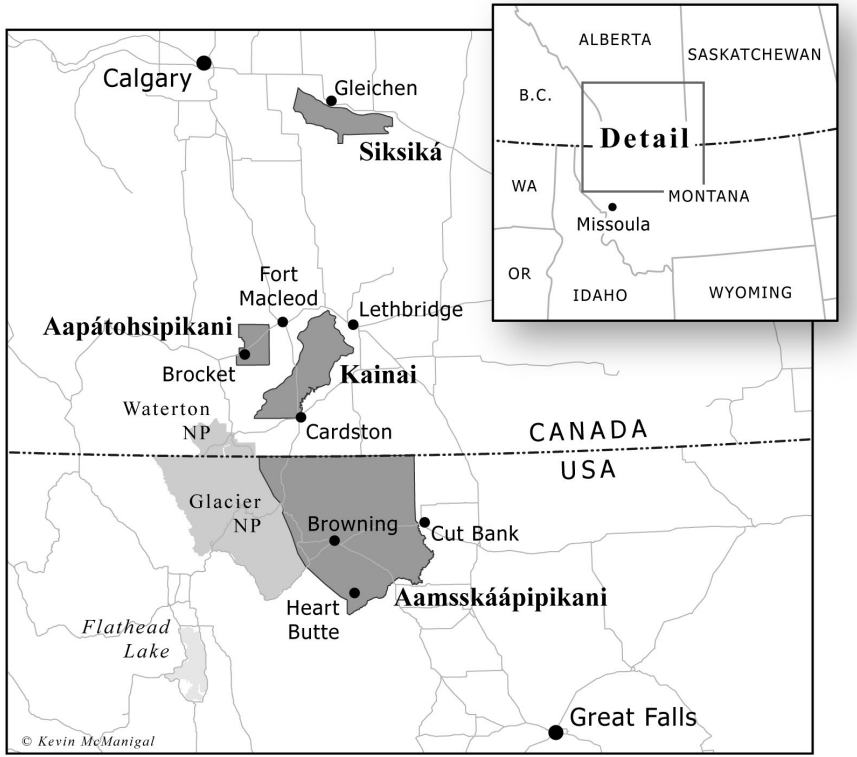
1. Background & Challenges
2. Database Structure
3. Categories
4. New Discoveries
5. Future Projects

01
Background & Challenges

Blackfoot Language



Map by Eric Leinberger



(Frantz 2017; Taylor 1969; Uhlenbeck 1938)

Morphological Templates

Verbs: per–preverb*–[stem]–infl

itáóhpokso’kaamiiwa
it–a–ohpok–**[[yo’k–aa]–m]**–ii–wa
then–IP□v–with–**[[sleep–AI]–TA]**–DIR–3
‘he sleeps with her’

- [stem] = initial–(medial)–final
- Stems can be recursive

Nouns: per–prenoun–[stem]–infl

isttohkisoka’simi
isttohk–[isoka’sim]–i
thin–[garment]–IN.□G
‘shirt’

- [stem] = usually monomorphemic
- Stems can compound

Challenge 1: Stem Recursivity

- How do we relate words that share subparts of form + meaning?

(1)

	OriginalWord	WordTranslation	Source
a.	**a-sú-kas-sĭm**	‘shirt’	Curtis (1911: 170)
b.	E-stok**e-so-char-sim**	‘A Shirt’	Umfreville (1790: 202)

- Umfreville (1790) does not list **e-so-char-sim** as a unique stem.

- The connections between (1a) and (1b) would be lost in a database that only contained words or (maximal) stems.

Challenge 2: Multiple Templatic Positions

- How do we track distribution across templatic positions?

(2)

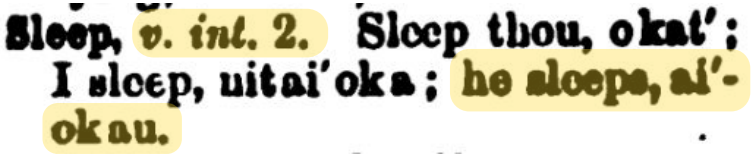
	OriginalWord	WordTranslation	Source
a.	**E-stok**[e-so-char-sim]	'A Shirt'	Umfreville (1790: 202)
b.	nits-[**ĭstok**-itsi]	'I lie down'	Tims (1889: 145)

- Database should encode two types of info:
  - morpheme position in each word token
  - which morphemes can freely vary across templatic positions

Challenge 3: Choosing category labels

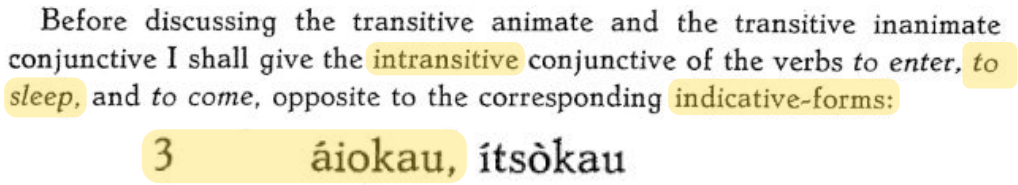
Categorization in sources differ; some sources do not use them at all

Tims (1889:168)



Sleep, *v. int. 2.* Sleep thou, okat'; I sleep, uitai'oka; he sleeps, ai'-okau.

Uhlenbeck (1938:161)



Before discussing the transitive animate and the transitive inanimate conjunctive I shall give the intransitive conjunctive of the verbs *to enter*, *to sleep*, and *to come*, opposite to the corresponding indicative-forms:

3 áiokau, ítsòkau

ai'okau, “he sleeps”

(v. int. 2)

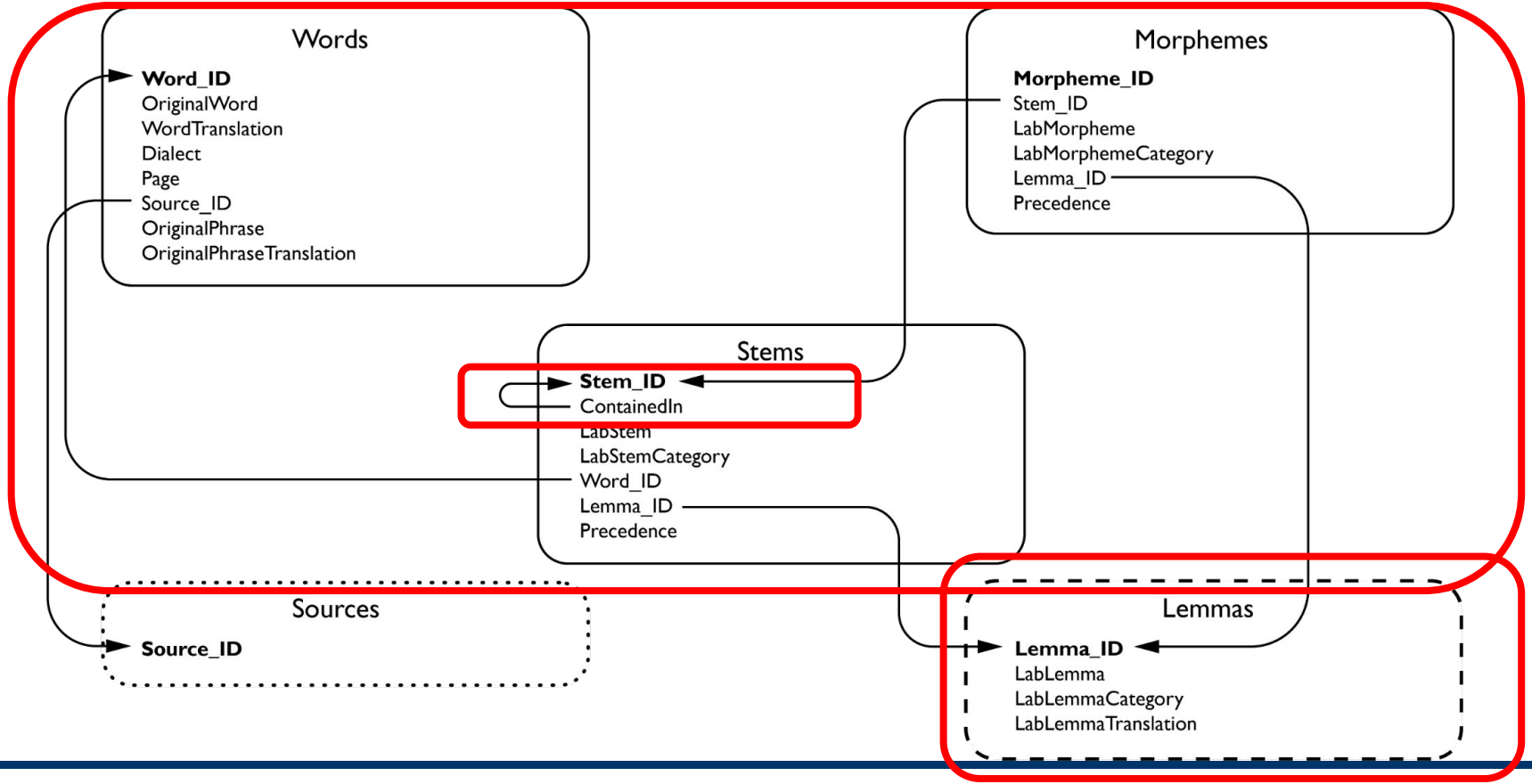
áiokau, “he sleeps”

(intransitive, indicative)

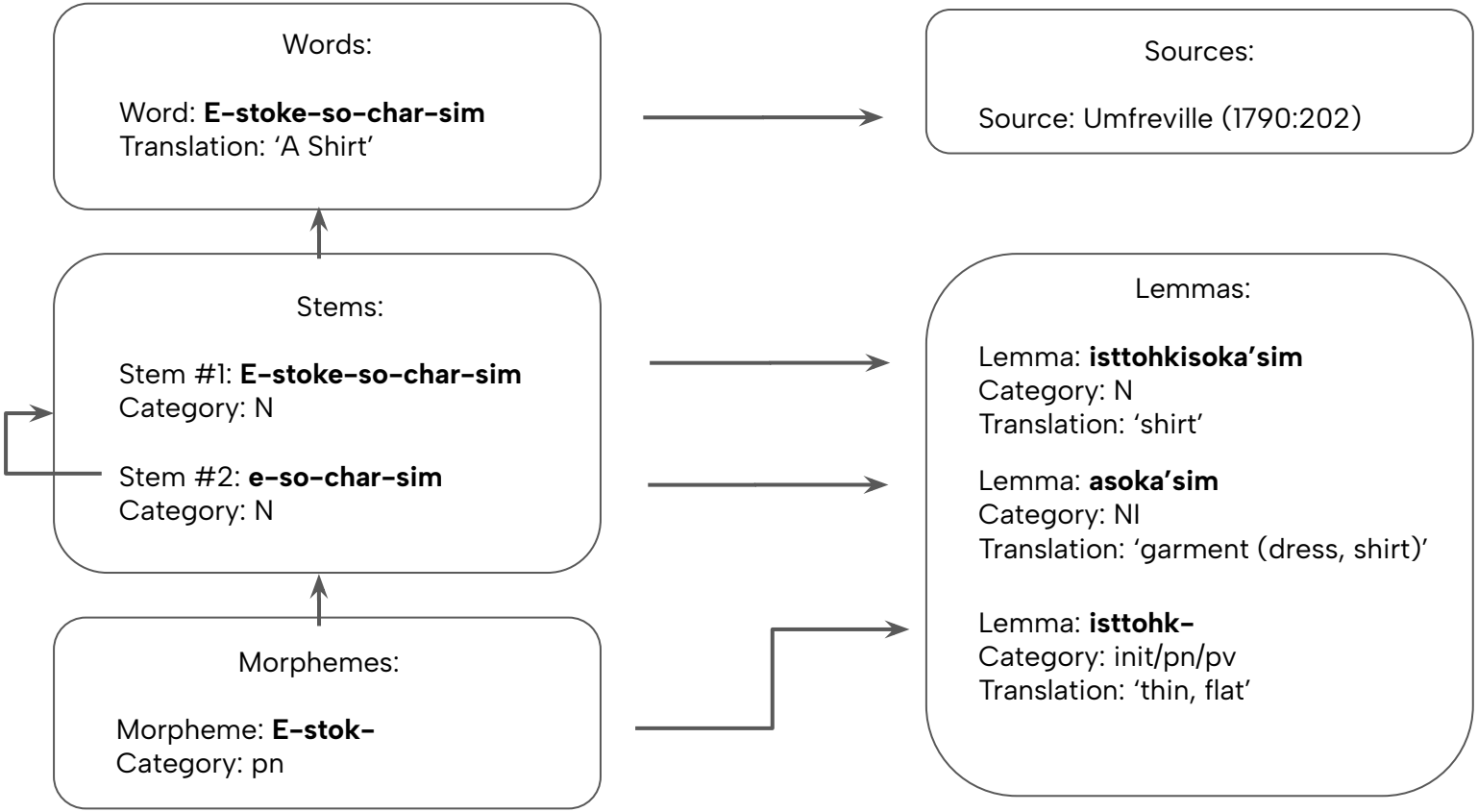
02
Database Structure

(Weber et al. 2023)

Structure: mySQL relational database



Illustration



Solution 1: Recursivity

- To encode relationships between subparts, we analyze each word into stems/ morphemes within separate tables
  - Maintain the original orthography for each entry
  - Link them via the ContainedIn or Stem_ID field:

Stems table

Stem_ID	LabStem	LabStemCategory	ContainedIn	Lemma_ID
stem-00004152	E-stoke-so-char-sim	N	NULL	lemma-0000304
stem-00004669	e-so-char-sim	N	**stem-00004152**	lemma-0000226

Morphemes table

Morpheme_ID	LabMorpheme	LabMorphemeCat.	Stem_ID	Lemma_ID
morph-00000264	E-stok-	pn	**stem-00004152**	lemma-0002388

Solution 2: Multiple Templatic Positions

- LabMorphemeCategory lists the position within each word token.
- LabLemmaCategory lists all positions the lemma is found in.
- Future research: investigate the distribution of stems and morphemes.

Morphemes table

Morpheme_ID	LabMorpheme	LabMorphemeCat.	Stem_ID	Lemma_ID
morph-00000264	E-stok-	**pn**	stem-00004152	lemma-0002388
morph-00000ABC	ĭstok-	**init**	stem-0000WXYZ	lemma-0002388

Lemmas table

Lemma_ID	LabLemma	LabLemmaCat.	LabLemmaTranslation
lemma-0000304	isttohkisoka'sim	N	shirt
lemma-0000226	asoka’sim	NI	garment (dress, shirt, coat)
lemma-0002388	isttohk-	**pn/pv/init**	thin

03
Categories

Solution 3: Choosing Category Labels

Abbrev.	LabWordCategory
N	Noun
PN	Proper Name
Pro	Pronoun
D	Demonstrative
V	Verb
Num	Numeral
Adj	Adjective
Adv	Adverb
Voc	Vocative
Conj	Conjunction
Int	Interjection
Part	Particle

Abbrev.	LabStemCategory
NA	Animate noun
NI	Inanimate noun
N	Noun
NDA	Dependent animate noun
NDI	Dependent inanimate noun
ND	Dependent noun
D	Demonstrative
D-wh	Interrogative demonstratives
D-one	‘One’-replacement demonstratives
VAI	Animate intransitive verb
VII	Inanimate intransitive verb
VTA	Transitive animate verb
VTI	Transitive inanimate verb
Num	Numeral
Voc	Vocative
Conj	Conjunction
Int	Interjection

Abbrev.	LabMorphemeCategory
fna	animate noun final
fni	inanimate noun final
fn	noun final
fnda	dependent animate noun final
fndi	dependent inanimate noun final
fnd	dependent noun final
fai	animate intransitive verb final
fii	inanimate intransitive verb final
fta	transitive animate verb final
fti	transitive inanimate verb final
init	initial
med	medial
pn	prenoun
pv	preverb

Categories: Words

- We record source lexical categories as “**OriginalWordCategory**” and “**OriginalPartialWordCategory**” for words and morphemes respectively
- The database has its own unified & uniform **word**, **stem**, and **morpheme** categories
- “**LabWordCategory**” determined by inflection, semantic differences, orthography, and phonology

Abbrev.	LabWordCategory	Inflection
N	Noun	nominal
PN	Proper Name	nominal
Pro	Pronoun	nominal
D	Demonstrative	nominal
V	Verb	verbal
Num	Numeral	verbal
Adj	Adjective	verbal
Adv	Adverb	(various)
Voc	Vocative	uninflected
Conj	Conjunction	uninflected
Int	Interjection	uninflected
Part	Particle	uninflected

Categories: Stems

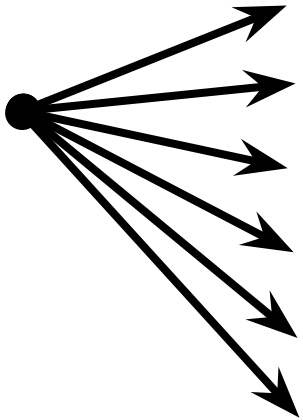
- **LabWordCategory** may be related to zero, one, or several stem categories
- **LabStemCategory** largely follows conventions within Algonquian studies (Bloomfield 1946; Goddard 1990)

Abbv.	Verb stem type	Subject	Object
VAI	Animate intransitive verb	animate	—
VII	Inanimate intransitive verb	inanimate	—
VTA	Transitive animate verb	—	animate
VTI	Transitive inanimate verb	—	inanimate

Many-to-Many Relationship

Word to Stem

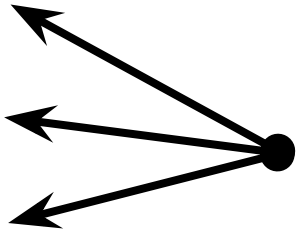
LabWordCategory	
N	Noun
PN	Proper Name
Pro	Pronoun



LabStemCategory	
NA	Animate noun
NI	Inanimate noun
N	Noun
NDA	Dependent animate noun
NDI	Dependent inanimate noun
ND	Dependent noun

Many-to-Many Relationship

LabWordCategory	
N	Noun
PN	Proper Name
Pro	Pronoun



Stem to Word

LabStemCategory	
NA	Animate noun
NI	Inanimate noun
N	Noun
NDA	Dependent animate noun
NDI	Dependent inanimate noun
ND	Dependent noun

04
New Discoveries

Distribution: subclasses of morphemes

- Some roots *only* occur in a particular position.
- Not always given a separate entry in Frantz & Russell (2017) = **new analysis**

Lemma_ID	LabLemma	LabLemmaTranslation	LabLemmaCategory
lemma-0002606	pain-	painful, sensitive ache	**init**
lemma-0000840	-sspin-	cheek, chin	**med**
lemma-0002260	aahk-	might	**pv**

Distribution: subclasses of morphemes

- Other stems and morphemes occur in *many* positions.
- LabLemmaCategory includes all possible positions.
- **Future research:** studies of medials and other positions.

Lemma_ID	LabLemma	LabLemmaTranslation	LabLemmaCategory
lemma-0002314	ssp-	high	**init/pn/pv**
lemma-0000078	-ota's	horse, dog	**NDA/med**
lemma-0001385	mokaki	s.o. is wise, careful	**VAI/pv**

- Confirms earlier research that Algonquian roots differ in distribution (Déchaine & Weber 2018)

Differences between verb and noun structure

- Verbs can stack many preverbs (in yellow below).
- So far: nouns can only have one prenoun.

Structure of word-AT1969-2888 itɔ́xpoksooʔkaamiiʔwa he sleeps with her

Word	ID	Category	Precedence	ContainedIn
itɔ́xpoksooʔkaamiiʔwa	word-AT1969-2888	V	—	NULL
↳itɔ́xpoksooʔkaam	stem-00003147	VTA	—	NULL
↳it-	morph-00002356	pv	1	stem-00003147
↳ɔ́-	morph-00002284	pv	2	stem-00003147
↳xpok-	morph-00002440	pv	3	stem-00003147
↳sooʔkaam	stem-00006184	VTA	4	stem-00003147
↳sooʔkaa	stem-00006291	VAI	1	stem-00006184
↳-m	morph-00002893	fta	2	stem-00006184

(Taylor 1969: 252)

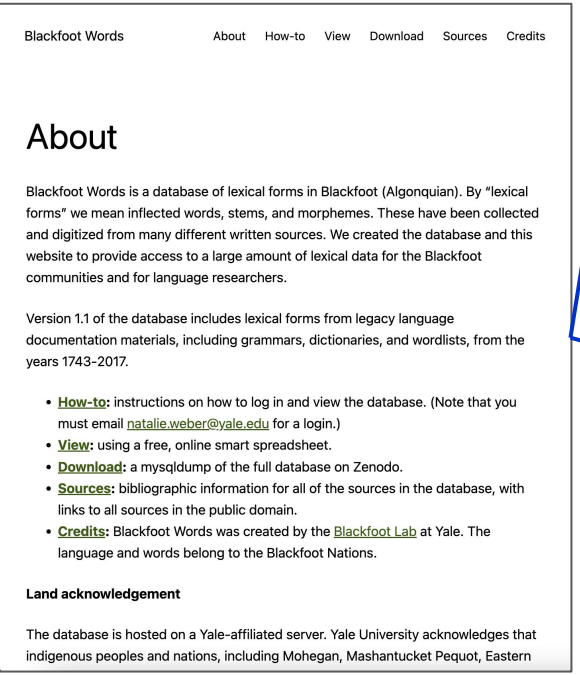
05
Future Projects

Suitable projects

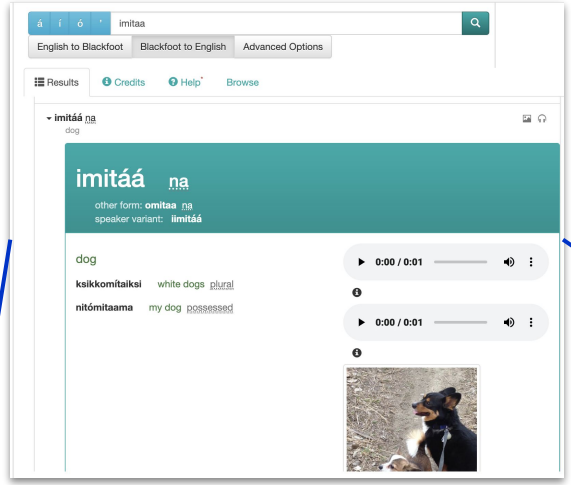
- Morphophonological alternations: across different templatic positions
- Historical linguistics: sources span nearly 300 years
- Dialectal variation: phonetic, phonological, and lexical variation
- Language maintenance and revitalization

Goal: portable digital resource (Bird & Simons, 2003) with multipurpose use for pedagogical and research materials (L'Homme & Cormier 2014)

Future growth

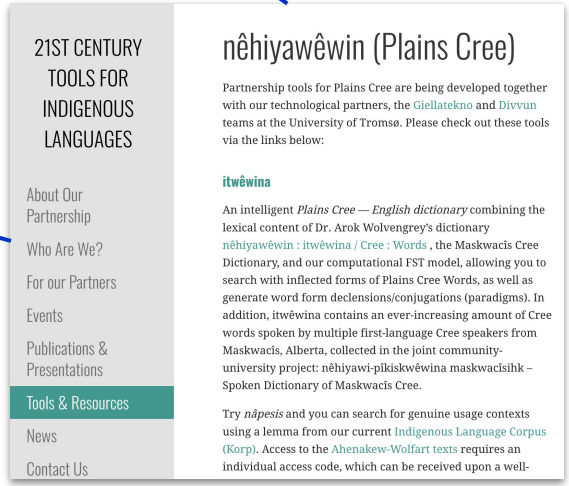


Blackfoot Words



Blackfoot Digital Dictionary
https://blackfoot.algonquianlanguages.ca/
(Genee & Junker 2018)

https://21c.tools/



Summary

Blackfoot Words = a database of tokens and types

- (tokens) tokenizes each word into stems and morphemes, preserving the original source orthography and encoding some hierarchical structure
- (types) lemmatizes at the stem and morpheme levels

	Challenge:		**Solution**:
1.	Stem recursivity	→	Recursive stem table
2.	Multiple templatic positions	→	LabLemmaCategory
3.	Lexical categories	→	Developed categories at 3 levels

Nitsíkohtaahsi’takihpinnaan!

(Thank you!)

References

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3): 557–582.

Bloomfield, Leonard. 1946. Algonquian. In *Linguistic structures of Native America, Hoijer, Harry* (ed.), 85–129. (Publications in Anthropology 6). New York: Viking Fund.

Curtis, Edward S. 1911. *The North American Indian: Vol. 6. The Piegan. The Cheyenne. The Arapaho*. Norwood: The Plimpton Press.

Déchaine, Rose-Marie, & Natalie Weber. 2018. Root syntax: Evidence from Algonquian. In Monica Macaulay, & Meg Noodin (Eds.), *Papers of the Forty-seventh Algonquian Conference*, 57–82. Michigan State University Press.

Frantz, Donald G. 2017. *Blackfoot grammar*. 3rd edn. University of Toronto Press.

Genee, Inge, and Marie-Odile Junker. 2018. The Blackfoot Language Resources and Digital Dictionary project: Creating integrated web resources for language documentation and revitalization. *Language Documentation & Conservation* 12: 298–338.

Goddard, I. (1990). Primary and secondary stem derivation in Algonquian. *International Journal of American Linguistics* 56(4): 449–483. https://doi.org/10.1086/466171

References

L'Homme, Marie-Claude & Monique C. Cormier. 2014. Dictionaries and the Digital Revolution: A Focus on Users and Lexical Databases. *International Journal of Lexicography* 27(4): 331–340. https://doi.org/10.1093/ijl/ecu023

Taylor, Allan. 1969. A Grammar of Blackfoot. Berkeley: University of California, Berkeley Dissertation.

Tims, John W. 1889. *Grammar and dictionary of the Blackfoot language in the dominion of Canada*. London: Society for Promoting Christian Knowledge.

Uhlenbeck, Cornelius C. 1938. *A concise Blackfoot grammar based on material from the Southern Peigans*. (Verhandelingen der Koninklijke Akademie van Wetenschappen, Afdeeling Letterkunde, Nieuwe Reeks, d. 41.) N.V. Noord-Hollandsche Uitgevers-Maatschappij.

Weber, Natalie, Tyler Brown, Joshua Celli, McKenzie Denham, Hailey Dykstra, Nico Kidd, Rodrigo Hernandez-Merlin, Evan Hochstein, Pinyu Hwang, Diana Kulmizev, Hannah Morrison, Matty Norris, and Lena Venkatraman. Blackfoot Words: A lexical database of Blackfoot legacy sources. *Language Resources and Evaluation* 57: 1207–1262. https://doi.org/10.1007/s10579-022-09631-2. Available at https://www.blackfootwords.com/..