Estimating sound similarities

Michael Cysouw, Jelena Prokić, Steve Moran

Sound Similarities

- Manually specified
  (Kondrak 2002; Heeringa 2004)
- Hidden Markov Models
  (Bhargava & Kondrak 2009)
- Regular multi-alignment
  (Prokić 2010; Steiner, Stadler & Cysouw 2011)
- Bayesian inference
  (Prokić 2010)
- Investigating almost identical words
  (Holman, Brown & Wichmann 2011)

Graphemic Normalization

- Widespread idea:
  “Convert everything into IPA”
- IPA is just another orthography !
  (only approximation of sound)
- Still: sound-based normalization is practical
  (but there are strong differences !)
- But: can we do without ?

Estimating sound similarities

Michael Cysouw, Jelena Prokić, Steve Moran

Estimating symbol similarities

Michael Cysouw, Jelena Prokić, Steve Moran

Graphemic parsing

- **Unicode normalization**
  ṍ vs. o ˜ ´

- **Orthographic parse**
  (separate orthographic units as used in the source:“graphemes”)

- **Orthographic normalization**
  (research specific!)

Graphemic parsing

- **Code points** (7) t s ʰ o ̃ ́ ː
- **Characters** (4) t sʰ ṍ ː
- **Graphemes** (2) tsʰ ṍː

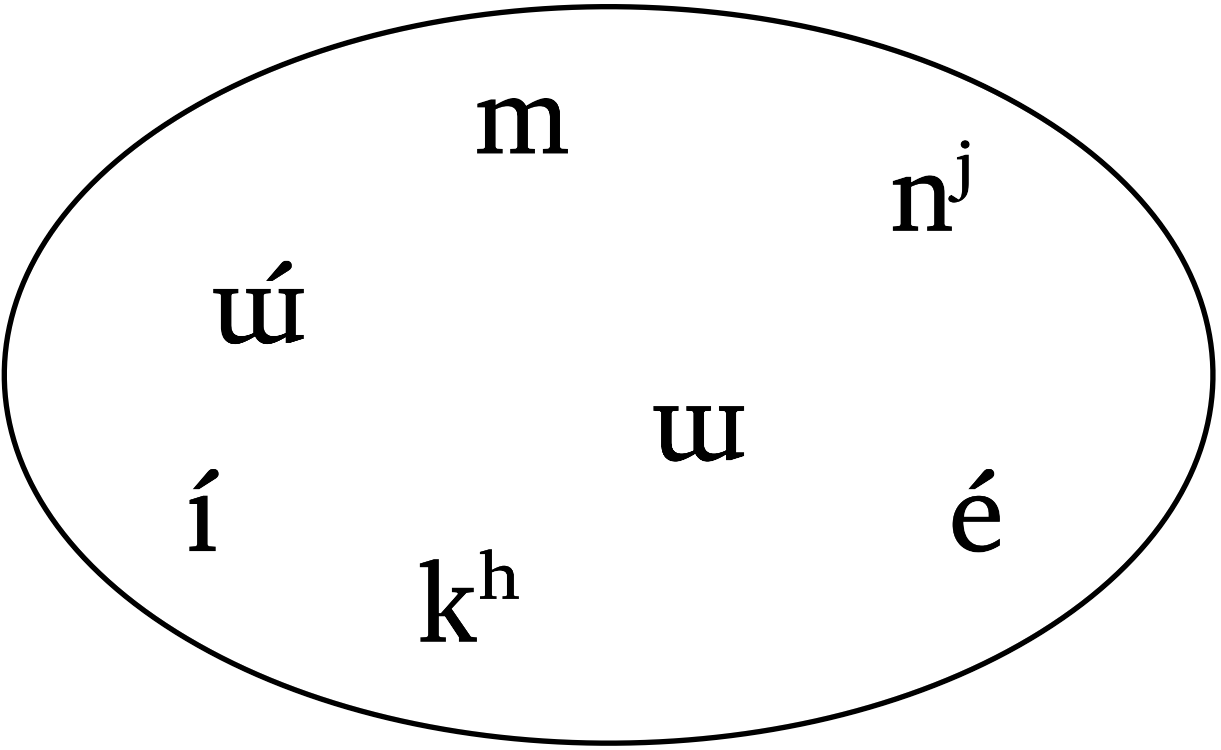
Cross-script mapping

E	R	freq	dice
r	р	184	0.88874745
n	н	115	0.8461936
l	л	104	0.79646295
s	с	114	0.7927922
t	т	165	0.7701921
m	м	47	0.7699933
o	о	184	0.7510106
k	ть	21	0.74458015
p	п	50	0.7388723
i	и	102	0.7034591
a	а	221	0.6866478
u	у	40	0.6449104
c	к	77	0.6251676
e	е	219	0.59066784
b	б	32	0.525643
w	в	46	0.46787763
d	д	42	0.381996
⋮	⋮	⋮	⋮

Cysouw, Michael & Hagen Jung. 2007. Cognate identification and alignment using practical orthographies. Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, 109-116.

‘bag of symbol’’ approach

mín^{j}éék^{h}ɯɯ́



m

n^j

ɯ́

ɯ

í

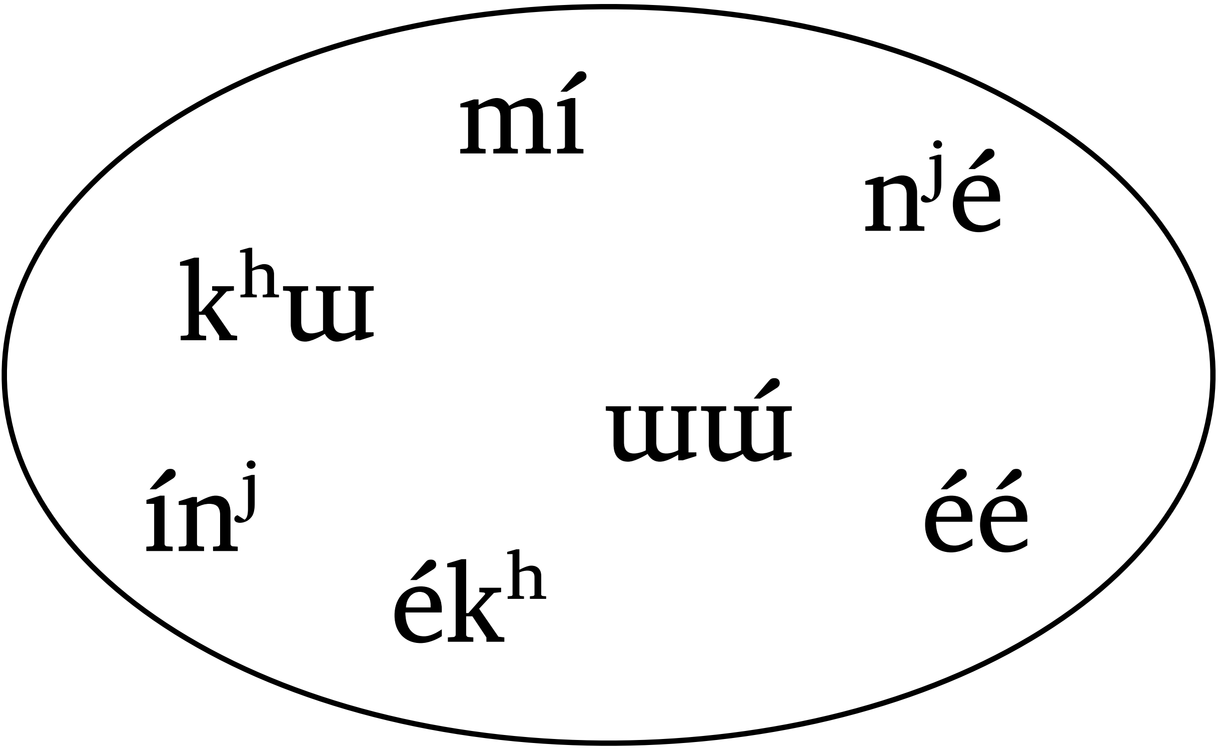
é

k^h

(1-grams)

‘bag of symbol’’ approach

mín^{j}éék^{h}ɯɯ́



mí

n^{j}é

k^{h}ɯ

ɯɯ́

ín^{j}

éé

ék^{h}

(2-grams)

Cross-script mapping

- Ignore linear structure of words
  “bag of symbols” approach
- Use parallel wordlist to estimate co-occurrences of n-grams
- N-grams that have a high probability of co-occurrence in parallel meaning are interested for historical linguistics

Chocó†

DR	hɨwá hã
CT	huwá hã
CM	húa hã
TD	húa hã
EP	húa hã
BA	húa hã
WM	hua hã

Chibcha

IK	gúnnɨ
KO	káu-xəldə
DM	gúla
CL	háttakɾaˀ
TN	átkara
BI	ˀakdurə, ˀaxdurə

Barbacoa

PA	kuse
GU	tasik
TR	'kwal
AW	tʃtɨ
TP	tȩde
CH	tʲaapa

Kamsá

KS	kukuá-tçe

Quechua

IN	máki

Arawak

WY	wahápɨ†
AC	-káahi
CR	no-kapi
PP	wa-káapi
YC	wa-játeˀela
TO	paséwa†
CA	kaapi†
BN	-káapi
RE	-ké

Tucano

TC	õbṍ-kã
WN	wã'bõ bã'kã
PY	õ'bõ̰-kã
WA	ã́'bṍ
BR	ã'bõ
TY	wã'bṍ-pãbã
YR	'wã́bõ
DE	bõhõ'tṍõ
SR	bõ'hṍ
TA	wã'bṍ
CP	wã'bṍ
MA	ãbõ
BS	ãbõ
TM	pita-'ka
CU	pɨ'rɨ́
KG	'hɨ̃tʰɨ
SI	ɨ̃'tɨ sa'da†
SE	'hɨ̃'tɨ
OR	hɨ̃tɨ

Carib

CJ	ee'nʲari
YK	óma

Guahibo

PL	pe-kóbe
GH	pe-kóbe:
CI	pé-kobe, pe-kóbe
JT	pe-kó
GY	peh keˀé

Sáliba-Piaroa

SL	

Macú-Puinave

PU	mo lap
NK	teiˀ^{3}
KK	tejˀ^{2}-ja^{4}
JU	depũj

Witoto

MR	ono-ʤɯ, onoɯ
MN	ónoɯ
NP	ónoɯ
OC	on̥õõ(po)
MU	úse
BO	(mé)-ˀóxtsʰɨɨ́
MÑ	óhtsʰɨɨ́, íˀóhtsʰɨɨ́

	Bora	**Muinane**
down	ʧɨ́nʲe, paári	báari, gíino
bee	íímɯ́ʔóexpʰi, téʔtsʰipa	nɨ́ɨ́bɨri, mɨ́ɨbɨriʔɨ
sharp	tsʰɯ́ʔxɨβáne	sɨ́ɨ́xéβano
…	…	…

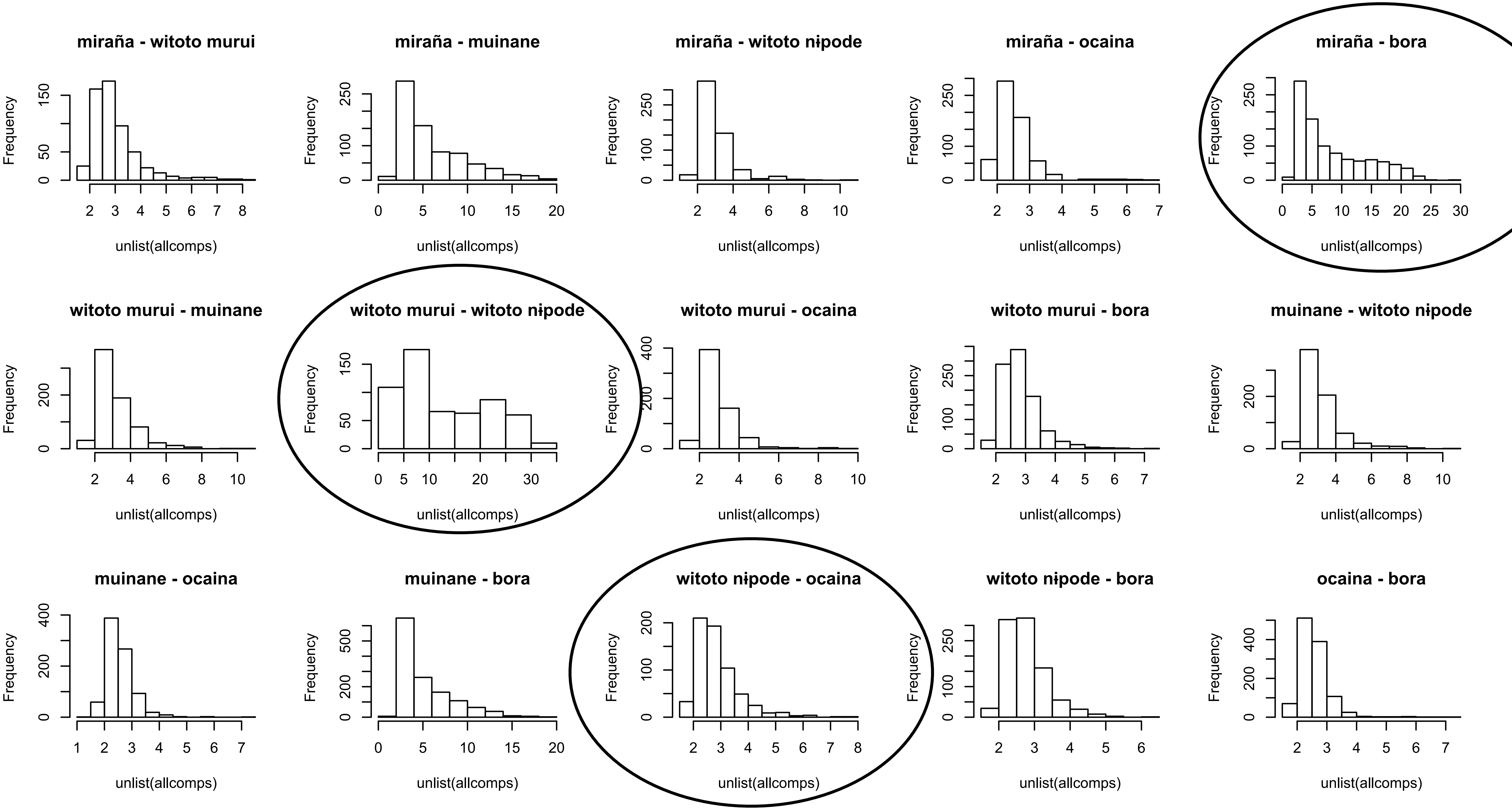
Bora	Muinane	Bora	Muinane	Bora	Muinane
#k	#k^{h}	#i	#i	#n	#n
kɨ	k^{h}ɯ	#a	#a	#m	#m
se	ts^{h}ɨ	di	ti	mɨ	mɯ
xe	xɨ	du	to	nɨ	nɯ
ga	k^{w}a	#d	#t	us	ts^{h}ɨ
ba	pa	#s	#ts^{h}	#t	#t^{h}
#b	#p	gi	ʧi	ɨg	ɯk^{w}
e#	ɨ#	ni	ni	#ɸ	#p^{h}

Using bigram matching

- Bora ‘two’: mín^{j}éék^{h}ɯɯ́
- Muinane ‘two’: mííηokɨ

	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m	22	3	2	2	2	2	2	2
mi	4	12	2	2	5	1	1	1
in^{j}	2	1	5	9	3	1	1	2
n^{j}e	1	1	5	5	4	1	1	2
ee	3	3	3	3	6	2	2	2
ek^{h}	1	2	1	1	4	2	3	2
k^{h}ɯ	2	2	2	2	2	1	23	2
ɯɯ	2	2	3	3	2	2	4	4
ɯ#	2	2	3	2	3	1	3	4

	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m	22	3	2	2	2	2	2	2
mi	4	12	2	2	5	1	1	1
in^{j}	2	1	5	9	3	1	1	2
n^{j}e	1	1	5	5	4	1	1	2
ee	3	3	3	3	6	2	2	2
ek^{h}	1	2	1	1	4	2	3	2
k^{h}ɯ	2	2	2	2	2	1	23	2
ɯɯ	2	2	3	3	2	2	4	4
ɯ#	2	2	3	2	3	1	3	4



miraña - witoto murui
miraña - muinane
miraña - witoto nɨpode
miraña - ocaina
miraña - bora
witoto murui - muinane
witoto murui - witoto nɨpode
witoto murui - ocaina
witoto murui - bora
muinane - witoto nɨpode
muinane - ocaina
muinane - bora
witoto nɨpode - ocaina
witoto nɨpode - bora
ocaina - bora
Frequency
unlist(allcomps)

	Ocaina	Witoto Murui
HAND	on̥õõ	onoʤɯ
WE	xoxo	koko
HOUSE	ɸoo	ɸo
DOG	hõʔxo	hɯko
JAGUAR	hõʔxo	hɯko
FATHER	mõõ	moo
HUMMINGBIRD	ɸaʔtííʔtʲo	ɸiθido
TREE	am̥ɯ̃ɯ̃ɲ̊a	amena
STICK	am̥ɯ̃ɯ̃ɲ̊a	amena
WHO	bṍ	bu
SLEEP	ɯ́ɯ́nõ	ɯnɯ
AGOUTI	ɸɯ́ɯ́tʲo	ɸɯdo
THIS	bĩ́ĩ́	bie
THIS	baʔi	bie
NAME	maam̥ɯ	mamekɯ
DAY	mooɲ̊a	aremona
BOW	tsipóxatʲa	θɯkuira
HEAR	xaaxa	kakade
DAY	moɲ̊amó	aremona

	Ocaina	Witoto Murui
GREASE	ɸahĩ́ĩ́	ɸare
YOU (PLURAL)	mõʔ	omoɯ
THIS	bɯ	bie
ARROW	oɯdʲáátʲa	dukɯraθɯ
SPEAR	oɯdʲáátʲa	dukɯraθɯ
LIP	ɸaʔóóʔko	ɸue igoɯ
GREEN	moxóóso	mokorede
I	xõ	kue
ONE	tʲa	dahe
WE	xo	kaɯ
TOOTH	aʔtiiʔtʲo	iθido
MOUTH	ɸooɯ	ɸue
BELLY	gááho	hebe
FATHER	mõõhõ	moo
YOU (PLURAL)	mõʔxo	omɯko
SWAMP	xonɯ́ɯ́βaga	kɯnere
RAT	mɯɲṍṍko	miɲɯe
PATH, TRAIL	naahõ	naɯθo
OWL	mṍṍɲ̊õhõ	monuiθɯ