

Presenter: Jennifer Proctor

AI for Repair of Multilingual Records incorporating compound transliteration and diacritic restoration



an Education and Data Science powered relationship



WorldCat®



The Problem:

In Micro

THE MUSIC IN THE BOX

Language complicates search

24

43. Юмореска

В. ГЕВИКСМАН

Не очень быстро, шутливо, грациозно

The image shows a page of musical notation for a piece titled '43. Юмореска' by V. Geyksman. The page number '24' is in the top left corner. The tempo and mood instructions are 'Не очень быстро, шутливо, грациозно'. The music is written in bass clef with a key signature of one flat (B-flat) and a time signature of 2/4. The first staff begins with a dynamic marking 'p' and a fingering '2'. The second staff ends with a first ending bracket labeled '1' and a '4' indicating the number of measures. The notation includes various rhythmic values, slurs, and articulation marks.

No results match your search for 'юмореска гевиксман'.
Having trouble? Help us improve our search and [let us know](#) what you're trying to find.
Or, [save this search](#) for future reference.

Search Library Items Lists Contacts Libraries

Search for an item in libraries near you:

Enter a title, subject or author: юмореска гевиксман

Search [Advanced Search](#)

Home Search Create lists, bibliographies and reviews: [Sign in](#) or [create a free account](#)

WorldCat юмореска гевиксман **Search**

[Advanced Search](#) [Find a Library](#)

Search results for 'юмореска гевиксман'

Results 1-1 of about 1 (.09 seconds) << First < Prev 1 Next >

[Select All](#) [Clear All](#) Save to: [New List] Save Sort by: Relevance Save Search

1.  [Dve p'esy dliā fagota i fortepiano](#)
by Vitalii Geviksman
 Musical score
Language: No Linguistic Content
Publisher: Moskva : Gos. muzykal'noe izd-vo, 1950.

Results 1-1 of about 1 (.09 seconds)

Publication: Moskva : Gos. muzykal'noe izd-vo, 1950.
OCLC Number/Unique Identifier: 14100617
Language: No linguistic content
Physical Description: 1 score (9 pages) + 1 part (3 pages) ; 29 cm
Number of Musical Instruments or Voices: Woodwinds - Bassoon (01) Keyboard - Piano (01)
Contents: 1. Melodīia-- 2. **Юмореска**.
Subjects: [Library of Congress Subject Headings](#)
[Bassoon and piano music Scores and parts.](#)
[Show More](#) ▾
Notes (General): Cover title.
Romanized record.
Label on cover: Sole selling agent, Leeds music, New York.
Staff View (MARC Record): [MARC Record](#)
Genre: Scores and parts.
Partitions et parties.

Access Options

Request from Libraries Worldwide

Request Item through Interlibrary Loan

*But wouldn't the Cyrillic work if the record was just complete?

A Problem with Many Roots

Ю	ĪŪ	ю	īū
Я	ĪĀ	я	īā
Ө	F	ө	f
У	Ÿ	у	ÿ
Ж	Zh	ж	zh
З	Z	з	z
И	I	и	i
І	Ī	і	ī
Й	Ī	й	ī
К	K	к	k
Л	L	л	l
М	M	м	m
Н	N	н	n
О	O	о	o
П	P	п	p
Р	R	р	r
С	S	с	s
Т	T	т	t
У	U	у	u
Ф	F	ф	f
Х	Kh	х	kh
Ц	ĪS	ц	īś
Ч	Ch	ч	ch
Ш	Sh	ш	sh
Щ	Shch	щ	shch
Ъ	" (hard sign)	ъ	" (hard sign)
Ы	Y	ы	y
Ь	' (soft sign)	ь	' (soft sign)
Ъ	ĪĒ	ъ	īē
Э	Ē	э	ē

Human

Factors

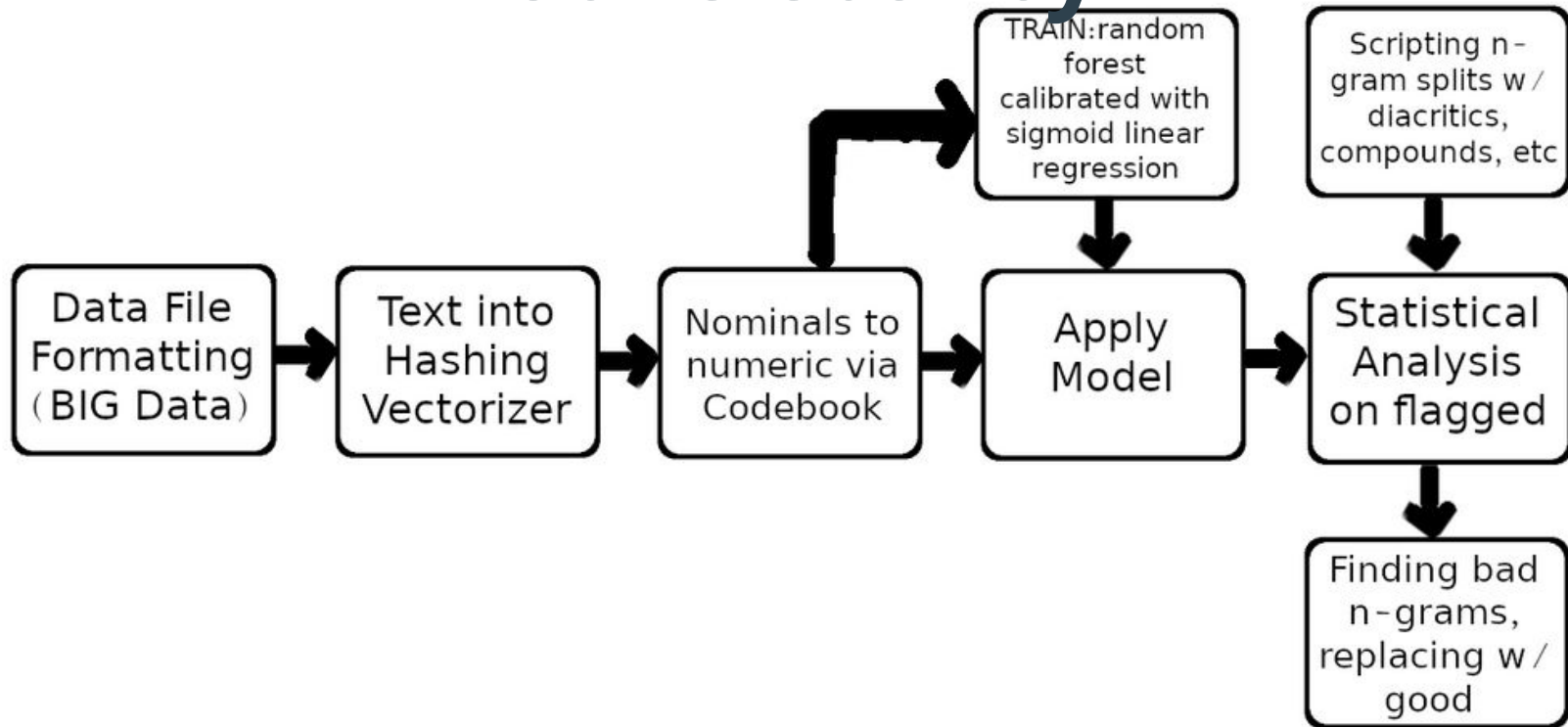
- Different priorities - ex: Russian libraries vs cataloguers who know no Cyrillic
- Different institutional and national norms leading to inconsistent data entry

Technology

- Inconsistent retention of diacritics, non-Roman characters, and stress marks
- Programs misinterpret characters like stress marks
- Transliteration is not 1-to-1

Macro: Modeling for

Consistency



Pre-processing, data transformation, Machine Learning, and Statistical Analysis were combined so that data (& search results) will be more predictable in future

TRAINING SUBSET

4.5 million

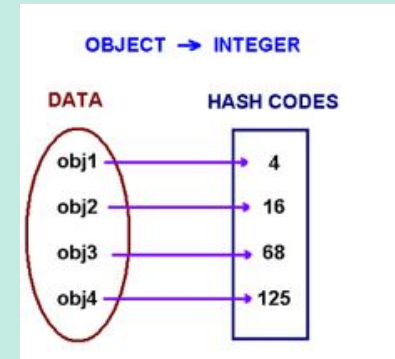
records labeled (OpenRefine, Python)

Text into Numbers

Computers only speak Math

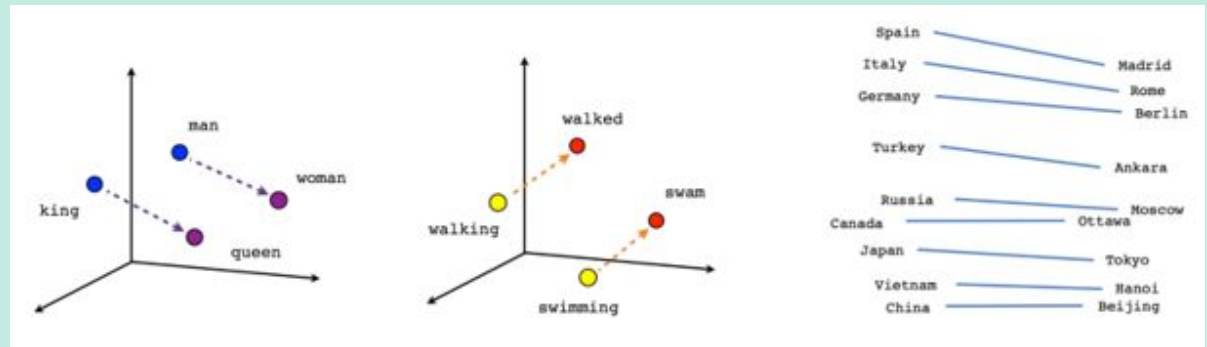
■ HASHING

Letters and words become numbers



■ WORD VECTORS

What specific number is influenced by relationships between words



■ NOMINALIZATION

Encoding Text Categories as Unique Numbers with a code book

	A	B
6	"rus"	4
7	"bul"	5
8	"it"	6
9	"er"	7
10	"nju"	8
11	"rb"	9
12	"oru"	10
13	"aa"	11
14	"quc"	12
15	"ch"	13
16	"flu"	14
17	"nke"	15

The Model

SUCCESS WITH

Random Forest with Linear
Regression Calibration

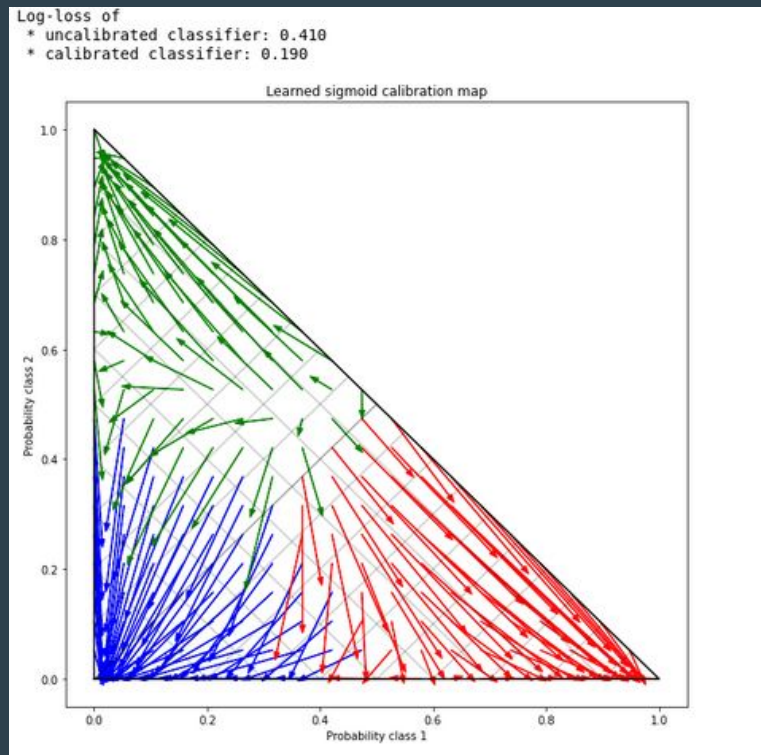


Table 2: Measurement of Model Success - ROC AUC

Macro 0.958257

Weighted 0.947674

Table 3: Measurement of Model Success - Confusion Matrix

	Probability of omissions or errors	Low	Medium	High
Low		363547	1062	1322
Medium		1187	40963	61
High		21973*	585	70172

While dozens of ML algorithms tested and scalability options were explored (b/c cannot allocate 33 TB RAM errors), ultimately the horizontal scalability issue made

Error Catching

And Correcting!

■ NGRAMS

Breaking apart transliterated text into ngrams was way more complicated than normal because 1 Cyrillic letter could equal as much as 6 or as few as 1 transliterated character (which may or may not be a letter)

Actual

	A	B
1	g4grams	g4gramsFreq
2	chesk	0.267
3	iches	0.257
4	ogg	0.256
5	nija	0.196
6	koj	0.189
7	skoj	0.175
8	pro	0.169
9	isto	0.159
10	noj	0.156
11	ist	0.149
12	stor	0.148
13	tori	0.145
14	nie	0.142
15	kij	0.138
16	ross	0.134

■ FREQUENCIES

Presence of common typos can help QCers hone in on error locations

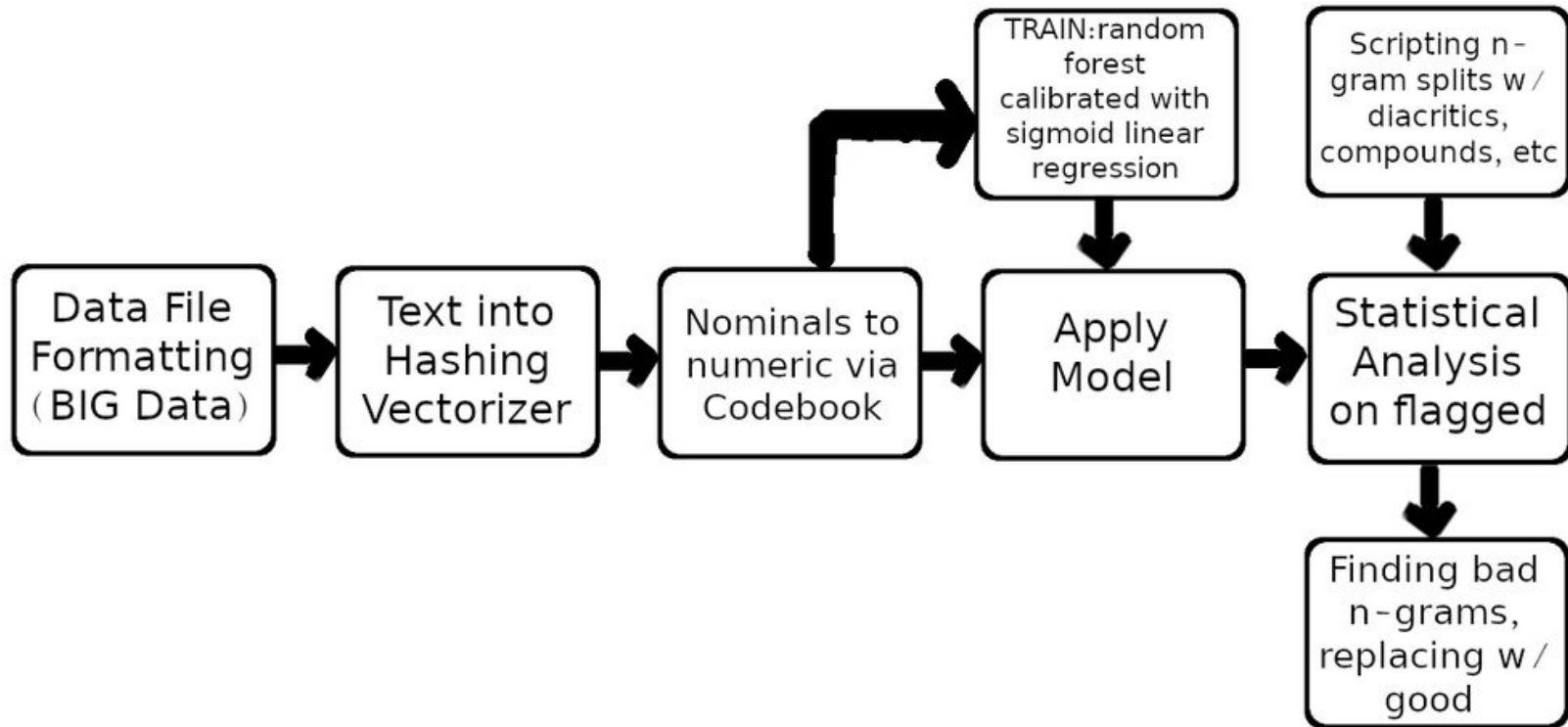
Typos

	A	B
1	b4gram	b4gramsFreq
2	al'n	0.037
3	l'no	0.026
4	tel'	0.023
5	l'ny	0.017
6	ul't	0.017
7	kul'	0.016
8	st'	0.016
9	tur	0.015
10	l'tu	0.015
11	el'n	0.012
12	el's	0.01
13	ost'	0.01
14	nal'	0.009
15	l'sk	0.009

■ SUBSTITUTIONS

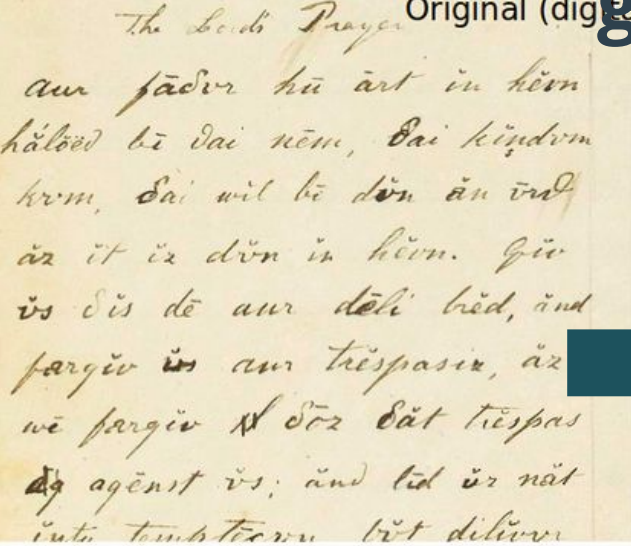
By matching some common typos to their best match in the actual Russian language frequencies, some replacements can be made automatically and generally once this process is done, a good context-aware Russian spellchecking tool is able to sort out the rest

Final Project Pipeline




A Shadow Problem: Where did the of Diacritics go?

Original (digitized)

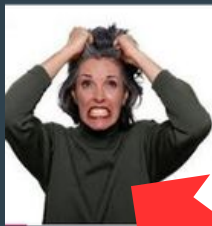


Carefully transcribed by expert volunteers to preserve meaning



Fate	fæt
far	fɔ:r
fall	fɔ:l
fat	fæt
me	mɪ
met	mɛt
pine	pɪn
pin	pɪn
no	nɔ
move	mʊv
nor	nɔ:r
not	nɒt
tube	tju:b
tub	tʌb
bull	bʊl
thin	θɪn
this	ðɪs
oil	ɔɪl
house	haʊs

The Lord's Prayer
Our fɔ:dʊr hʊ ɔ:st in hɛvn
hæ'ləəd bɪ θɔɪ nɛm, θɔɪ kɪŋdɒm
kɒm, θɔɪ wɪl bɪ dʌn ʌn ɛrə əs it ɪz dʌn in hɛvn. Gɪv ʌs ðɪs dʌs ɔ:ur dɛli brɛd, and
fɔ:rgɪv ʌs ɔ:ur trɛspɔ:sɪz, əz wɛ fɔ:rgɪv ~~ðɪs~~ ðɔ:z ðæt trɛspɔ:
~~ag~~ ʌg ~~g~~ŋst s; ʌnd lɪd ʊz nʌt ɪntu
tempɔ:ʃn, bʌt dɪlɪvɔ: ʌs frəm ɪvl. fɔ: ʌn ɪz ðɪ kɪŋdɒm and θɪ pɔ:ʊɔ: and ðɪ glɔ:ri.
fɔ: ɛvɔ: ʌnd ɛvɔ:. Əmɛn



Post-processing Gibberish

The Lord's Prayer
ur f:ɔ:ɹ h :rt in hvn
hæ'lɛd b i nm, θi kdm
km, θi wl b dn n ɔ:r æs t z dn n hvn. Gv s θs d ur dli brd, ænd frgv s aur
trpsiz, æz w frgv ~~S~~ ðz ðæt trps
~~ag~~ gŋst s; ænd ld z nt ntu temptn, bt dilvr
s frm vl, fr ðain z θi kdm ænd thi p'ur ænd θi glri, fr vr nd vr. mn

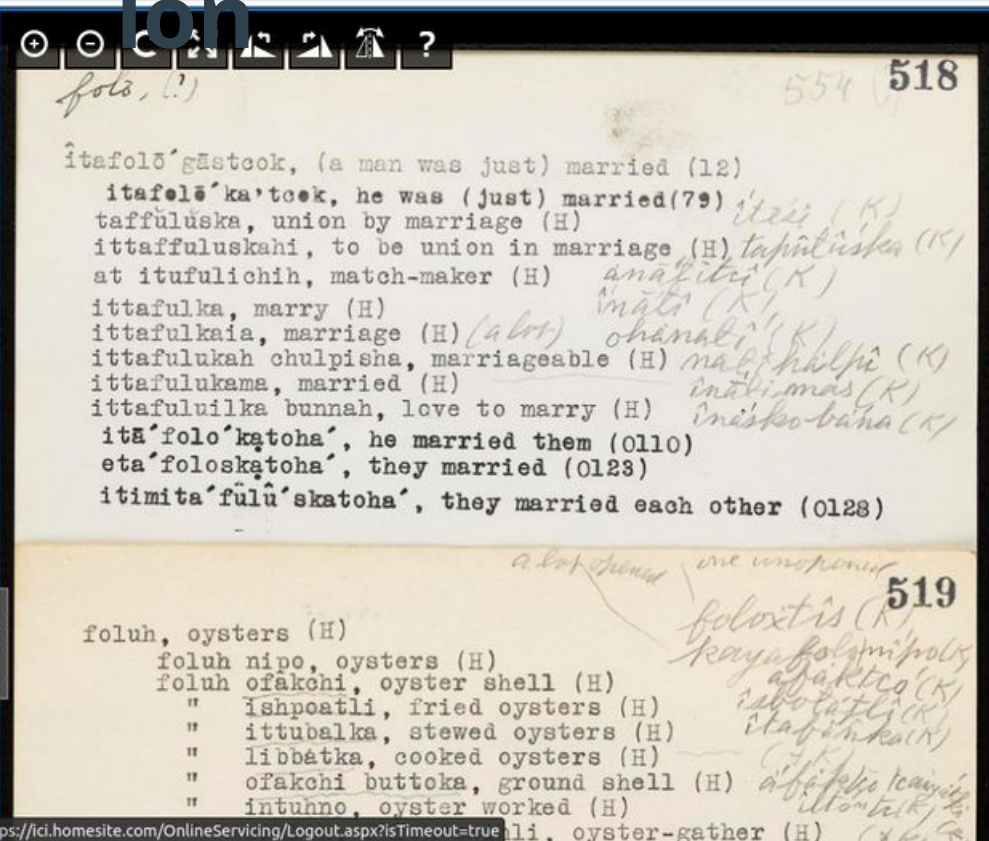
Hidden deep inside the software we use there is a set of assumptions - valuations - about what languages, and therefore what letters, should be included. This goes far beyond ensuring you can understand the GUI - it impacts what data is saved and how, and what happens to data that is NOT saved

The Importance of

Representation

Today, the Alabama Language is spoken by less than 100 people, largely on the Alabama-Coushatta Reservation in Texas

This Alabama-English dictionary contains meanings and pronunciations of tens of thousands of words and phrases digitized to make them more accessible to Indigenous Language Learners. Unfortunately, export from the transcription software erases many of



Navigation icons: back, forward, home, search, page #, print, share (Facebook, Twitter).

i This transcription has been completed. [Contact us with corrections.](#)

foto, (?)	554 [handwritten]	518 [typed]
itafolo' gastcok, (a man was just) married (12)		
itafolo'ka'tcok, he was (just) married (79)		
taffūlūska, union by marriage (H)		^[[i'tēsi(K)]]
ittaffuluskahi, to be union in marriage (H)		^[[tapūtūška (K)]]
at itufulichih, match-maker (H)		^[[ānā'lītci(K)]]
ittafulka, marry (H)		^[[i'nā'ti (K)]]
ttafulukah chulpisha, marriageable (H)		^[[natihālpī (K)]]
ittafulukama, married (H)		^[[inātimas (K)]]
ittafuluklka bunnah, love to marry (H)		^[[ināskobāña (K)]]
itā'folo'katoha', he married them (0110)		
eta'foloskatoha', they married (0123)		
itimita'fūlū' skatoha', they married each other (0128)		
^[[a lot opened one unopened]]		
foluh, oysters (H)		519 [typed]
foluh nipo, oysters (H)		^[[foloxtis (K)]]
foluh ofakchi, oyster shell (H)		^[[kayafolopsi'po (K)]]
" ishpoatli, fried oysters (H)		^[[afaktco' (K)]]
" ittubalka, stewed oysters (H)		^[[isbotátli (K)]]
" libbátka, cooked oysters (H)		^[[itafánka (K)]]
" ofakchi buttoka, ground shell (H)		^[[ā'fā'kālgaiv
" intuhno, oyster worked (H)		
" foluh, oyster-gather (H)		