A Python Toolkit for Universal Transliteration

Ting Qian[1], Kristy Hollingshead[2], Su-youn Yoon[3],
Kyoung-young Kim[4], Richard Sproat[5]

University of Rochester[1], OHSU[2], ETS[3], UIUC[4], OHSU[5]
ting.qian@rochester.edu[1], hollingk@cslu.ogi.edu[2], syoon9@gmail.com[3],
kkim36@illinois.edu[4], rws@xoba.com[5]

LREC, Malta
May 21, 2010

Transliteration Examples from the Web

- ಕ್ಯಾಮರು Guru гуру ગુરુ ഗുരു ગુરૂ
- 카메라 камеру камеры Камер Camera Камера カメラ
- мелия Melia Μελία Мелиа მელია Мелия मेलिया
- Хотэль хотел Хотел Хоутел Hotel отель
- Рома роми ロウマ Roma Ρώμα ΡΩΜΗ
- Ванила ヴァニラ วานิลา Ванили Vanilla Ванилия バニラ वनिला
- Карма カルマ Karma 카르마 Кармы
- 브라운 Brown БРАУН Брайан Браун
- ტაიმი タイム Тайм TIME
- Сунил సునీల్ సునిల్ Sunil सुनिल

Basic Issues

- Cooccurrence - e.g. temporal correlation:
  - In parallel/comparable corpora we expect related concepts/terms to have similar distributions over space and time
- Edit distance:
  - Phonetic similarity
  - Graphical similarity
- Our goal: techniques for extracting plausible transliteration candidates for comparable corpora in n-tuples of languages that use different scripts.

Previous Work

- Transliteration: Knight & Graehl 1998; Meng et al. 2001; Gao et al. 2004; inter alia.
- Comparable corpora: Fung, 1995; Rapp 1995; Tanaka and Iwasaki, 1996; Franz et al.,1998; Ballesteros and Croft, 1998; Masuichi et al., 2000; Sadat et al., 2003; Tao and Zhai, 2005.
- Mining transliterations from multilingual web pages: Zhang & Vines, 2004
- Sproat, Tao & Zhai, ACL 2006:
  - Trained phonetic distance, similarity in temporal distribution

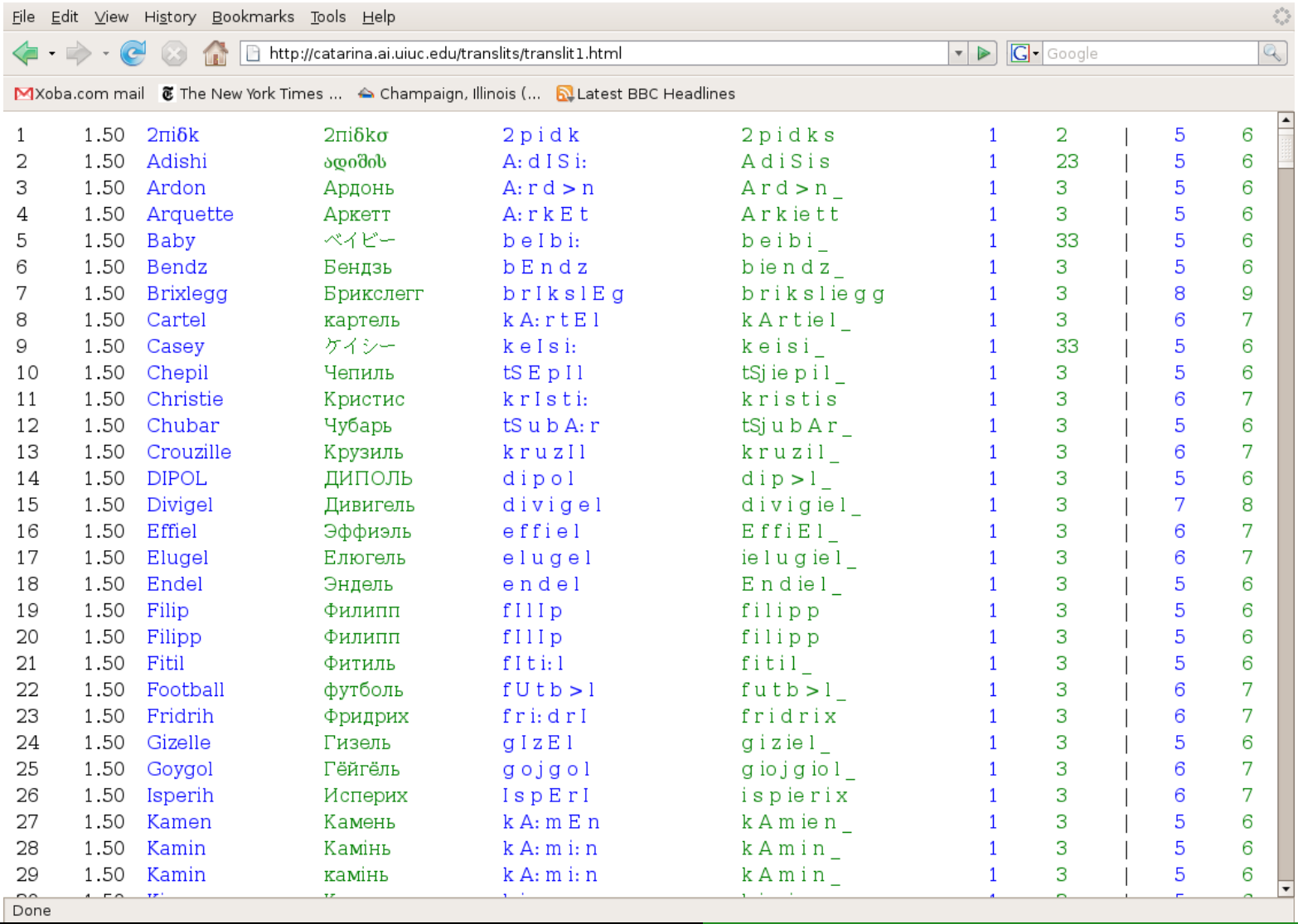
Previous Work

- Klementiev and Roth:
  - Discriminative model using letter n-gram features, and temporal distribution
- Tao et al, EMNLP 2006:
  - Untrained phonetic model and temporal distribution
- Yoon, Kim and Sproat, ACL 2007:
  - Untrained vs. discriminatively trained phonetic models
    - Unitran: Provides pronunciations for scripts in Basic Multilingual Plane
    - Hand-built phonetic model uses phonetic features as well as "pseudofeatures" derived from second-language learner errors
- Recent NEWS 2009 workshop (colocated with ACL in Singapore) highlighted a number of approaches to transliteration

Web Transliterations using Unitran/Handbuilt Distance Model

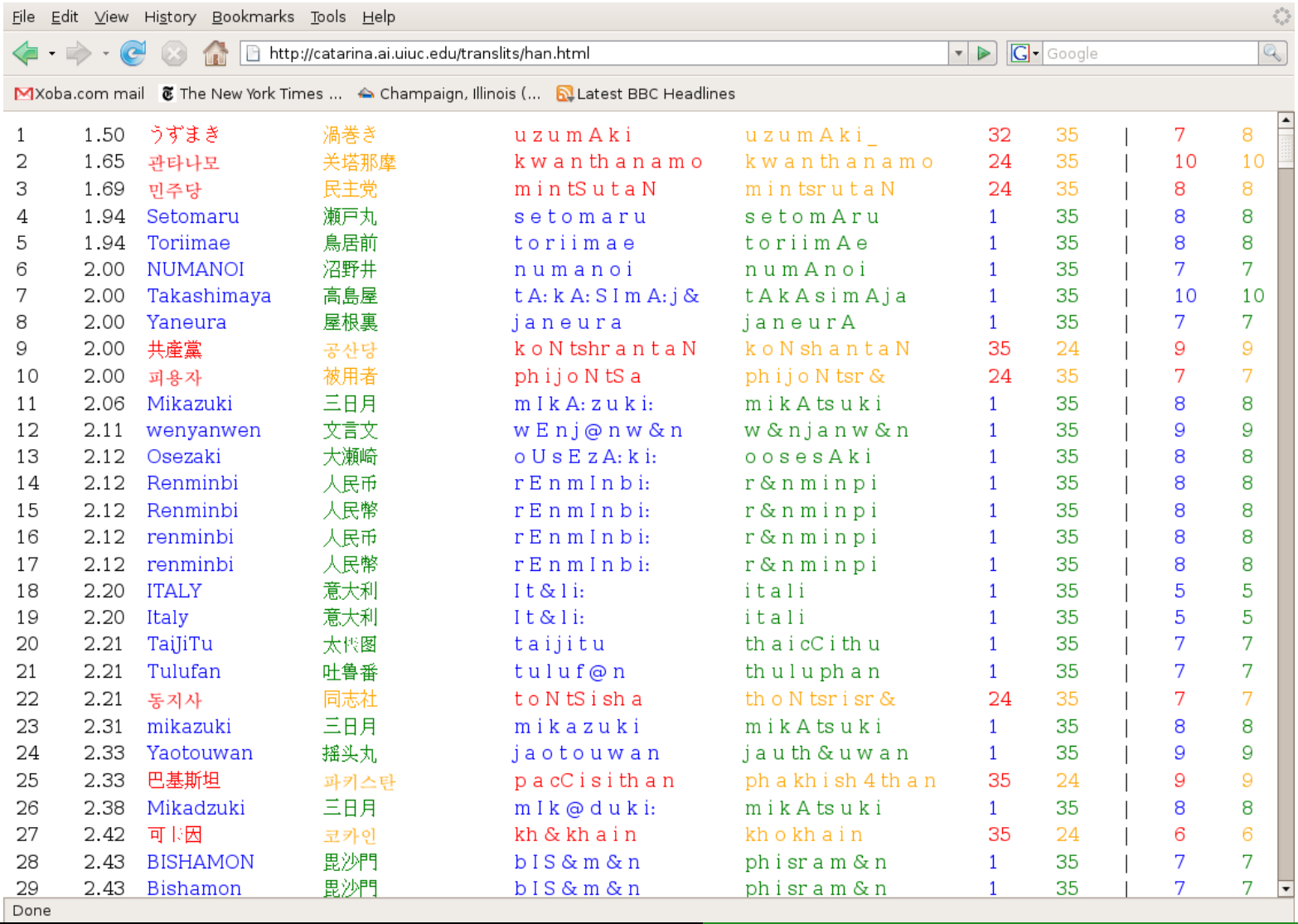
- Find patterns of form $x_i x_{i+1} x_{i+2} \ldots (y_i y_{i+1} y_{i+2} \ldots)$ where at least some of $y_i y_{i+1} y_{i+2}$ are in a script different from $x_i x_{i+1} x_{i+2}$
- Use Unitran to guess pronunciations for most strings:
- Festival for "English"
- Special tables for:
  - Chinese (Mandarin)
  - Kanji (kunyomi)
  - Extended Latin-1
- Rank by (untrained) phonetic edit distance

Web Transliterations using Unitran/Handbuilt Distance Model



1	1.50	2пiδk	2пiδkσ	2 p i d k	2 p i d k s	1	2	5	6
2	1.50	Adishi	ადიშის	A: d I S i:	A d i S i s	1	23	5	6
3	1.50	Ardon	Ардонь	A: r d > n	A r d > n _	1	3	5	6
4	1.50	Arquette	Аркетт	A: r k E t	A r k i e t t	1	3	5	6
5	1.50	Baby	ベイビー	b e I b i:	b e i b i _	1	33	5	6
6	1.50	Bendz	Бендзь	b E n d z	b i e n d z _	1	3	5	6
7	1.50	Brixlegg	Брикслегг	b r I k s l E g	b r i k s l i e g g	1	3	8	9
8	1.50	Cartel	картель	k A: r t E l	k A r t i e l _	1	3	6	7
9	1.50	Casey	ケイシー	k e I s i:	k e i s i _	1	33	5	6
10	1.50	Chepil	Чепиль	tS E p I l	tSj i e p i l _	1	3	5	6
11	1.50	Christie	Кристис	k r I s t i:	k r i s t i s	1	3	6	7
12	1.50	Chubar	Чубарь	tS u b A: r	tSj u b A r _	1	3	5	6
13	1.50	Crouzille	Крузиль	k r u z I l	k r u z i l _	1	3	6	7
14	1.50	DIPOL	ДИПОЛЬ	d i p o l	d i p > l _	1	3	5	6
15	1.50	Divigel	Дивигель	d i v i g e l	d i v i g i e l _	1	3	7	8
16	1.50	Effiel	Эффиэль	e f f i e l	E f f i E l _	1	3	6	7
17	1.50	Elugel	Елюгель	e l u g e l	i e l u g i e l _	1	3	6	7
18	1.50	Endel	Эндель	e n d e l	E n d i e l _	1	3	5	6
19	1.50	Filip	Филипп	f I l I p	f i l i p p	1	3	5	6
20	1.50	Filipp	Филипп	f I l I p	f i l i p p	1	3	5	6
21	1.50	Fitil	Фитиль	f I t i: l	f i t i l _	1	3	5	6
22	1.50	Football	футболь	f U t b > l	f u t b > l _	1	3	6	7
23	1.50	Fridrih	Фридрих	f r i: d r I	f r i d r i x	1	3	6	7
24	1.50	Gizelle	Гизель	g I z E l	g i z i e l _	1	3	5	6
25	1.50	Goygol	Гёйгёль	g o j g o l	g i o j g i o l _	1	3	6	7
26	1.50	Isperih	Исперих	I s p E r I	i s p i e r i x	1	3	6	7
27	1.50	Kamen	Камень	k A: m E n	k A m i e n _	1	3	5	6
28	1.50	Kamin	Камінь	k A: m i: n	k A m i n _	1	3	5	6
29	1.50	Kamin	камінь	k A: m i: n	k A m i n _	1	3	5	6

Web Transliterations using Unitran/Handbuilt Distance Model



File Edit View History Bookmarks Tools Help

http://catarina.ai.uiuc.edu/translits/han.html

Xoba.com mail | The New York Times ... | Champaign, Illinois (... | Latest BBC Headlines

1	1.50	うずまき	渦巻き	u z u m A k i	u z u m A k i _	32	35	7	8
2	1.65	관타나모	关塔那摩	k w a n t h a n a m o	k w a n t h a n a m o	24	35	10	10
3	1.69	민주당	民主党	m i n t S u t a N	m i n ts r u t a N	24	35	8	8
4	1.94	Setomaru	瀬戸丸	s e t o m a r u	s e t o m A r u	1	35	8	8
5	1.94	Toriimae	鳥居前	t o r i i m a e	t o r i i m A e	1	35	8	8
6	2.00	NUMANOI	沼野井	n u m a n o i	n u m A n o i	1	35	7	7
7	2.00	Takashimaya	高島屋	t A: k A: S I m A: j &	t A k A s i m A j a	1	35	10	10
8	2.00	Yaneura	屋根裏	j a n e u r a	j a n e u r A	1	35	7	7
9	2.00	共産黨	공산당	k o N ts h r a n t a N	k o N sh a n t a N	35	24	9	9
10	2.00	피용자	被用者	p h i j o N t S a	p h i j o N ts r &	24	35	7	7
11	2.06	Mikazuki	三日月	m I k A: z u k i:	m i k A ts u k i	1	35	8	8
12	2.11	wenyanwen	文言文	w E n j @ n w & n	w & n j a n w & n	1	35	9	9
13	2.12	Osezaki	大瀬崎	o U s E z A: k i:	o o s e s A k i	1	35	8	8
14	2.12	Renminbi	人民币	r E n m I n b i:	r & n m i n p i	1	35	8	8
15	2.12	Renminbi	人民幣	r E n m I n b i:	r & n m i n p i	1	35	8	8
16	2.12	renminbi	人民币	r E n m I n b i:	r & n m i n p i	1	35	8	8
17	2.12	renminbi	人民幣	r E n m I n b i:	r & n m i n p i	1	35	8	8
18	2.20	ITALY	意大利	I t & l i:	i t a l i	1	35	5	5
19	2.20	Italy	意大利	I t & l i:	i t a l i	1	35	5	5
20	2.21	TaiJiTu	太[illegible]图	t a i j i t u	th a i cC i th u	1	35	7	7
21	2.21	Tulufan	吐鲁番	t u l u f @ n	th u l u ph a n	1	35	7	7
22	2.21	동지사	同志社	t o N tS i s h a	th o N ts r i sr &	24	35	7	7
23	2.31	mikazuki	三日月	m i k a z u k i	m i k A ts u k i	1	35	8	8
24	2.33	Yaotouwan	摇头丸	j a o t o u w a n	j a u th & u w a n	1	35	9	9
25	2.33	巴基斯坦	파키스탄	p a cC i s i th a n	ph a kh i sh 4 th a n	35	24	9	9
26	2.38	Mikadzuki	三日月	m I k @ d u k i:	m i k A ts u k i	1	35	8	8
27	2.42	可[illegible]因	코카인	kh & kh a i n	kh o kh a i n	35	24	6	6
28	2.43	BISHAMON	毘沙門	b I S & m & n	ph i sr a m & n	1	35	7	7
29	2.43	Bishamon	毘沙門	b I S & m & n	ph i sr a m & n	1	35	7	7

Done

Web Transliterations using Unitran/Handbuilt Distance Model

66	2.80	Fritz	فريتس	f r I t s	f r j t s	1	6	|	5	5
67	2.80	MRirt	مريرت	m r i r t	m r j r t	1	6	|	5	5
68	2.80	ويلمز	Wilms	w j l m z	w I l m z	6	1	|	5	5
69	2.80	كلينت	Clint	k l j n t	k l I n t	6	1	|	5	5
70	2.81	زلتوّن	zilatuun	D I l E t w U n	z i l a t u u n	6	1	|	8	8
71	2.88	bird	بؤرد	b & r d	b o r d	1	6	|	4	4
72	2.88	black	بلاك	b l @ k	b l & k	1	6	|	4	4
73	2.88	mustt	مَست	m ^ s t	m E s t	1	6	|	4	4
74	2.88	مرج	Merge	m I r dZ	m & r dZ	6	1	|	4	4
75	2.90	shogol	شُغّل	S > g & l	S U G a~ l	1	6	|	5	5
76	2.90	مُسلِمِينَ	Muslimiyna	m U s l I m I j n E	m ^ z l I m I I n &	6	1	|	10	10
77	2.92	FRIENDS	فريندز	f r E n d z	f r j n d z	1	6	|	6	6
78	2.92	Friends	فريندز	f r E n d z	f r j n d z	1	6	|	6	6
79	2.92	Surendar	سريندر	s r E n d r	s r j n d r	1	6	|	6	6
80	2.92	فريندز	Friends	f r j n d z	f r E n d z	6	1	|	6	6
81	2.94	فهلميهكي	Välimäki	v h l j m h k j	V _ l i m _ k i	6	1	|	8	8
82	3.00	Gül	گل	G _ l	g _ l	1	6	|	3	3

Web Transliterations using Unitran/Handbuilt Distance Model

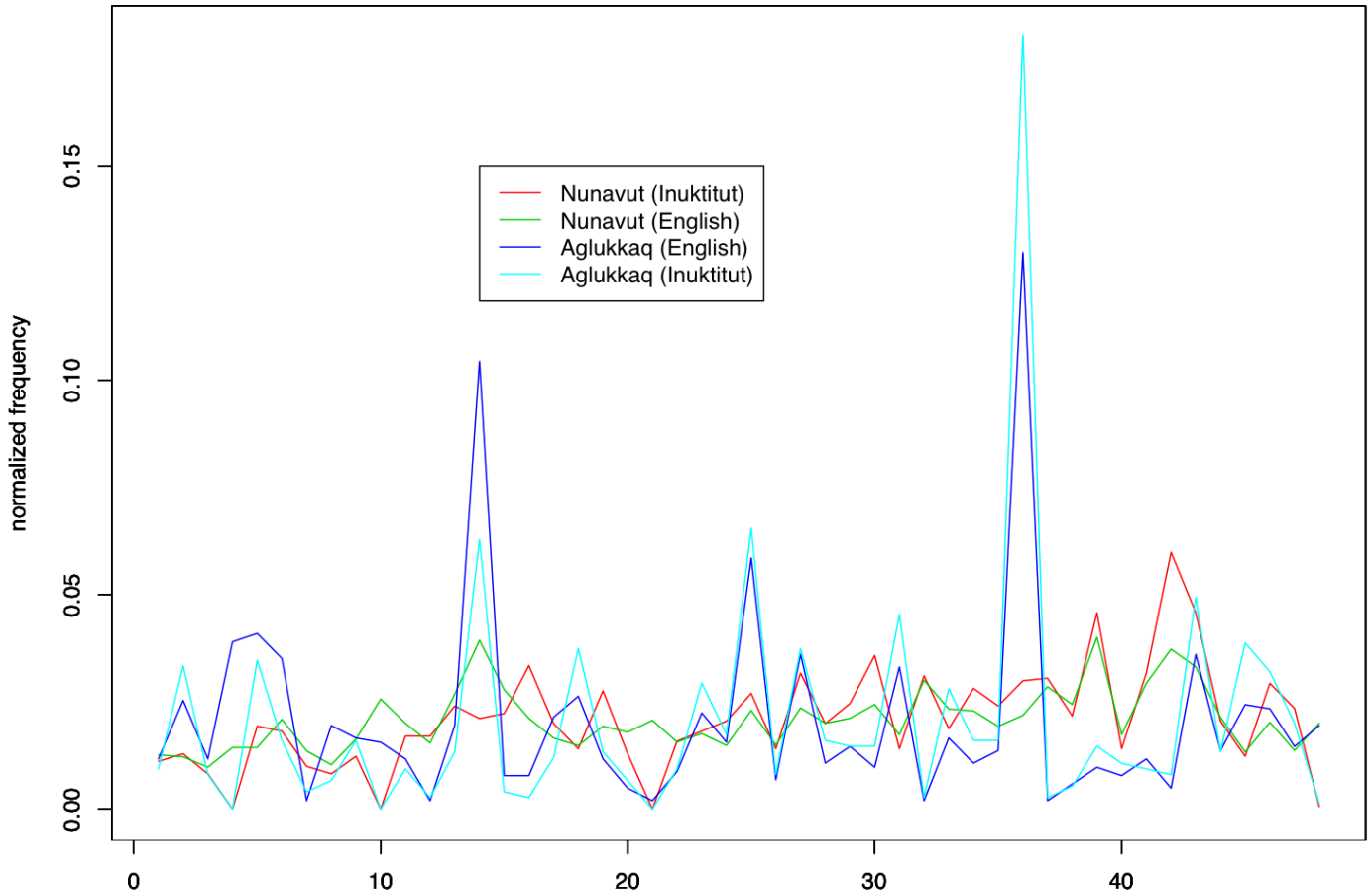
1	2.25	ᎩᎵᏏ	gilisi	k i l i s i	g i l i s i	26	1	|	6	6
2	2.33	ᏏᎴᏀ	yonega	j o n e k A	j o n e g a	26	1	|	6	6
3	2.56	ᎢᎳᏛᏃ	Italiano	i t h A l i A n o	I t A: l i: A: n o U	26	1	|	8	9
4	2.67	Kituwah	ᎩᏚᏣ	k I t S u w &	k i t u w A	1	26	|	6	6
5	2.83	ᎳᎻᏛ	தமிழ்	th A m i l i	t[A m i l A _	26	14	|	6	7
6	2.93	ᏚᏛᏗ	dulisdi	t u l i s t i	d u l & s d i:	26	1	|	7	7
7	3.08	ᏕᏫᏂ	Suomi	s u o m i	s u o U m i:	26	1	|	5	6
8	3.33	Kanuga	ᏫᏋᏏ	k @ n u g &	kh A n u k A	1	26	|	6	6
9	3.75	Keetoowah	ᎩᏚᏣ	k i: t u &	k i t u w A	1	26	|	5	6
10	3.88	adatasti	ᎠᏓᎳᏛᏗ	@ d & t @ s t i:	A t A th A s th i	1	26	|	8	8
11	4.04	dideloquasdi	ᏗᏘᏀᏚᏛᏗ	d I d & l o U k w @ z d i:	t i t e l o k w A s t i	1	26	|	13	11
12	4.04	ᏗᏘᏀᏚᏛᏗ	dideloquasdi	t i t e l o k w A s t i	d I d & l o U k w @ z d i:	26	1	|	11	13
13	4.08	limon	ᎵᎹᏂ	l I m & n	l e m A n i	1	26	|	5	6
14	4.19	ᏅᏓᏔᏕ	unulahi	u n u l A h i	& n j u l A: h i:	26	1	|	7	8
15	4.28	ᏗᏩᏯᏗ	diyohidi	t i j o h i t i	d I a I & h i: d i:	26	1	|	8	9
16	4.36	ugista	ᏅᏂᏀᏔ	u dZ I s t &	u ts i ts A th A	1	26	|	6	7
17	4.42	Cherokee	ᏀᏔᎩ	tS E r & k i:	ts A l A k i	1	26	|	6	6
18	4.42	ᏫᏔᏔ	Català	kh A th A l A	C a t a l _	26	1	|	6	6
19	4.42	ᏀᏔᎩ	Cherokee	ts A l A k i	tS E r & k i:	26	1	|	6	6
20	4.44	ᏔᏓᎢ	nudadequa	n u t A t e k w A	n & d A: d I k w &	26	1	|	8	9

Web Transliterations using Unitran/Handbuilt Distance Model

27	2.75	amaruq	ᐊᒪᕇᖅ	A: m A: r ^ k	A m A r u q	1	27	|	6
28	2.75	nuliaq	ᓄᓕᐊᖅ	n u l i: @ k	n u l i A q	1	27	|	6
29	2.75	pualuuk	ᐳᐊᔪᒃ	p u A: l & k	p u A l u: k	1	27	|	6
30	2.86	imiqtuq	ᐃᒥᖅᑐᖅ	I m I k t ^ k	i m i q t u q	1	27	|	7
31	2.90	Susan	ᔪᓚᓐ	s u z & n	s u: s A n	1	27	|	5
32	2.94	Qupanuaq	ᕿᑯᓄᐊᖅ	k u p a n u a k	q u p A n u A q	1	27	|	8
33	2.94	Sanirajak	ᓴᓂᕆᔭᒃ	s & n I r A: dZ & k	s A n i r A j A k	1	27	|	9
34	2.94	qikturiaq	ᖿᒃᑐᑦᓀᐊᖅ	k I k tS U r i: @ k	q i k t u r i A q	1	27	|	9
35	3.00	Igluligaarjuk	ᐃᒡᓗᓕᒫᔪᒃ	I g l ^ l i: g A: r dZ & k	i G l u l i G A: r j u k	1	27	|	12
36	3.00	Inuktitut	ᐃᓄᒃᑎᑐᑦ	I n ^ k t & t ^ t	i n u k t i t u t	1	27	|	9
37	3.00	arnaq	ᐊᕐᓇᖅ	A: r n @ k	A r n A q	1	27	|	5
38	3.00	inuktitut	ᐃᓄᒃᑎᑐᑦ	I n ^ k t & t ^ t	i n u k t i t u t	1	27	|	9
39	3.00	panik	ᐸᓂᖅ	p @ n I k	p A n i q	1	27	|	5
40	3.00	qaniq	ᖿᓇᓂᖅ	k A: n i: k	q A n i q	1	27	|	5
41	3.00	sukak	ᔪᒃᖅ	s u k & k	s u k A q	1	27	|	5

Temporal correlation: Nunavut Hansards

Nunavut ᓄᓇᕗᑦ
Aglukkaq ᐊᖏᔪᒃᑲᖅ



Synopsis

1. Given comparable corpora, such as newswire text, in a pair of languages that use different scripts:
   - `ScriptTranscriber` provides an easy way to mine transliterations from comparable texts.
   - Particularly useful for underresourced languages
2. `ScriptTranscriber` is an open source package that allows for ready incorporation of more sophisticated modules
3. Available as part of the nltk_contrib source tree at `http://code.google.com/p/nltk/`

Overview

- Approx. 7,500 lines of object-oriented Python
- Requires PySNoW
- Modules:
  - Document structure and XML representation.
  - Extractor: extracts terms from text. Specializations:
    - Capitalization-based extractor
    - Chinese foreign name extractor
    - Chinese personal name extractor
    - Thai extractor
  - Morph analyzer
  - Pronouncer. Specializations:
    - Unitran — UTF-8 pronouncer
    - English pronouncer
    - Hanzi (Chinese character) pronouncer
  - Comparator. Specializations:
    - Hand-built phonetic comparator
    - Time correlation comparator
    - Perceptron-based comparator

XML Fragment

埃及总统穆巴拉克、叙利亚总统阿萨德和沙特阿拉伯国王法赫德
28日和29日在埃及亚历山大市举行首脑会议。

Egyptian President Hosni Mubarak, Syrian president Hafez al-Assad and king Fahd of Saudi Arabia held a meeting in the northern Egyptian port city of Alexandria just before the end of last year.

```
<?xml version="1.0" encoding="UTF-8"?>
<doclist>
 <doc>
   <lang id="zh">
     ...
     <token count="1" morphs=""
      prons="sr a th &amp; a l a p o ;
      s u n A DUM g u m A k u d A k u DUM">沙特阿拉伯</token>
     <token count="1" morphs=""
      prons="f a x &amp; t &amp; ; n o r i A k A i o s i e">法赫德</token>
     <token count="1" morphs=""
      prons="m u p a l a kh &amp; ;
      j a w A r A g u d o m o e k u d A k u g A ts u">穆巴拉克</token>
     <token count="1" morphs="" prons="a s a t &amp;
      ; k u m A DUM o s i e">阿萨德</token>
   </lang>
   <lang id="en">
     <token count="1" morphs="" prons="@ l &amp; g z @ n d r i: &amp;">Alexandria</token>
     <token count="1" morphs="" prons="&amp; r e I b i: &amp;">Arabia</token>
     <token count="1" morphs="" prons="&amp; s A: d">Assad</token>
     <token count="1" morphs="" prons="I dZ I p S &amp; n">Egyptian</token>
     <token count="1" morphs="" prons="f A: d">Fahd</token>
     ...
     <token count="1" morphs="" prons="m u b A: r I k">Mubarak</token>
     ...
     <token count="1" morphs="" prons="s &gt; d i:">Saudi</token>
     ...
   </lang>
 </doc>
</doclist>
```

Sample Program

```
#!/bin/env python
# -*- coding: utf-8 -*-

"""Sample transcription extractor based on the LCTL Thai parallel
data. Also tests Thai prons and alignment.
"""

__author__ = """
rws@uiuc.edu (Richard Sproat)
"""

import sys
import os
import documents
import tokens
import token_comp
import extractor
import thai_extractor
import pronouncer
from __init__ import BASE_

## A sample of 10,000 from each:

ENGLISH_        = '%s/testdata/thai_test_eng.txt' % BASE_
THAI_           = '%s/testdata/thai_test_thai.txt' % BASE_
XML_FILE_       = '%s/testdata/thai_test.xml' % BASE_
MATCH_FILE_     = '%s/testdata/thai_test.matches' % BASE_
```

Sample Program

```
BAD_COST_      = 6.0

def LoadData():
  t_extr = thai_extractor.ThaiExtractor()
  e_extr = extractor.NameExtractor()
  doclist = documents.Doclist()
  doc = documents.Doc()
  doclist.AddDoc(doc)
  #### Thai
  lang = tokens.Lang()
  lang.SetId('th')
  doc.AddLang(lang)
  t_extr.FileExtract(THAI_)
  lang.SetTokens(t_extr.Tokens())
  lang.CompactTokens()
  for t in lang.Tokens():
    pronouncer_ = pronouncer.UnitranPronouncer(t)
    pronouncer_.Pronounce()
  #### English
  lang = tokens.Lang()
  lang.SetId('en')
  doc.AddLang(lang)
  e_extr.FileExtract(ENGLISH_)
  lang.SetTokens(e_extr.Tokens())
  lang.CompactTokens()
  for t in lang.Tokens():
    pronouncer_ = pronouncer.EnglishPronouncer(t)
```

Sample Program

```
    pronouncer_.Pronounce()
  return doclist


def ComputePhoneMatches(doclist):
  matches = {}
  for doc in doclist.Docs():
    lang1 = doc.Langs()[0]
    lang2 = doc.Langs()[1]
    for t1 in lang1.Tokens():
      hash1 = t1.EncodeForHash()
      for t2 in lang2.Tokens():
        hash2 = t2.EncodeForHash()
        try: result = matches[(hash1, hash2)] ## don't re-calc
        except KeyError:
          comparator = token_comp.OldPhoneticDistanceComparator(t1,  t2)
          comparator.ComputeDistance()
          result = comparator.ComparisonResult()
          matches[(hash1, hash2)] = result
  values = matches.values()
  values.sort(lambda x, y: cmp(x.Cost(), y.Cost()))
  p = open(MATCH_FILE_, 'w') ## zero out the file
  p.close()
  for v in values:
    if v.Cost() > BAD_COST_: break
    v.Print(MATCH_FILE_, 'a')
```

Sample Program

```
if __name__ == '__main__':
  doclist = LoadData()
  doclist.XmlDump(XML_FILE_, utf8 = True)
  ComputePhoneMatches(doclist)
```

Interactive Use

```
>>> import pronouncer
>>> import tokens
>>> t1 = tokens.Token('WWJD')
>>> t2 = tokens.Token('拉拉瓜')
>>> p = pronouncer.UnitranPronouncer(t1)
>>> p.Pronounce()
>>> t1
#<WWJD 1 [] [' l A l A k u A'] >
>>> p = pronouncer.HanziPronouncer(t2)
>>> p.Pronounce()
>>> t2
#<拉拉瓜 1 [] ['l a l a k w a', 'k u d A k u k u d A k u u r i'] >
>>> import token_comp
>>> c = token_comp.OldPhoneticDistanceComparator(t1, t2)
>>> c.ComputeDistance()
>>> c.ComparisonResult()
#<comparator: WWJD <-> 拉拉瓜, 3.2857, "l A l A k u A <-> l a l a k w a">
>>> c.ComparisonResult().Cost()
3.2857142857142856
```

Summary

- `ScriptTranscriber` is a toolkit for extracting transliteration pairs from comparable corpora.
  - Works with any script in the Unicode Basic Multilingual Plane
  - Easy to extend the modules
- Available from the nltk_contrib source tree at `http://code.google.com/p/nltk/`.

Acknowledgments

Work reported here was partially funded by NBCHC040176 from the US Department of the Interior, a Google Research Award, and the National Science Foundation under grant #0705708 to the Center for Language and Speech Processing at the Johns Hopkins University.