CENTER FOR ADVANCED
STUDY OF LANGUAGE

HLT Tools for Low-Resource Languages – Language Nation

Nikki Adams, Jay Ritch, Chris Green, Corey Miller,
Evan Jones and Mark Vinson

Technology LEARN Workshop, August 6, 2014

HLT Tools for Language Nation

- Why HLT tools?
  - o Boost working capability
  - o Increase accuracy
  - o Reduce time spent on particular task
- Challenges for low-resource languages
  - o Scarce dictionaries for building dictionary-based tools
  - o Scarce grammatical descriptions or pedagogical materials used in building tools
  - o Above often requires fieldwork before and while building tools

Somali Deep-Tagged Corpus and Word Frequency List

- Hand-tagged ~8000-word corpus, includes not only PoS, but morphological parse
- Words tagged for frequency and by domain
- Can be used for automatic fetching of domain-specific texts for learners in online environment
- Used to test accuracy of morphological parser
- Can be used in building syntactic parser

Somali – Place names

- The same set of place names – Somali cities and regions as well as major cities in neighboring countries – recorded by Somali speakers of different dialects
- Allows users to compare pronunciation of place names to help determine origin of speakers.
- Also allows users to help determine “out of place” speakers

Somali frequency list with deep-tagged words

token	root	count	Part of Speech	News vs Wikipedia Dominance	News Normalized Frequency	Pedagogy Normalized Frequency	Wikipe
oo	oo	34131	<CONJ>		29459	8877	27551
ka	ka	28506	<Adposition>		26975	10279	20697
ku	ku	24546	<Adposition>/<2.SG.OP>		23028	18455	17811
ay	ay	23030	<3.SP>		27905	15029	10774
ee	ee	22961	<CONJ>		22451	4049	16074
iyo	iyo	21852	<CONJ>		14482	13082	21644
ah	ahow	21337	<Verb><PRES.RED.ABS>		18082	3971	17574
u	u	17677	<Adposition>		15697	18767	13554
in	in	17507	<COMP>/<Noun>		22550	10590	6937
ayaa	ayaa	16927	<Focus>		24300	1635	4516
uu	uu	15308	<3.SG.M.SP>/<3SG.M.SP>		14925	7320	10660
la	la	13200	<PRON.IMP>/<Adposition>		12708	9266	9283
soo	soo	11594	<DEIC>		11213	7008	8129
waxaa	waxa	10857	<Focus.cleft>/<Focus.Cleft>/<Focus.clef>		8056	545	10060
waa	waa	9396	<Declarative>		1705	21414	13272
waxa	waxa	7862	<Focus.Cleft>/<Focus.cleft>		4929	6930	8008
ugu	u+ku	7295	<Adposition><-AC>/<3.SG.M.SP><-AC>		6085	1869	6090
lagu	la+ku	7104	<PRON.IMP><-AC>		6951	3893	4913
mid	mid	6429	<Noun>		4379	779	6320
waxay	waxa+ay	6220	<Focus.cleft><-PC.3>/<Focus.Cleft><-PC.3.PL>/<Foc		4896	1869	5465
kale	kale	5581	<Adjective>		4255	4594	4973
badan	bada	5519	<Noun>/<Adjective>		3587	2414	5553

Somali MaDYM

- Morphologically-Aware Did You Mean?
  - Components
    - Zorc and Osman dictionary in XML form
    - Spell-checking for above dictionary
    - Morphological analyzer
  - Enables electronic dictionary access for misspelled words: shakal
  - Provides analysis for (possibly misspelled) morphologically-complex words: niinka

Persian - MaDYM

- Morphologically-Aware Did You Mean?
  - o Components
    - Haim and Steingass dictionaries in XML form
    - Morphological analyzer
  - o Enables electronic dictionary access for misspelled words: كطاب
  - o Provides analysis for (possibly misspelled) morphologically-complex words: كطابحا

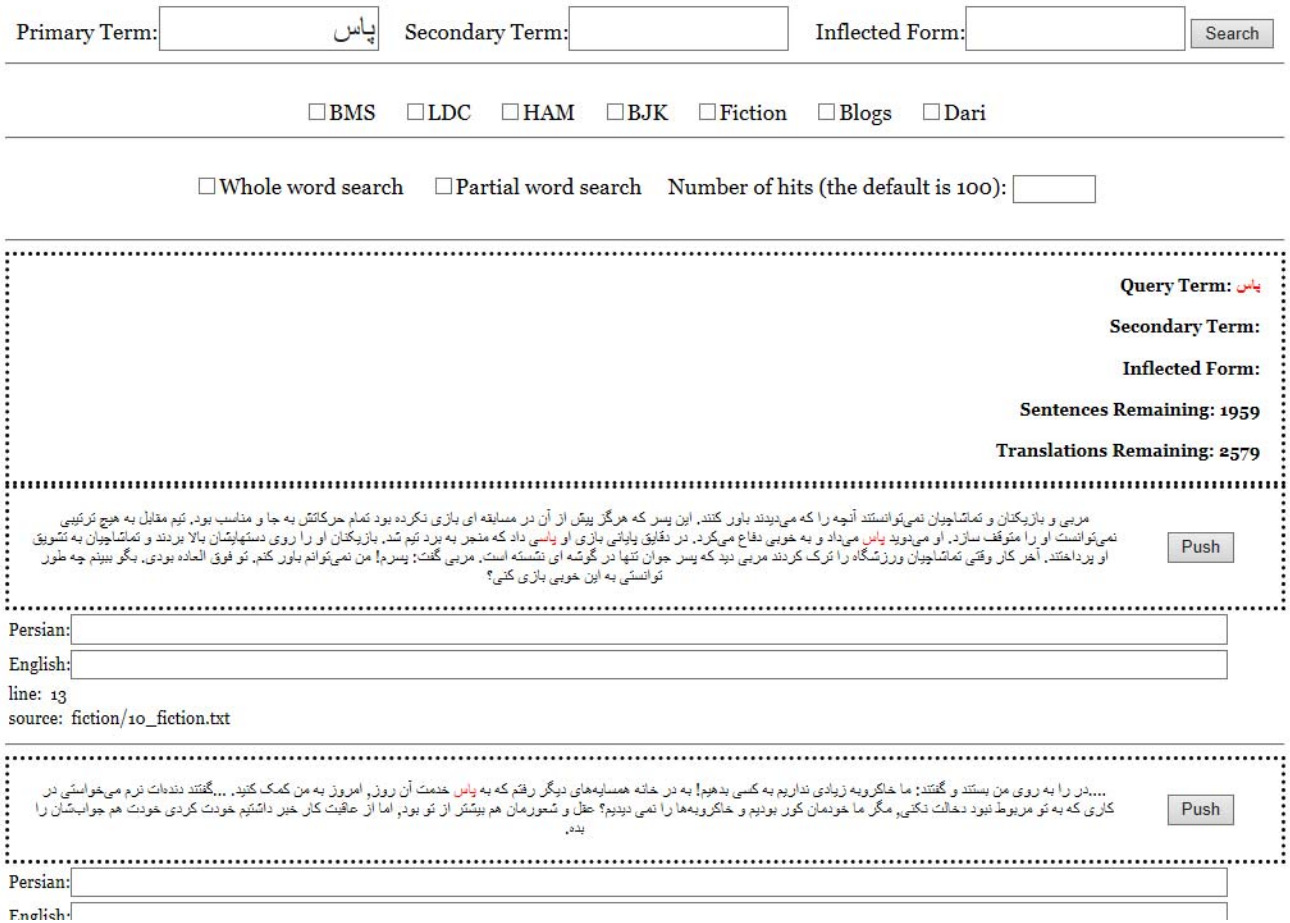
Persian Word Frequency

- Identify most frequently occurring words
  - Ensure coverage in curriculum: texts, exercises, tests
  - Assist with automatic ILR leveling of class materials
- Create corpus of news, fiction, blogs including Farsi and Dari
- Analyze corpus to identify word families including inflections and stylistic variants: ...رفت، میرود، میره
- Built concordancer to identify example words in context

Persian frequency list with sample words in context

General Rank	Word Family	Pronunciation	Gloss	Arabic Root	Written or Spoken	Attested Word Forms	Cumulative Coverage	Words in Context - Persian	Words in Context - English
32	کار	kar	work			کار (Singular Noun), کار	35.49%	کار پیدا کردی؟	Did you find work?
33	رفتن	ræftæn	go			رفت (Singular Noun/...nse	35.69%	با آن پسر رفت بیرون چند مرتبه	She went out with that guy several times
34	بایستن	bajestæn	be required			باید (Verb Auxiliary), نباید	35.92%		
35	روز	ruz	day			روز (Singular Noun), رو	36.13%	سه روز دیگه برمیگردند	They return in three more days.
36	هر	hær	each			هر (Quantifier)	36.34%	هر جای دنیا هم که باشی این به دردت میخورد	Wherever you are in the world, this will help you.
37	آمدن	ɑmædæn	come			آمده (...erb Present Perfect)	36.52%	پسر تازه از ایران آمده بود	The boy had recently come from Iran.
38	نیز	niz	also			نیز (Conjunction)	36.74%		
39	اما	æmmɑ	but			اما (Conjunction)	36.96%	پول خوبی در میآورند میگوید اما زندگی اهمیت ندارد	He says they make good money, but life has no importance.
40	رسیدن	residæn	reach, achieve			رسید (Verb Past Tense), ر	37.16%	رسید به سیصد تومن	It reached three hundred toman.
41	دست	dæst	hand			دست (Singular Noun), دس	37.36%	دست خالی به خانه باز نمی گشت	He wouldn't return home empty handed again.
42	بیش	biʃ	more			بیشتر (...dverb/Quantifier)	37.55%	معمولا تو از من خیلی بیشتر انگلیسی حرف میزنی	Usually you speak far more English than I do
43	بعد	bæʔd	then, after	ب ع د		بعد (...ective/Conjunction)	37.70%	دفعهٔ بعد هم باید بیای	You have to come next time, too
44	یا	jɑ	or			یا (Conjunction)	37.91%		
45	قرار	gærɑr	agreement	ق ر ر		قرار (Singular Noun)	38.10%	قرار شد نامه هایی که برای شما می نویسم به او نشان بدم تا	It was decided that I would show him the letters I write so th
46	حال	hal	state	ح و ل		حال (Singular Noun), حال	38.28%	حال شما چه طوره؟	How are you?
47	خوب	xub	good			خوب (Adjective/Adverb)	38.38%	انگلیسیتش هم خوب نبود	His English also wasn't good.
48	دیدن	didæn	see			دید (...ense/Singular Noun)	38.54%	تو کنسرت هم من را دید	He also saw me at the concert.
49	نه	næ, noh	not, no			نه (...ction/Singular Noun)	38.66%	در ظرف نه ماه	In the span of nine months
50	مردم	mærdom	people			مردم (Singular Noun), مر	38.88%	خیلی از مردم دنباله کار اند	Many people are looking for work
51	سه	se	three			سه (Singular Noun), سوم	39.04%	لازم نیست سه هفته تمام تو بیمارستان بماند	It's not necessary for him to stay three whole weeks in the h
52	همه	hæme	every			همه (...uantifier/Pronoun)	39.61%	متأسفانه همه چیز روسی نوشته شده	Unfortunately, everything has been written in Russian
53	دانستن	dɑnestæn	know			دانست (Verb Past Tense)	39.73%	با نگرانی منتظر دانستن علت احضارش پایه پا میتمد.	He paced around, anxiously awaiting to know the reason for
54	پس	pæs	then			پس (...ective/Conjunction)	39.88%	پس خودت الان گرسنه ات نیست	So, you are not hungry now.
55	او	u	he, she			او (Pronoun Personal)	40.09%	آیا میدانستید او نابینا بود؟	Didn't you know that she was blind?
56	چه	tʃe	what			چه (...tion/Singular Noun)	40.24%	تو نمیدانی چه صبحانه شیک واقعاً مجللی دادند	You don't know what a chic and truly luxurious breakfast the
57	مورد	mored	case	و ر د		مورد (Singular Noun), مو	40.41%	این موردش فرق می کند، به شماها ربطی ندارد!	This case is different, it has nothing to do with you.
58	اگر	ægær	if			اگر (IF), اگه (IF)	40.56%	اگر اجازه دهید همین چند ماه را در خدمتتان باشم	If you'll allow me I'd like to serve you in these few months.
59	دولت	dowlæt	government	د و ل		دولت (Singular Noun), دو	40.74%	بعد حساب کردم من اگر بخواهم یه سال دیگه بخوانم حدوده	Next I calculated that if I want to study for one more year, I'd
60	هزار	hæzɑr	thousand			هزار (Singular Noun), ه	40.93%	بیست هزار دلار دارد	He has twenty thousand dollars.
61	همین	hæmin	the same, this			همین (...Pronoun Personal)	41.04%	آدم بلاخره میتواند یه نیم ساعت همین جوری حرف بزند	A person can talk in this way for a half an hour.
62	اول	ævvæl	first	ا و ل		اول (...tive/Singular Noun)	41.18%	مرد اول از خدا خانه، لباس و غذای بیشتری خواست	The first man asked God for more clothing, food, and housin
63	جهان	dʒæhɑn	world			جهان (Singular Noun), ج	41.35%	یکی از فرماندهان برجسته امریکایی در جنگ جهانی دوم ب	He was one of the prominent American commanders in the S
64	زدن	zædæn	touch, strike			زد (Verb Past Tense), زد	41.47%		
65	تو	to, tu	you			تو (Pronoun Personal),	41.60%	تو چکار میکنی	What do you do?
66	نظر	næzær	opinion	ن ظ ر		نظر (Singular Noun)	41.74%		
67	اصل	æsl	principle	ا ص ل		اصلی (Adjective), اصل (ا	41.84%	من اصلا حوصله اش را نداشتم	I didn't have any motivation for it.
68	چیز	tʃiz	thing			چیزی (Singular Noun), چ	41.93%	بعد داخل شکمم چیز شده	Then something happened inside my stomach
69	اسلام	eslɑm	Islam	س ل م		اسلامی (Adjective), اسلام	42.11%	رحم از فرایض اسلام است	Mercy is one of the practices of Islam.

Persian Concordancer



For more information

Contact

Corey Miller

Research Scientist

301-226-8834

cmiller@casl.umd.edu

www.casl.umd.edu