THE FINNO-UGRIC LANGUAGES AND THE INTERNET PROJECT

Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén
Department of Modern Languages

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS

FIN-CLARIN
KONEEN SÄÄTIÖ

INTRODUCTION

The ”Finno-Ugric Languages and The Internet” project started at the beginning of 2013 as part of the Kone Foundation Language Programme [1] and the international CLARIN cooperation. The main goal of the project is to build a prototype of a system that will crawl the internet and gather texts written in small Uralic languages. The largest Uralic languages Hungarian, Finnish, and Estonian are outside the scope of the project. The gathered texts will be collected into sentence and word corpora for each language and the links to the associated web-pages into link collections. The corpora will act as a source for linguists and the link collections will hopefully spread the knowledge of the existence of relevant pages to interested parties.

LANGUAGE IDENTIFICATION

We are using a language identifier, developed within the project [2], which uses relative frequencies of n-grams of characters together with tokens and token-based backoff. The language identifier recognizes 285 languages from all around the world, including 34 Uralic languages. The amount of training material differs considerably between languages ranging from 19000 characters in Ume Sami to over 400 million characters in the Hungarian material.

The average identification accuracy for Uralic languages is generally slightly lower than for all languages. This is due to some of the languages being very close varieties of each other, especially within the Finnic languages.

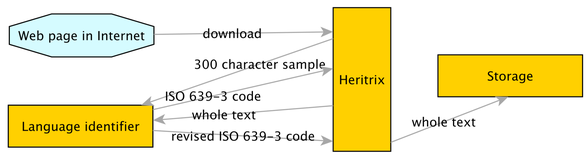
	ekk	fin	fit	fkv	izh	krl	lud	olo	vep	vot	vro
ekk	99.3 %	0.1 %									0.1 %
fin		**85.6** %	**10.6** %	0.9 %	0.7 %	*1.5* %		0.4 %			
fit		*5.1* %	**81.2** %	**11.2** %	*1.0* %	0.7 %		0.5 %			
fkv		*4.0* %	**10.3** %	**83.4** %	*1.2* %	0.1 %	0.3 %				0.1 %
izh	0.3 %	*5.3* %	*3.3* %	*3.4* %	**81.0** %	*2.3* %	0.5 %	*1.4* %		0.3 %	*1.0* %
krl		*7.7* %	*1.4* %	0.5 %	0.4 %	**80.1** %	0.1 %	*4.0* %	0.9 %		0.4 %
lud	*1.0* %	*3.2* %	0.1 %	0.1 %	*2.7* %	*1.3* %	**69.7** %	**12.8** %	*5.2* %	0.4 %	0.3 %
olo	0.3 %	*1.8* %	0.2 %	0.2 %	0.2 %	*3.7* %	*3.8* %	**86.2** %	0.6 %		0.4 %
vep	0.7 %	0.6 %	0.1 %		0.2 %		*2.8* %	*1.9* %	**90.3** %		0.6 %
vot	*3.2* %	*3.7* %	*4.8* %	0.3 %	*5.1* %	0.8 %	0.3 %	*1.3* %	0.8 %	**69.2** %	*5.6* %
vro	*3.2* %				0.1 %		0.7 %	0.4 %		0.2 %	**94.6** %

Confusion matrix of Finnic languages.

For the test length of 20 characters, the overall average identification recall is 93.9% whereas the average is 90.5% for the Uralic languages. Almost all languages attain 100.0% recall at 150 characters.

CRAWLING THE NATIONAL DOMAINS

In order to crawl for pages written in small Uralic languages, we use Heritrix [3], a web archiving system developed by the Internet Archive. The version we are currently using downloads all text files as well as pdf files it finds from within the domain in question.



A diagram showing how Uralic web pages are processed during a crawl.

We have chosen to start collecting the material by crawling the national domains most likely to contain material written in small Uralic languages: .ee, .fi, .no, .ru, and .se.

	URLs	LI-1 URLs	LI-2 URLs	domains	LI-1 domains	LI-2 domains
.fi	354 000 000	89 166	39 056	450 000	2 824	1 465
.se	308 000 000	16 687	14 979	1 500 000	676	439
.no	358 000 000	137 059	133 513	800 000	636	586
.ru	172 000 000	18 122	8 585	1 400 000	3 243	909
.ee	108 000 000	22 785	13 496	100 000	500	232

Statistics for the crawls of the five national domains.

We will be doing new crawls for them all as most of the crawls ended before the domains were really exhausted. It is actually far from trivial to define when we have exhausted a national domain. There are many sites that dynamically generate an infinite number of web-pages and even sub-domains, which makes each of the national domains infinite in size if we are calculating the number of pages or sub-domains.

THE LINK COLLECTION

The link collection that is available at http://suki.ling.helsinki.fi/sites has been curated by hand from the pages of the .fi crawl. It contains links to 266 sites from which text was found in 19 of the 31 small Uralic languages searched. The links have not been verified by experts or native speakers. We are planning to incorporate a simple crowd-sourcing platform to be able to get feedback from those who are more familiar with the languages. Our goal is to make the creation of the link collection as automated as possible, avoiding manual link curation.

SENTENCE CORPORA

When we are creating a sentence corpora, one of the greatest problems we have at the moment is that many of the downloaded pages are multilingual. We are currently making a survey of the methods for language identification in multilingual documents and in future we will incorporate a multilingual detection method in the system. We did a separate language identification for all the lines of all the files containing small Uralic languages in order to see which ones are, indeed, written in the language indicated by the identification of the file as a whole.

FUTURE WORK

Language identification methods will be further developed in order to improve the robustness of the language identifier we use. We will, furthermore, try to increase the speed of the crawler in order to crawl more widely and more often. The most important national domains in regard to the Uralic language speakers will be re-crawled with more depth and more frequency. We also intend to look into crawling the .com and .org domains. We would also like to extract text from other binary files than pdfs.

	#unique lines	#words	#characters
Northern Sami (sme)	312 150	3 209 570	30 314 461
Võro (vro)	167 997	3 239 365	22 940 862
Ingrian (izh)	98 743	2 960 322	22 054 552
Eastern Mari (mhr)	132 692	1 626 001	20 586 975
Western Mari (mrj)	137 739	882 884	10 115 581
Southern Sami (sma)	86 856	814 187	9 264 654
Udmurt (udm)	41 133	570 633	7 554 055
Erzya (myv)	29 742	503 107	6 911 773
Lule Sami (smj)	53 734	376 067	3 436 123
Inari Sami (smn)	35 740	352 319	3 428 425
Tornedalen Finnish (fit)	21 133	384 037	3 137 644
Moksha (mdf)	15 931	202 740	2 853 814
Komi-Zyrian (kpv)	13 139	205 243	2 374 729
Skolt Sami (sms)	23 354	188 873	2 010 098
Livvi (olo)	6 622	112 560	940 632
Liv (liv)	12 194	85 171	602 979
Kven Finnish (fkv)	3 414	57 199	500 600
Ludian (lud)	2 078	53 094	485 457
Khanty (kca)	7 244	38 704	378 562
Veps (vep)	5 480	29 691	324 504
Komi-Permyak (koi)	4 370	19 982	186 543
Karelian (krl)	950	11 550	103 498
Mansi (mns)	319	4 997	60 811
Votic (vot)	702	5 895	42 657
Kildin Sami (sjd)	332	2 751	32 409
Ume Sami (sju)	194	2 636	21 703
Nenets (yrk)	209	1 165	14 370
Selkup (sel)	639	1 486	12 849
Nganasan (nio)	195	428	6 950
Tundra Enets (enh)	6	14	112

The number of lines, words, and characters in small Uralic languages after language identifying each individual line.

REFERENCES

[1] Kone Foundation. The language programme 2012-2016. http://www.koneensaatio.fi/en, 2012.

[2] Tommi Jauhiainen and Krister Lindén. Identifying the language of digital text. In review, submitted 08/14, 2015.

[3] Gordon Mohr, Michael Stack, Igor Rnitovic, Dan Avery, and Michele Kimpton. Introduction to heritrix. In *4th International Web Archiving Workshop*, Bath, 2004.

CONTACT INFORMATION

firstname.lastname@helsinki.fi

http://suki.ling.helsinki.fi