The Finno-Ugric Languages and the Internet project

Jauhiainen, Heidi

Septentrio Academic Publishing
2015-01-15

Jauhiainen , H , Jauhiainen , T & Linden , K 2015 , The Finno-Ugric Languages and the Internet project . in T Pirinen , F Tyers & T Trosterud (eds) , First International Workshop on Computational Linguistics for Uralic Languages : Proceedings of the Workshop . vol. 2 , Septentrio Conference Series , no. 2 , vol. 2015 , Septentrio Academic Publishing , Tromsø , þÿpp. 87 98 , International Workshop on Computational Linguistics for U Tromsø , Norway , 16/01/2015 . https://doi.org/10.7557/scs.2015.2

http://hdl.handle.net/10138/159402
https://doi.org/10.7557/scs.2015.2

Proceedings of 1st International Workshop in Computational Linguistics for Uralic Languages (IWCLUL 2015); ‹http://dx.doi.org/10.7557/scs.2015.2›

The Finno-Ugric Languages and The Internet project

Heidi Jauhiainen
University of Helsinki
Department of Modern Languages
heidi.jauhiainen@helsinki.fi

Tommi Jauhiainen
University of Helsinki
Department of Modern Languages
tommi.jauhiainen@helsinki.fi

Krister Lindén
University of Helsinki
Department of Modern Languages
krister.linden@helsinki.fi

December 12, 2014

Abstract

This paper describes a Kone Foundation funded project called ”The Finno-Ugric Languages and The Internet” together with some of the achieved results. The main activity of the project is to crawl the internet and gather texts written in small Uralic languages. The sentences and words of the found texts will be assembled into a freely available corpus. Crawling is done using the open source crawler Heritrix, which is developed by the Internet Archive. Heritrix crawls through the pages and passes the found texts to a language identifier. We are using a state of the art language identifier, which has been further developed within the project and has been evaluated using 285 languages. We describe the language

identification evaluation results concerning the 34 Uralic languages known by the language identifier. We also describe the initial observations and results from the first five large crawls which were done in the national internet domains of Finland, Sweden, Norway, Russia, and Estonia.

1 Introduction

"The Finno-Ugric Languages and The Internet" -project[1] started at the beginning of 2013 as part of the Kone Foundation Language Programme [1]. The project is located at the Department of Modern Languages[2] at the University of Helsinki and is part of the international CLARIN[3] 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. Web crawling has been used to collect text corpora for a variety of languages [2, 3], but to our knowledge, this project is the first collecting the texts in small Uralic languages. The largest Uralic languages Hungarian, Finnish, and Estonian are outside the scope of the project. We are using language identification software to detect the language of the crawled web-pages. 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. Due to the copyrights connected with longer texts, we are only publishing corpora of up to sentence-length text snippets with links to the original text on the internet. The negotiations for free use of copyrighted texts would be far beyond the resources of this project. We aim at making the complete work-flow from the crawling to the creation of corpora as automated as possible.

In Section 2, we talk about the language identifier used in the project with respect to Uralic languages. Section 3 deals with the five large crawls done so far, i.e. in the national domains of Finland, Sweden, Norway, Russia and Estonia. In Sections 4 and 5, we describe the link collection and the sentence corpora respectively.

2 Language Identification for Uralic Languages

We are using an extended version of the language identifier described in [4]. The extended version of the language identification method will be described in a forthcoming journal article [5], where it is evaluated together with the methods presented

[1] http://suki.ling.helsinki.fi
[2] http://www.helsinki.fi/modernlanguages/
[3] http://clarin.eu

in [6], [7], [8], [9], [10], and [11]. The language identifier uses relative frequencies of *n*-grams of characters together with tokens and token-based backoff. The evaluated language identifier recognizes 285 languages from all around the world. The definition of a language is taken from Ethnologue [12] and the division of the languages is as in the ISO 639-3 standard[4]. The 285 languages include 34 Uralic languages: Hungarian, Khanty, Mansi, Estonian, Finnish, Kven Finnish, Tornedalen Finnish, Ingrian, Karelian, Liv, Livvi-Karelian, Ludian, Veps, Votic, Võro, Hill Mari, Meadow Mari, Erzya, Moksha, Udmurt, Komi-Permyak, Komi-Zyrian, Inari Sami, Kildin Sami, Skolt Sami, Ume Sami, Lule Sami, North Sami, South Sami, Nenets, Nganasan, Forest Enets, Tundra Enets and Selkup. The current version of the language identifier only knows one orthography per language, but this will be corrected in future versions.

The evaluation of the language identifier was done in tests using sequences from 5 to 150 characters in length and the recall figures for Uralic languages can be seen in Table 1. The character sequences are random parts of the test corpus, always beginning from the beginning of a word. For most of the languages the test set consists of the texts of the Universal Declaration of Human Rights in that particular language and the training set is the text from Wikipedia. The aim was that the test set would always be a text from a different domain than the training text. This was easily possible for most of the larger languages, but quite difficult for some of the smaller Uralic languages. In some cases, such as Forest and Tundra Enets, the test set is from a different section of the same document as the training text. The amount of training material differs considerably between languages ranging from 19 000 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. The language identifier has not been optimized to perform better with Uralic languages or even with closely related languages. In the test length of 20 characters, the overall average is 93.9% whereas the average is 90.5% for the Uralic languages. Almost all languages attain 100.0% recall at 150 characters. Table 1 also includes the recall figures attained by the widely used method described in [6].

Table 2 is a confusion matrix showing the kind of mistakes that were made in the language identifications between most of the Finnic languages in the 20 character sized tests. Notable problem pairs are those of Finnish (fin) and Tornedalen Finnish (fit), Tornedalen Finnish (fit) and Kven Finnish (fkv), as well as Ludian (lud) and Livvi-Karelian (olo). The Samic languages are not as easily confused and no table is presented for them.

[4] http://www.sil.org/iso639-3/

	SUKI	# char.				C&T	# char.			
639-3	**5**	**20**	**40**	**80**	**150**	**5**	**20**	**40**	**80**	**150**
ekk	55.0%	99.3%	100%	100%	100%	20.4%	94.5%	99.4%	100%	100%
enf	65.2%	91.7%	98%	100%	100%	10.0%	86.5%	98.9%	100%	100%
enh	56.2%	82.5%	88.9%	97%	100%	8.3%	79.7%	94.7%	99.4%	99.8%
fin	29.1%	85.6%	97.9%	100%	100%	12.6%	80.5%	92.2%	95.8%	99.2%
fit	39.8%	81.2%	94.4%	97.8%	100%	11.7%	75.3%	93.5%	97.1%	100%
fkv	52.8%	83.4%	90.4%	97.9%	100%	3.6%	53.1%	80.2%	94.9%	100%
hun	84.6%	99.8%	100%	100%	100%	17.4%	97.2%	99.9%	100%	100%
izh	42.3%	81%	95%	99.6%	100%	8.7%	84.6%	97.5%	100%	100%
kca	73.2%	94.5%	99.5%	100%	100%	47.3%	90.4%	98.4%	100%	100%
koi	72.0%	98.9%	100%	100%	100%	37.3%	98.3%	99.9%	100%	100%
kpv	71.0%	92.5%	96.9%	100%	100%	29.9%	87.8%	97.3%	100%	100%
krl	31.0%	80.1%	91.9%	98.6%	100%	3.1%	48.7%	71.2%	90.9%	95.8%
liv	72.7%	91.7%	99.5%	100%	100%	46.4%	96.0%	100%	100%	100%
lud	40.0%	69.7%	85.5%	97.4%	100%	34.6%	78.7%	88.4%	93.5%	100%
mdf	62.7%	89.4%	98.2%	99.2%	100%	9.5%	73.5%	95.3%	100%	100%
mhr	62.9%	95.6%	99.3%	100%	100%	44.0%	93.2%	98.4%	99.9%	100%
mns	65.0%	91.4%	98.5%	100%	100%	0.6%	70.1%	95.5%	99.8%	100%
mrj	85.6%	99.1%	100%	100%	100%	47.6%	98.0%	100 %	100%	100%
myv	65.1%	92.7%	98.2%	99.6%	99.8%	30.7%	91.5%	97.7%	99.4%	99.6%
nio	89.3%	99.9%	100%	100%	100%	20.4%	99.1%	99.9%	100%	100%
olo	38.2%	86.2%	93.7%	99.7%	100%	6.7%	62.4%	86.1%	93.5%	96.8%
sel	73.8%	88.3%	93.2%	94.7%	100%	9.1%	79.5%	92.3%	95.7%	100%
sjd	77.9%	95.3%	98.6%	99.8%	100%	42.7%	92.7%	97.1%	99.4%	100%
sju	74.9%	94.9%	98.9%	100%	100%	6.5%	89.9%	98.4%	100%	100%
sma	59.4%	90.1%	98.1%	100%	100%	30.1%	92.8%	99.5%	100%	100%
sme	48.8%	96.1%	99.2%	99.2%	100%	18.0%	94.1%	99.4%	100%	100%
smj	72.2%	99.5%	100%	100%	100%	48.6%	97.5%	100%	100%	100%
smn	68.6%	89.7%	98.1%	99.8%	100%	40.9%	91.2%	99.8%	100%	100%
sms	81.0%	96.6%	99.3%	100%	100%	59.5%	96.4%	99.8%	100%	100%
udm	63.6%	95.2%	98.4%	99.8%	100%	50.2%	95.3%	99.0%	100%	100%
vep	48.6%	90.3%	96.9%	100%	100%	15.0%	87.1%	98.1%	100%	100%
vot	35.9%	69.2%	84.4%	93.4%	99%	29.4%	80.8%	95.5%	99.6%	100%
vro	49.5%	94.6%	99.1%	100%	100%	33.4%	89.6%	96.0%	100%	100%
yrk	68.1%	89.5%	97.1%	100%	100%	5.8%	59.8%	91.3%	100%	100%
Average	**61.1%**	**90.5%**	**96.7%**	**99.2%**	**100%**	**24.7%**	**84.9%**	**95.6%**	**98.8%**	**99.7%**

Table 1: Recalls of Uralic languages obtained by the two language identifiers for test lengths between 5 and 150 characters. Percentages are averages over 1000 sample sequences of each length. The figures on the left are for the identifier developed within the project and the figures on the right are for an identifier using the well-known method of Cavnar & Trenkle.

	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** %

Table 2: Confusion matrix of Finnic languages. The Finnic languages were mistaken also as other languages, and if the figures of the other languages would be added to the table, the rows would add to 100.0%.

Table 3 shows the confusions between most of the Uralic languages written in Cyrillic script. Forest Enets seems to dominate over Tundra Enets as does Komi-Permyak over Komi-Zyrian. The language model for Komi-Permyak is based primarily on Wikipedia and the one for Komi-Zyrian on a bible translation. The test material for Komi-Zyrian is also from the bible, which should in fact make it easier to identify but, nevertheless, 6.9% of the 20 character extracts are identified as Komi-Permyak.

	enf	enh	kca	koi	kpv	mdf	mhr	mrj	myv	udm
enf	**91.7** %	*3.2* %	0.2 %	0.3 %		0.2 %	0.1 %			
enh	**11.5** %	**82.5** %	0.1 %				0.1 %			0.1 %
kca		0.1 %	**94.5** %		0.2 %					0.3 %
koi				**98.9** %	1.1 %					
kpv	0.1 %			6.9 %	**92.5** %					
mdf			0.7 %	0.7 %	0.4 %	**89.4** %		0.1 %	4.4 %	0.1 %
mhr			0.3 %	0.8 %			**95.6** %	1.7 %		0.2 %
mrj							0.2 %	**99.1** %		
myv			0.2 %	0.3 %	0.1 %	2.4 %	0.2 %	0.4 %	**92.7** %	0.5 %
udm			0.2 %	0.3 %	0.7 %	0.2 %	0.6 %	0.2 %	0.2 %	**95.2** %

Table 3: Confusion matrix of some of the Uralic languages written in Cyrillic script. As with the Table 2, the rows in this table would add to 100.0% if the figures for all the 285 languages would be shown.

3 Crawling the National Domains

In order to crawl for pages written in small Uralic languages, we use Heritrix [13], a web archiving system developed by the Internet Archive[5]. We chose to use Heritrix

[5] http://www.archive.org

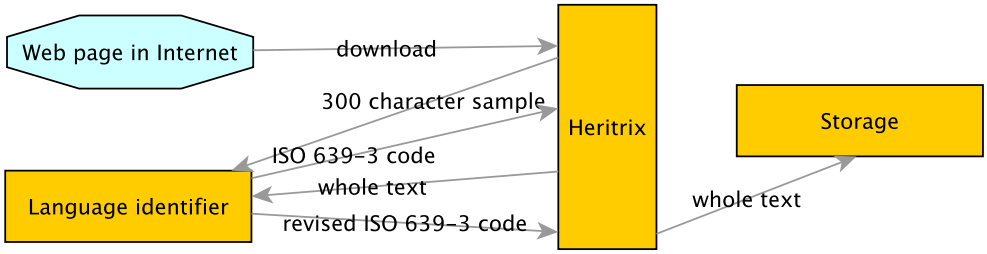

Figure 1: A diagram showing how Uralic web pages are processed during a crawl.

after considering several available crawlers. Heritrix is the outcome of many years of development by the Internet Archive and it is still being maintained. In the beginning it was a product of cooperation between Internet Archive and the Nordic national libraries and it is still used by several national libraries around the world to collect national web archives. It has also been successfully used for collecting similar corpora by [2].

The goal of the Internet Archive is to archive the sites as usable collections for future generations. In this project, we are only interested in collecting the textual material in the small Uralic languages. The version we are currently using downloads all text files as well as pdf files it finds from within the domain in question. We have made some custom changes to the code of the crawler so that, when a file has been downloaded, the running text is extracted from is. The crawler sends an excerpt of 300 characters from the middle of text to the language identifier which responds with the ISO-639-3 code of the language of the text. If the language is one of the Uralic languages we are interested in, the crawler sends the whole text to be re-identified. If this identification still points to a small Uralic language, the whole text of the page is archived. The address and the identification results of all crawled pages, including the ones rejected, are stored.

We have chosen to start collecting the material by crawling the national domains most likely to contain material written in small Uralic languages, i.e. .ee, .fi, .no, .ru, and .se. Table 4 shows the statistics for each of the five national domain crawls. The first column "**URLs**" indicates the total number of downloaded files during the crawl and the fourth column "**domains**" indicates how many subdomains were crawled. For the Russian crawl "**domains**" numbers only the top level domains as our crawling tactic had changed when the crawl started. The second column "**LI-1 URLs**" gives the number of pages identified to contain small Uralic languages during the crawl. The third column "**LI-2 URLs**" is the number of pages still identified as Uralic after a more

precise language identification which was done after the crawl. The fifth and sixth columns indicate the number of subdomains in Uralic languages before and after the more precise identification.

	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

Table 4: Statistics for the crawls of the five national domains.

In the following paragraphs we make a few notes of the individual national domain crawls. 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. Currently we have set the crawler to accept only up to 100 000 pages per top-domain. Even this does not allow us the luxury to just wait for the exhaustion of the queued URLs, as some of the sites are very slow to serve the pages and waiting for them to reach the page limit could take months or even years. We are now trying to determine if the speed of the crawl could be used as an additional indication of domain exhaustion. We could consider, for example, that if the hourly average speed drops below 10% of the average speed of the first week of the crawl, the domain is exhausted. As we have not yet stabilized our criteria for exhaustion, the current figures can not really be compared with each other and do not give a realistic picture of the size of the national domains.

Finnish .fi domain In the crawl of the Finnish internet we downloaded around 354 million files. The Finnish crawl was terminated as the crawler was running out of disk space and the speed had slowed down to around 25 pages per second. The average speed for the first week of the crawl was 282 pages per second, so we could consider the crawl exhausted. The Finnish language model used in the language identifier is derived from the Finnish Wikipedia, which is mostly written in the official form of written Finnish. However, many people do write texts using the forms of their respective dialects. The written forms of Tornedalen Finnish, Kven Finnish, and Ingrian are much closer to these written Finnish dialects than the official written Finnish. This creates a problem as a great number of texts in dialectal Finnish are identified as these three languages. This could be corrected by creating separate language models

for written dialects of Finnish.

Swedish .se domain In the crawl of the Swedish internet we downloaded around 308 million files and it was terminated after the speed of the crawl had slowed to around 20 pages per second, which is well below 10% of the 238 pages per second average for the first week of the crawl. In this crawl, the library systems which were localized for Northern Sami turned out to be a problem. Over 100 domains dedicated to library systems were found by the crawler, the largest of them, bibliotek.nora.se, with 2 322 pages in Northern Sami. Not only do the library catalogue links expire quickly, they usually include the same text over and over again. We will have to incorporate a double-checking mechanism before creating the link collections and corpora in order to avoid collecting the same text many times. Some methods for removing doubles are introduced in [2] and [3].

Norwegian .no domain, Russian .ru domain, and Estonian .ee domain All of the three crawls ended in problems with either software, hardware or the crawl strategy used. The national domains were far from exhausted by any criteria we have considered. As most of the pages written in Uralic languages have been found in our crawl of the Norwegian internet, we are including the statistics for these crawls in the Tables 4 and 5.

4 The link collection

The link collection that is available at the time this was written has been curated by hand from the pages of the .fi crawl[6]. 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 crowdsourcing platform to be able to get feedback from those who are more familiar with the languages. The links lead to the actual pages currently found on the internet, so it is certain that some of the links will break while time passes. We will not remove the broken links completely from the database, but move them elsewhere and, if possible, make links to corresponding pages in the Internet Archive.

Our goal is to make the creation of the link collection as automated as possible, avoiding manual link curation. The list of sites from the Finnish crawl available at the moment includes only the front page of some sites although more pages were found during the crawl. In the future all the links found when crawling will be in the list of

[6] http://suki.ling.helsinki.fi/sites

links. The greatest problems will arise from the pages, which are written in a correctly identified language, but are near-doubles of other pages as in the case of the Swedish library systems mentioned above.

5 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. The first column of Table 5 shows the number of unique lines identified as written in the respective language. The columns 2 and 3 show the total number of words and characters in these lines.

Even though the language identification used is state of the art, it is far from perfect. The collections have not been checked by experts in the corresponding languages, but some things are clear even for a layman. The three smallest collections Selkup, Nganasan, and Tundra Enets do not actually contain the intended language at all, but are mostly some sort of lists of model numbers in Cyrillic for Nganasan and Tundra Enets. The Selkup collection consists of pages frequented with the word ”Статья”, ”article”, which is a very frequent word in the training text used for Selkup[7]. The Ingrian collection contains mostly hyphenated or otherwise broken Finnish or Finnish dialects written as spoken. Especially south-western Finnish dialects seem to be identified as Ingrian. Northern Finnish dialects are identified as either Kven or Tornedalen Finnish. In order to fix these problems with dialectal Finnish, we will try to include separate language models for dialectal Finnish in the future. Some long lists of names from the Swedish crawl have been identified as Ludian and are now polluting the collection of the language. The Khanty collection is polluted by long lists of model numbers written in Cyrillic script.

Future work

Language identification methods will be further developed in order to improve the robustness of the language identifier we use. We will also try to enhance the language models in order to more efficiently distinguish small languages from various dialects

[7] www.yamalchild.ru/docs/konv_selkup.doc

	#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

Table 5: The number of lines, words, and characters in small Uralic languages after language identifying each individual line.

and to identify languages in multilingual documents. The material found during the already performed crawls will be of assistance for this.

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.

Acknowledgments

We are thankful for the support of the Kone Foundation and to Jack Rueter for sharing his invaluable resources in Finno-Ugric languages. We also thank the anonymous reviewers for their suggestions and references.

References

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

[2] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[3] Vladimír Benko. Aranea: Yet another family of (comparable) web corpora. In *Proceedings of 17th International Conference, TSD 2014*, pages 247–256, Brno, Czech Republic, 2014.

[4] Tommi Jauhiainen. Tekstin kielen automaattinen tunnistaminen. Master's thesis, University of Helsinki, Helsinki, 2010.

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

[6] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, 1994.

[7] Erik Tromp and Mykola Pechenizkiy. Graph-based n-gram language identification on short texts. In *Benelearn 2011 - Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning*, pages 27–34, The Hague, 2011.

[8] John Vogel and David Tresner-Kirsch. Robust language identification in short, noisy texts: Improvements to liga. In *The Third International Workshop on Mining Ubiquitous and Social Environments*, pages 43–50, Bristol, 2012.

[9] Josh King and Jon Dehdari. An n-gram based language identification system. The Ohio State University, 2000.

[10] Ralf D. Brown. Selecting and weighting n-grams to identify 1100 languages. In *Text, Speech, and Dialogue 16th International Conference, TSD 2013 Pilsen, Czech Republic, September 2013 Proceedings*, pages 475–483, Pilsen, 2013.

[11] Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. Language identification of short text segments with n-gram models. In *LREC 2010, Seventh International Conference on Language Resources and Evaluation*, pages 3423–3430, Malta, 2010.

[12] M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, editors. *Ethnologue: Languages of the world, seventeenth edition.* SIL International, Dallas, Texas, 2013.

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