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

Language Diversity in the Newspaper and Periodical Corpus of the National Library of Finland (NLF)

Tommi Jauhiainen
Jussi Piitulainen
Erik Axelson
Krister Lindén

We took a relatively large sample with which to experiment with different methods for determining the languages for the paragraphs and the texts. The sample contains texts digitized by the NLF from July 2013 to December 2021. The texts came in separate half-year packages, 17 in total. We created one sample for Training and one sample for Evaluation. For the Training sample, we randomly picked 1,000 of these bindings from each package. In the training sample, the number of texts, e.g. pages, is 241,360.

Initial #texts	179,324 comb. Languages	1st rule #texts	95,926 comb. Languages	2nd rule #texts	69,102 comb. Languages
17,027	fin	31,323	fin	45,880	fin
4,023	no language	4,080	*no language*	6,220	swe
2,176	swe	3,926	swe	5,039	fin, olo
1,439	fin, olo	3,206	fin, olo	4,670	no language
1,037	*cym, fin*	2,146	*fin, sgs*	2,658	fin, swe
797	fin, sgs	2,077	*cym, fin*	2,277	eng, fin
586	fin, swe	1,613	fin, swe	2,007	est, fin
543	*est, fin*	1,554	eng, fin	1,598	cym, fin
533	eng, fin	1,377	est, fin	1,527	*fin, smn*
479	fin, sms	1,297	fin, sms	1,514	fin, sms
458	*fin, tca*	1,060	fin, smn	1,113	nno, swe
441	fin, smn	955	*fin, slk*	1,027	dan, swe
356	*fin, sme*	912	~~fin, tca~~	970	*fin, pol*
328	fin, slk	813	fin, pol	950	fin, sme
327	fin, gsw	782	*fin, gsw*	941	fin, slk
316	fin, pol	727	nno, swe	918	*fin, sgs*
308	*ces, fin*	663	fin, sme	870	fin, gsw
276	*fin, mlt*	641	*ces, fin*	831	fin, hbs
274	*fin, izh*	601	dan, swe	687	fin, mlt
264	*fin, gla*	577	~~cym, fin, sgs~~	620	nob, swe
245	*sgs, swe*	569	fin, hbs	619	*fin, msa*
241	fin, hbs	568	*fin, kal*	607	*fin, izh*

The difference between the initial situation and after applying the 1st rule was the restriction of the number of languages per text. Example sentence affected: "A AAA ovh 4.D 9U !" identified as Czech (ces).

Rule 2: Minimum sentence length in alphabetical characters, example: "2,49 / kg" identified as Samogitian (sgs).

Rule 3: Minimum sentence length in words including alphabetical characters, example: "Nuori-" identified as Ingrian (izh).

3rd #texts	54,063 comb. Languages	4th rule #texts	15,071 comb. Languages	5th rule #texts	1,361 comb. Languages
55,705	fin	109,530	fin	167,937	fin
7,728	swe	16,818	swe	32,521	swe
5,469	*fin, olo*	6,576	fin, swe	8,782	*fin, swe*
5,074	no language	6,050	no language	7,798	no language
3,057	fin, swe	5,132	*fin, olo*	6,355	eng, fin
2,723	eng, fin	4,916	eng, fin	1,637	eng, fin, swe
1,955	est, fin	2,630	*est, fin*	1,172	eng
1,911	*cym, fin*	1,958	~~fin, sms~~	565	*fin, olo*
1,786	fin, sms	1,710	*nno, swe*	565	dan, swe
1,734	fin, smn	1,698	*fin, smn*	533	eng, swe
1,330	nno, swe	1,644	dan, swe	487	rus
1,225	dan, swe	1,456	*cym, fin*	487	bul, rus
1,127	fin, sme	1,169	~~fin, sme~~	391	*est, fin*
1,106	*fin, pol*	1,083	*fin, sgs*	364	fin, pol
1,092	*fin, slk*	976	*fin, mlt*	352	nno, swe
1,039	*fin, gsw*	939	fin, pol	272	fin, gsw
1,030	fin, sgs	837	*fin, nno, swe*	253	*fin, mlt*
956	*fin, hbs*	806	fin, gsw	252	*cym, fin*
768	fin, mlt	780	eng, fin, swe	246	nob, swe
717	*nob, swe*	761	eng	226	*fin, smn*
665	~~dan, nno, swe~~	760	fin, hus	226	fin, hus
621	*fin, gla*	760	*dan, fin, swe*	222	deu, fin
592	*fin, lus*	706	fin, hbs	206	*fin, hbs*
588	~~fin, ksh~~	690	nob, swe	202	*dan, fin, swe*

Rule 4: Minimum confidence score for a sentence, example: "Enjoy -03 46 000 km met hopea ." identified as Kölsch (ksh).

Rule 5: Combined confidence score and length in words, example: "STT HELSINKI" identified as Skolt Sami (sms).

Rule 6: Combined confidence score and length in words for sentences with no lowercase letter starting words, example: "Ekman , Hans Ernback Ja Monica Ekman ." identified as Swedish (swe).

Rule 7: Ratio of number of words to number of unique words, example: "• ■ fl ■ ■■■■■■ > » » » » fl » fl fl » » » » » » » » » » »" identified as Maltese (mlt).

Rule 8: Sentence confidence and length for languages with one similar identification in neighbourhood, example: "lii Hill Supi Moisioni ; » tn M) , Tumppra Puh. ! * ? ," identified as Inari Sami (smn).

Rule 9: Sentence confidence and length for languages without similar identification in neighbourhood, example "Heimo Holopainen ." identified as Livvi (olo).

Rule 10: Sentence confidence and length for languages without similar identification in neighbourhood, second to last sentence in text, example "NEU VO STOL lIT ON METALL ITEOLLISUUS ." identified as Swiss German (gsw).

Rule 11: Sentence confidence and length for languages without similar identification in neighbourhood, last sentence in text.

Rule 12: Sentence length for languages without similar identification in direct neighbours.

Rule 13: Sentence confidence comparison in middle of one other language.

Rule 14: No lowercase words in sentence without similar identification in direct neighbours.

Rule 15: Ratio of uppercase starting words to lowercase starting words without similar identification in direct neighbours.

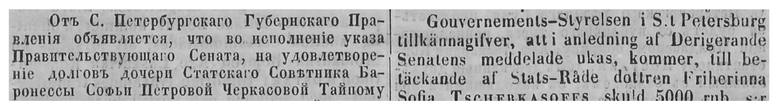
Repeating context sensitive rules 8 - 15.

For the Evaluation sample, from each package, we randomly picked 1,000 from those bindings that were not selected for the Training sample.

Train. #texts	comb. Languages	Eval. #texts	comb. Languages	All #texts	comb. Languages
180,145	fin	182,872	fin	363,027	fin
36,255	swe	34,909	swe	71,164	swe
14,228	no language	14,555	no language	28,768	no language
4,868	fin, swe	4,606	fin, swe	9,475	fin, swe
1,675	rus	1,742	rus	3,417	rus
1,285	eng, fin	1,400	eng, fin	2,685	eng, fin
1,261	eng	1,075	eng	2,336	eng
291	eng, fin, swe	282	eng, fin, swe	576	eng, fin, swe
267	eng, swe	281	eng, swe	548	eng, swe
103	deu	108	rus, swe	207	deu
85	deu, swe	104	deu	183	rus, swe
75	rus, swe	93	deu, fin	168	deu, fin
75	deu, fin	67	deu, swe	152	deu, swe
52	fin, fra	65	fin, fra	117	fin, fra
52	dan, swe	51	dan, swe	103	dan, swe
48	deu, eng, fin, swe	40	fra, swe	88	deu, eng, fin, swe
48	deu, eng, fin	40	eng, fin, fra, swe	86	eng, fin, fra, swe
46	eng, fin, fra, swe	40	deu, eng, fin, swe	85	deu, eng, fin
35	fra, swe	37	deu, eng, fin	76	fra, swe
34	fin, olo	34	fin, olo	68	fin, olo
34	est	23	eng, fra, swe	47	fin, sme
27	fin, sme	20	fin, sme	42	deu, fin, swe
23	fin, fra, swe	20	deu, fin, swe	41	fin, fra, swe
22	deu, fin, swe	19	nob, swe	41	eng, fra, swe
20	nob, swe	19	fin, lat	39	nob, swe
20	lat, swe	19	eng, fin, fra	39	fin, lat
20	fin, lat	18	fin, fra, swe	36	est
18	eng, fra, swe	18	est, fin	33	eng, fin, fra
14	eng, fin, fra	17	deu, eng, fin, fra, swe	31	lat, swe
13	est, fin	11	lat, swe	31	est, fin

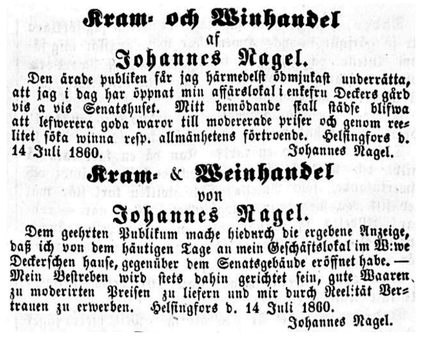
For reporting the decade level language use, we combined these samples.

1780	1800	1840	1850	1860	1870	1880	1890
swe	swe	swe	swe	swe	swe	swe	swe
	eng, swe	fin	rus, swe	rus, swe	fin	fin	fin
			rus	fin	rus, swe	fin, swe	fin, swe
				fin, swe	fin, swe	rus, swe	est
				deu, swe	lat, swe	lat, swe	deu, swe
				fra, swe		deu, swe	lat, swe
				fra		dan, swe	fra, swe


Отъ С. Петербургскаго Губернскаго Правленія объявляется, что во исполненіе указа Правительствующаго Сената, на удовлетвореніе долговъ дочери Статскаго Совѣтника Баронессы Софьи Петровой Черкасовой Тайному

Gouvernements-Styrelsen i S:t Petersburg tillkännagifver, att i anledning af Derigerande Senatens meddelade ukas, kommer, till betäckande af Stats-Räde döttren Friherinna Sofia Tscherkasoffs skuld 5000 rub. sar

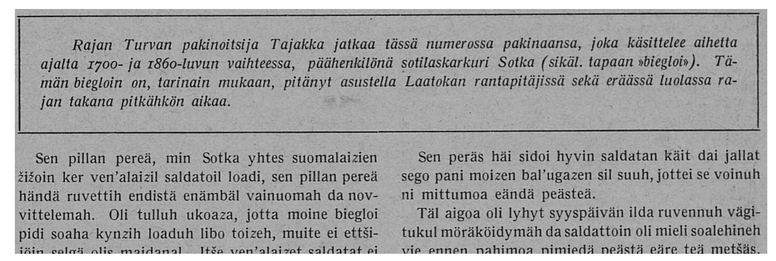
Russian - Swedish combination from "Tillkännagifvanden ifrån embets-myndigheterne i Wiborgs län, 14.03.1857, nr 12, p. 2"


Kram- och Winhandel
af
Johannes Nagel.
Den ärade publiken får jag härmedelst ödmjukast underrätta, att jag i dag har öppnat min affärslokal i enkefru Deckers gård vis a vis Senatshuset. Mitt bemödande skall städse blifwa att lefwerera goda waror till modererade priser och genom reellitet söka winna resp. allmänhetens förtroende. Helsingfors d. 14 Juli 1860.
Johannes Nagel.

Kram- & Weinhandel
von
Johannes Nagel.
Dem geehrten Publikum mache hiedurch die ergebene Anzeige, daß ich von dem häutigen Tage an mein Geschäftslokal im W:we Deckerschen hause, gegenüber dem Senatsgebäude eröffnet habe. — Mein Bestreben wird stets dahin gerichtet sein, gute Waaren zu moderirten Preisen zu liefern und mir durch Reellität Vertrauen zu erwerben. Helsingfors d. 14 Juli 1860.
Johannes Nagel.

German - Swedish combination from "Helsingfors Tidningar, 24.07.1860, nr 87, p. 4"

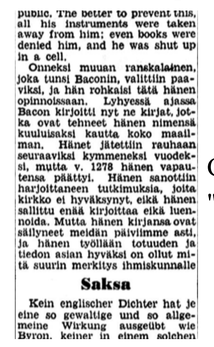
1900	1910	1920	1930	1940	1950	1960
fin	fin	fin	fin	fin	fin	fin
swe	swe	swe	swe	swe	eng	fin, swe
rus	rus	fin, swe	fin, swe	fin, swe	swe	eng
fin, swe	fin, swe	eng, fin	eng	eng	fin, swe	eng, fin
eng, fin	eng, fin	rus	eng, fin	rus	eng, fin	swe
deu	deu	eng	eng, swe	deu	deu, fin	fin, fra
deu, swe	fra, swe	eng, swe	deu, fin	eng, fin	fin, fra	eng, fin, swe
eng, swe	eng, swe	deu	deu, swe	dan, swe	eng, fin, swe	nob, swe
deu, fin	deu, swe	fin, fra	fin, olo	deu, fin	deu, eng, fin	deu, eng, fin
fin, fra	dan, swe	deu, fin	fin, fra	eng, swe	dan, swe	est, fin


Rajan Turvan pakinoitsija Tajakka jatkaa tässä numerossa pakinaansa, joka käsittelee aihetta ajalta 1700- ja 1860-luvun vaihteessa, päähenkilönä sotilaskarkuri Sotka (sikäl. tapaan »biegloi»). Tämän biegloin on, tarinain mukaan, pitänyt asustella Laatokan rantapitäjissä sekä eräässä luolassa rajan takana pitkähkön aikaa.

Sen pillan pereä, min Sotka yhtes suomalaizien žižoin ker ven'alaizil saldatoil loadi, sen pillan pereä händä ruvettih endistä enämbäl vainuomah da novvittelemah. Oli tulluh ukoaza, jotta moine biegloi pidi soaha kynzih loaduh libo toizeh, muite ei ettši-

Sen peräs häi sidoi hyvin saldatan käit dai jallat sego pani moizen bal'ugazen sil suuh, jottei se voinuh ni mittumoa eändä peästeä.
Täl aigoa oli lyhyt syyspäivän ilda ruvennuh vägitukul möräköidymäh da saldattoin oli mieli soalehineh vie ennen pahimoa pimiedä peästä eäre teä metsäs.

Finnish - Livvi combination from "Rajan turva : Sortavalan suojeluskuntapiirin äänenkannattaja, 01.11.1939, nr 11-12, p. 14"


public. The better to prevent this, all his instruments were taken away from him; even books were denied him, and he was shut up in a cell.
Onneksi muuan ranskalainen, joka tunsi Baconin, valittiin paaviksi, ja hän rohkaisi tätä hänen opinnoissaan. Lyhyessä ajassa Bacon kirjoitti nyt ne kirjat, jotka ovat tehneet hänen nimensä kuuluisaksi kautta koko maailman. Hänet jätettiin rauhaan seuraaviksi kymmeneksi vuodeksi, mutta v. 1278 hänen vapautensa päättyi. Hänen sanottiin harjoittaneen tutkimuksia, joita kirkko ei hyväksynyt, eikä hänen sallittu enää kirjoittaa eikä luennoida. Mutta hänen kirjansa ovat säilyneet meidän päivilimme asti, ja hänen työlään totuuden ja tiedon asian hyväksi on ollut mitä suurin merkitys ihmiskunnalle

Saksa

Kein englischer Dichter hat je eine so gewaltige und so allgemeine Wirkung ausgeübt wie Byron, keiner in einem solchen

German - English - Finnish combination from "Etelä-Suomen Sanomat, 05.09.1954, nr 203, p. 5"

1970	1980	1990	2000	2010
fin	fin	fin	fin	fin
fin, swe	swe	swe	swe	swe
eng	fin, swe	fin, swe	eng, fin	fin, swe
eng, fin	eng, fin	eng	fin, swe	eng, fin
eng, fin, swe	eng, swe	eng, fin	eng	eng, fin, swe
deu, fin	eng	eng, fin, swe	deu, eng, fin	eng
fin, pol	eng, fin, swe	eng, swe	eng, fin, swe	fin, sme
deu, eng, fin	fin, fra, swe	eng, fin, fra, swe	eng, fin, fra, swe	eng, swe
fin, fra	nob, swe	deu, eng, fin, swe	eng, fin, fra	eng, fin, fra, swe
est, fin	fra, swe	eng, fra, swe	eng, swe	fin, lat

German - English - Finnish - French - Swedish combination from "Etelä-Suomen Sanomat, 05.10.2001, nr 273, p. 35":

paluu (The Hitman's Run). Toimintaelokuva. USA 1999. 1.30 Hollywood Superstars: Tim Roth. 2.00 Murhaajan otteessa (Black & White). Jännityselokuva. USA 1998. 3.40 Gloria. Jännitysdraama. USA 1998. 5.30-7.00 B. Monkey Romanttinen trilleri. Englanti / Italia / USA 1998.

Canal+
8.00 James Caan. (Inside the Actors Studio: James Caan). Dokumentti. 8.50 Rakkautta ennen aamua. (Before Sunrise). Draama. USA 1995. 10.25 Steal This Movie! (Steal This Movie). Draama. USA 2000. 12.10 The Virginian. Lännenelokuva. USA 2000. 13.45 Quads!: Christmas Holidaze. Piirrossatiiri. Kanada 2001 14.10 Daydream Believers: The Monkees Story. Draama. USA 2000. 15.45 Harsh Realm: Inga fossa. Tieteissarja. 2000. 16.30 The Shadow Men. Tieteistrilleri. USA 2000. 18.00 Raja Hindustani. Rakkausdraama. Intia 1996. 20.05 Giorgio Armani. (Armani). Dokumentti. 2000 21.00 Amy ja Isabelle (ensi-ilta) (Amy and Isabelle). Draama. USA. 22.30 Gods and Monsters. Draama. USA 1998. 0.10 Looking for Alibrandi. Draama. Australia 2000. 1.50 Vihreä maili. (The Green Mile). Draama. USA 1999. 4.50 Harsh Realm: Three Percenters. Tieteissarja. 2000. 5.45 Kivisydän. (Hjärta av sten). Jännityselokuva. Ruotsi 2000. 7.05-7.57 Vahvempi voima: Tanska ja Chile. (A Force More Powerful: Denmark & Chile). Dokumentti. 2000.

Canal+ Kulta
8.00 The Last Days. Dokumentti. USA 1998. 9.25 Operaatio Mars. (Mission to Mars). Tieteis- ja toimintaelokuva. USA 2000. 11.15 Harsh Realm: Leviathan. Tieteissarja. 2000. 12.00 Quads!: The Fraud Quad. Piirrossatiiri. Kanada 2000 12.25 Sinisilmä-Mickey. (Mickey Blue Eyes). Romanttinen komedia. Englanti 1999. 14.05 James Dean: An Invented Life. Draama. USA. 15.40 Vihreä maili. (The Green Mile). Draama. USA 1999. 18.45 Seuralainen. (Mauvaise passe). Draama. Ranska 1999. 20.30 Veli 2. (Brat 2 (aka Brother 2)). Toimintaelokuva. Venäjä 2000. 22.35 Kun maailma ei riitä. (The World Is Not Enough).

RTL
7.00 Punkt 6. 7.30 Guten Morgen Deutschland. 8.00 Unter uns. 8.30 Gute Zeiten, schlechte Zeiten. 9.05 RTL Shop. 10.00 Punkt 9. 10.30 Meine Hochzeit. 11.00 Dr. Stefan Frank - Der Arzt, dem die Frauen vertrauen. 12.00 familien duell. 12.30 familien duell. 13.00 Punkt 12 - Das RTL-Mittagsjournal. 14.00 Die Oliver Geissen Show. 15.00 Bärbel Schäfer. 16.00 Der Schwächste fliegt. 17.00 Das Jugendgericht. 18.00 Die Nanny. 18.30 Unter uns. 19.00 Guten Abend RTL. 19.30 Exclusiv - Das Star-Magazin. 19.45 RTL aktuell. 20.10 Explosiv - Das Magazin. 20.40 Gute Zeiten, schlechte Zeiten. 21.15 Wer wird Millionär? 22.15 Mein Leben & Ich. 22.45 Ritas Welt. 23.15 Tage - 7 Köpfe. 0.15 Freitag Nacht News. 1.00 RTL-Nachtjournal. 1.30 Susan. 2.00 Veronica. 2.30 Der Mann an sich. 2.55 Mein Leben & Ich. 3.25 Ritas Welt. 3.50 RTL-Nachtjournal. 4.15 stern TV. 5.45 Freitag Nacht News. 6.15-6.30 Zeichentrickserie.

TV5
7.00 TV5, Le Journal. 7.15 Journal Mondial des Journaux. 7.30 Telematin. 9.00 Journal Radio-Canada. 9.30 Fete des Bebes. 10.00 TV5 Infos. 10.05 Zig Zag Cafe. 11.00 TV5, Le Journal. 11.15 Fiction: La kine. 13.00 TV5 Infos. 13.05 100% Questions. 13.30 Journal de France 3. 14.00 TV5 Infos. 14.05 Faxculture. 15.00 TV5, Le Journal. 15.15 Fiction: La kine. 17.00 TV5, Le Journal. 17.15 Le Journal de leco. 17.20 TV5, LInvite. 17.30 Griffe (RDI). 18.00 TV5 Infos. 18.05 Pyramide. 18.30 Questions pour un champion. 19.00 TV5, Le Journal. 19.15 Fiction: La kine. 20.55 Le Journal de leco. 21.00 Journal RTBF. 21.30 Journal France 2. 22.00 TV5 Infos. 22.05 Fiction Canadienne: Reseau (Serie 1). 23.15 Divertissement France 2. 1.30 Journal TSR. 2.00 Soir 3. 2.30 Le Canada Aujourdhui (RDI). 2.45 TV5 Infos. 2.50 On ne peut pas plaire a tout le monde. 4.00 TV5 Infos. 4.05 Fiction Canadienne: Reseau (Serie 1). 5.00 Journal RDI. 5.30 Journal Mondial des Journaux. 6.00 TV5 Infos. 6.05 TV5, LInvite. 6.30 Journal Mondial des Journaux.