Similecoj inter Radikoj, Bazita sur la Ofteco de Vortoj kiuj Uzas Ilin

Steve Eichblatt

21–a de januaro, 2019

1 Enkonduko

Esperanto estas tre regula lingvo. Vortoj estas konstruita el radikoj kaj modifaĵoj. Se oni imagas grandegan tabelon de ĉiu vorto konstuebla, t.e. ĉiu radiko kun ĉiu grupo de modifaĵo, ĉiuj vortoj estus laŭreglaj, sed la grandega parto estus neniam uzitaj. Vortoj kiel `dislegigon` aŭ `interlegatiĝi` ne havas sencon por homoj (Notu: anstataŭigu la radikon `leg` kun `kon`, kaj la du vortoj validiĝas.)

Do, la vortoj uzitaj, el ĉiuj la laŭreglaj ebloj, diras iun pri la homara sperto. En ĉi tiu raporto mi volas pripensi tiun fakton, kaj analizi la tekstojn esperante trovi kelkajn surprizojn.

2 Datumoj

Mi bezonis 2 fontojn de datumoj por ĉi tiu projekto, tekstaron de vortoj vere uzitaj, kaj ankaŭ liston de radikoj. La tekstaron elŝuteblan mi trovis ĉe `http://tekstaro.com/`. Por la listo de radikoj mi elŝutis la retan vortaron ĉe `http://reta-vortaro.de/tgz/index.html`, kaj mi facile tiris liston de radikoj el ĉi tiuj datumoj.

La tekstaro enhavas ĉirkaŭ 5 milionoj da vortoj, ĉirkaŭ 50 mil malsamaj vortoj. Kompreneble, la kvanto da radikoj estas multe malpli ol 50 mil, sed la vortaro ne indikas la radikojn de la vortoj.

3 Metodo

Ĉar la analizo bezonas la radikojn de ĉiu vorto, necesas programon por disigi la vortojn en pecojn. Feliĉe, pro la reguleco de Esperanto, tia programo estas relative facila por krei.

Ekzemple, la frazo "`malfeliĉe la tekstaro ne enhavas la radikojn de ĉiu vorto`" estiĝas "`mal,Feliĉ,e la tekst,ar,o ne en,Hav,as la radik,o,j,n de ĉiu vort,o.`" Notu la radiko estas ĉefliterita kiam ĝi ne komencigas la vorton.

La programo kiu faris tion estis relative simpla, kaj tute ne perfekta. Ĝia plej grava manko estas ke ĝi malkorekte disigas kombinaĵojn. Ekzemple, la vorto ‘‘matenmangi’’ estiĝas ‘‘matenmanĝ,i’’, anstataŭ ‘‘maten,Manĝ,i.’’ Feliĉe, la kombinaĵojn konsistigas malgrandan porcion el la tuta tekstaro. Se oni dezirus plibonigi la programon por trakti tiujn kazojn, mi taksas ke ĝi estus ebla sed malrapidega.

La programo konsideris nur vortojn kiuj aperas pli ol 3 fojoj en la tekstaro. Ankaŭ, se la programo disigis vorton en radikon kaj modifaĵon, kaj la modifaĵon aperas malpli ol 5 fojoj en la tuta tekstaro, ĉi tiu vorto estas forjetita.

De tie la programo trovis, ke estas 5,300 apartaj validaj radikoj, kaj 37 mil vortoj en la tekstaro konstruitaj el tiuj radikoj.

Tuj kiam oni havas la vortoj disigitaj, estas relative facila fari la analizon por trovi la similecojn inter radikoj. Ĉiu uzaĵo de vorto estas konsiderita kun radiko kaj modifaĵoj, kaj granda tabelo estas konstruita.

4 Analizo

4.1 Kvantoj

Oni povas tuj vidi kiujn el la 5,300 radikoj estas uzata en la plejmulta de vortoj. Jen la plejalta 52 (vidu tabelon 1) Ĉiu radiko devenas 7 vortjon, averaĝe. Ĉirkaŭ 1300 radikoj partoprenas en nur unu vorto (ekz. `kvankam, korpulent, ...`) La plej oftuzitaj radikoj partoprenas en ĉirkaŭ cent vortoj. Do oni vidas ke vortoj estas tre maldensa en la matrico de ebloj.

Kio estas la 109 vortoj konstruata el `kon`? Jen ili vicigata per ofteco: `kon,as kon,at,a kon,is re,Kon,is kon,at,a,j ne, Kon,at,a kon,i kon,at,iĝ,is re,Kon,i re,Kon,as kon,at,iĝ,i ek,Kon,i kon,at,o,j ne,Kon,at,o kon,ig,is kon,ig,i kon,at,e kon,at,o ek,Kon,is ne,Kon,at,a kon,o kon,at,a,j,n ne,Kon,at,a,n dis,Kon,ig,i ne,Kon,it,a kon,at,a,n kon,ant,e re,Kon,os ne,Kon,at,ul,o inter,Kon,a re,Kon,ebl,a kon,o,j,n re,Kon,u kon,ig,os kon,it,a kon,u kon,o,n kon,at,iĝ,o re,Kon,ant,e kon,at,ig,is re,Kon,o re,Kon,it,a kon,us kon,o,j ek,Kon,as re,Kon,ebl,a,j re,Kon,int,e kon,at,iĝ,u kon,ant,o kon,at,o,j,n kon,ig,u kon,at,iĝ,as ne,Kon,at,a,j,n ne,Kon,at,o,n ek,Kon,os ne,re,Kon,ebl,a dis,Kon,ig,o re,Kon,us dis,Kon,ig,is re,Kon,at,a kon,at,ig,i re,Kon,o,n ne,Kon,it,a,j kon,ig,as kon,os kon,ig,o ek,Kon,o dis,Kon,iĝ,is ne,Kon,at,o,j ek,Kon,u kon,it,a,j kon,at,ec,o kon,int,a re,Kon,int,a ne,Kon,o kon,at,o,n ek,Kon,int,e kon,ig,int,a kon,at,ec,o,n kon,iĝ,i inter,Kon,at,iĝ,o kon,at,iĝ,os kon,iĝ,is kon,ant,a kon,ant,a,j ek,Kon,us kon,iĝ,u kon,ig,it,a re,Kon,ebl,as dis,Kon,ig,o,n kon,int,e kon,at,in,o kon,at,iĝ,o,n kon,ad,o inter,Kon,at,iĝ,is kon,ant,o,j ne,Kon,it,a,n inter,Kon,at,iĝ,i ne,Kon,it,a,j,n`

radiko	N_{vortoj}	radiko	N_{vortoj}	radiko	N_{vortoj}	radiko	N_{vortoj}
ir	144	plen	75	liber	63	pens	59
ven	122	edz	72	sci	63	varm	58
don	121	skrib	72	tir	63	rid	57
kon	109	tim	70	lev	63	lum	56
labor	103	star	70	lig	62	romp	56
parol	97	ten	69	proksim	62	uz	55
am	96	met	68	fort	62	dorm	55
est	90	trov	68	kulp	61	memor	55
vid	88	prem	68	lern	61	dir	55
mort	87	mov	66	hav	61	flug	54
port	83	kur	65	aper	60	esperant	54
san	77	far	64	vetur	60	jun	54
viv	77	ferm	63	kompren	60	send	54

Tabelo 1: La radikoj uzata en multaj vortojn

`kon,iĝ,as ne,Kon,ad,o inter,Kon,a,n ne,Kon,at,ec,o dis,Kon,iĝ,i ne,Kon,ant,a dis,Kon,ig,as kon,ig,ant,e re,Kon,it,a,j dis,Kon,ig,ad,o.`

4.2 Interrilatoj

La vortoj organizita per radikoj ebligas nin taksi la similecon inter radikoj. Bazita sur la modifaĵoj uzita kun la du vortoj, kaj iliaj oftecoj, oni povas taksi empirie ilian similecon. Ĉar la program estas tro malrapida por kompari ĉiun vorton kun ĉiu alia vorto, mi komparis nur la 300 plej oftuzatan radikojn.

Ni komparu `ruĝ` kaj `blu`. Ili ŝajnas similajn, ĉu ne? La programo taksas ilian simileco ĉirkaŭ 80%. Kial? Nu, po 7%, ruĝ partoprenas en vorto kun `iĝ` au `ig`, ekz. `ruĝiĝas,ruĝiĝante`,k.t.p. La radiko `blu` nek iĝas nek igas. Klare, kun pligranda tekstaro, oni trovus tiujn vortojn, sed ne tre ofte.

Simile, `blu` kaj `verd` similas po nur 80% ankaŭ. Estas `verdaĵo,verduloj,verdeco`.

Fakte, `nigr` kaj `blank` eĉ pli similas, po 90%.

Kiuj estas la plej similaj radikoj? Per tiu analizo, estas la "atomvortoj", kiuj mem estas vortoj. Ekz. `sed, kvankam, baldaŭ, jes`, k.t.p, kiuj ĉiam, aŭ preskaŭ ĉiam aperas sen modifaĵojn.

Kiam ni forigi la atomvortojn el la tabelo, kaj serĉi similajn vortojn ni trovas, ke `nokt` kaj `vesper` la plej similaj. Tio estas kuraĝigante. Aliaj tre similaj paroj estas `(knab, vir), (infan, person), (mond, sun), (hor, monat), (hom, person), (lingv, popol), (famili, popol),`

`(oft, ĝust), (hor, tag), (est, hav), (pov, vol) (histori, lingv)`. Ĉi tiuj paroj certe diras ion pri la sperto homa!

Ni ŝategus kompari ĉiujn radikojn kun ĉiuj aliaj, sed tio estas tro malrapide. Tamen, ni povas uzi metodon nomita "spektra arigato" (angle: "spectral clustering") por rapidege arigi la radikojn en malgrandajn arojn. Tiam ni povas taksi la simileco inter la radikoj en ĉiu aro.

4.3 Radikaroj

Laŭ mi, la plej grava kriterio por disigi radikoj, kion ni jam vidis, estas se la radiko estas atomvortoj, se oni trovas multaj de vortoj konstruita el ĝi. Do, mi unue disigis la radikojn en 3 grupoj: la "vastaj", la "malvastaj" kaj la "atomoj". Vastaj radikoj faras peco de pli ol 7 vortoj (pli ol la averaĝo), kaj ne estas si mem vorto. Malvastaj radikoj faras peco de 7 vortoj aŭ malpli, kaj ankaŭ ne estas si mem vorto. Atomoj estas iam (aŭ ĉiam) si mem vorto. Atomoj ankaŭ devas aperi pli ol 100 fojo en la tekstaro.

4.3.1 Atomaj Radikoj

Tabelo 2 montras la grupoj el la atomaj radikoj per spektra arigato.

F	m_1	f_1	m_2	f_2	m_3	f_3	m_4	f_4	m_5	f_5	N_{tot}	plejoftaj radikoj
1	–	98	e	.	a	.	o	.	ege	.	68	la, kaj, de, en, ne
2	–	95	a	1	an	.	e	.	n	.	16	el, per, nur, ĉi, sin
3	–	90	a	2	e	2	ete	.	oj	.	11	kun, nun, dum, iom, jen
4	–	85	a	3	i	2	e	2	mal	1	6	pli, ĝi, plu, trans, hieraŭ
5	–	80	a	10	an	3	aj	2	e	1	8	mi, li, vi, ŝi, apud

Tabelo 2: La plej grandaj grupoj de atomaj radikoj

En tabelo 2, F indikas la numeron de la familio de radiko. m_n indikas la n-a plej oftan vortmodifaĵon, kaj f_n indikas ĝian oftecon. N_{tot} montras la tutan kvanton da radikoj en tiu familio. La kvin plej oftaj radikoj aperas en la plej dekstra kolumno.

Ĉiuj el tiuj grupoj havas – kiel plej ofta modifaĵo. Tio signifas, ke la radiko mem faras vorton. Grupo unu enhavas la radikojn `la,kaj`, *k.t.p.*, kiuj preskaŭ ĉiam estas senmodifaĵa. La sekvantaj grupoj en tabelo 2 havas pli kaj pli de aliajn modifaĵojn. Ni vidas, ke `mi, li, vi, ŝi` estas kune (estanta 80% atoma), sed `ĝi` estas pli ofta atoma. Notu, "si" trovas sin en la malvastaj radikoj, grupo 2; tiel malofte si aperas atome (ĉirkaŭ 12%).

4.3.2 Malvastaj Radikoj

Tabelo 3 montras la grupoj el la malvastaj radikoj per spektra arigato. La tabelo montras, ke la plej ofta modifajoj (en kolumno m_1) estas aŭ "a" aŭ "e" aŭ "is" aŭ "o" aŭ "oj". Do, la radikoj ariĝas en vortopecoj, kaj la plejparto de vortopecoj redividas en plurajn grupojn. Ni povas vidi, ke estas multegaj malvastaj radikoj, kaj la tabelo montras nur ĝiajn grandajn grupojn. Grupo 5 el tabelo ?? estas la plej granda, kaj la malplej interesa, havanta nur nomoj.

La -o finaĵo estas la plej ofta. Ni vidas, ke estas substantivoj kies dua plej ofta finaĵo estas -oj, kaj aliaj kies dua plej ofta finaĵo estas -on. Ni vidos tion ankoraŭ ĉe la vastaj radikoj.

F	m_1	f_1	m_2	f_2	m_3	f_3	m_4	f_4	m_5	f_5	N_{tot}	plejoftaj radikoj
1	a	99	aj	0	an	0	eco	0	oj	0	106	beat, jid, magr, ajmar, niz
2	a	36	e	16	aj	15	an	7	o	2	117	si, ĝeneral, konstant, sud, eventual
3	e	36	aj	28	a	8	an	6	as	1	110	kelk, subit, precip, plur, nepr
4	is	16	as	10	i	9	ojn	5	ado	4	382	ĉiu, foj, ekzempl, valent, ig
5	o	99	on	0	oj	0	a	0	as	0	735	johan, germani, fernand, franci, kristofor
6	o	80	on	6	oj	4	a	1	aj	0	152	faraon, petr, revu, viktor, litovi
7	o	70	on	23	oj	1	a	0	ojn	0	104	situaci, aer, komitat, brust, palac
8	o	59	oj	17	on	14	ojn	3	a	0	130	manier, poet, salon, numer, punkt
9	o	56	on	35	oj	1	ojn	1	e	0	105	mien, plank, frunt, spac, etos
10	o	53	on	7	a	6	e	4	aj	4	125	arme, moskv, pariz, rom, georg
11	o	43	on	25	oj	16	ojn	6	is	1	105	artikol, projekt, task, fraz, aŭt
12	o	40	oj	30	on	12	ojn	9	is	1	126	afer, templ, objekt, figur, event
13	oj	99	ojn	0	aj	0	on	0	o	0	155	juan, pice, nukleotid, flok, gulden

Tabelo 3: La plej grandaj grupoj de malvastaj radikoj

4.3.3 Vastaj Radikoj

Tabelo 4 montras la grupoj el la vastaj radikoj per spektra arigato. La vastaj radikoj estas la plej interesa grupo. Ĝia radikoj partoprenas en pli ol 7 vortoj, do ili havas multajn modifaĵojn.

Nun, ni povas kompari ĉiun radikon en tiuj relative malgrandaj grupoj, por vidi la plej similaj.

En grupo 2, la "adverbaj adjektivoj", la plej similaj radikparoj estas: `(brav, naiv)`, `(reciprok, simpl)`, `(brav, strang)`, `(malic, ĉef)` `(simpl, sincer)`, `(intim, serioz)`. La simileco inter tiuj paroj ĉirkaŭas 85%.

F	m_1	f_1	m_2	f_2	m_3	f_3	m_4	f_4	m_5	f_5	N_{tot}	plejoftaj radikoj
1	a	46	aj	15	an	8	e	5	ajn	3	55	ali, sol, propr, angl, sankt
2	a	31	e	30	aj	9	an	7	eco	2	38	bon, tut, sam, long, ĉef
3	a	27	aj	10	e	10	an	5	eco	4	55	grand, nov, bel, plen, grav
4	a	26	o	16	e	11	aj	9	an	5	34	fort, feliĉ, terur, saĝ, real
5	a	16	aj	6	e	5	mala	4	eco	4	33	jun, proksim, supr, liber, interes
6	anto	5	o	4	oj	3	a	3	ino	2	30	naci, san, mov, ofic, kapabl
7	e	50	a	11	aj	5	o	4	as	4	26	mult, ebl, ver, cert, rapid
8	e	18	as	16	a	14	is	7	aj	6	25	sekv, klar, facil, simil, neces
9	is	32	as	18	i	11	os	3	u	3	92	est, dir, pov, hav, dev
10	is	23	as	16	i	11	o	10	on	5	89	rigard, pens, komenc, sent, dezir
11	is	17	as	12	i	11	ado	3	ita	3	108	far, ven, ir, don, trov
12	is	13	oj	12	o	10	i	10	as	9	46	ag, rajt, serv, kant, organiz
13	is	9	i	6	as	5	ita	4	isto	3	81	kon, ferm, rid, ten, instru
14	o	67	on	11	oj	3	a	3	e	1	21	mond, moment, akv, princ, sun
15	o	55	oj	16	on	11	ojn	3	a	2	28	di, dom, voĉ, program, grup
16	o	51	on	19	oj	5	a	3	ojn	2	40	temp, sinjor, urb, kap, part
17	o	47	a	14	on	8	aj	5	e	5	29	vesper, nokt, histori, uson, pac
18	o	42	oj	21	on	12	ojn	6	a	2	42	lingv, tag, libr, ide, popol
19	o	37	on	14	ino	4	oj	4	is	2	45	esperant, viv, patr, fil, ŝip
20	o	36	oj	12	a	10	on	9	aj	4	30	lok, reĝ, ŝtat, famili, flank
21	o	31	a	19	e	10	on	8	aj	8	40	eŭrop, kultur, publik, natur, or
22	o	26	oj	14	on	10	ojn	6	ino	4	42	vir, amik, knab, frat, sign
23	o	26	on	12	is	9	as	7	oj	6	59	labor, nom, edz, tem, lum
24	o	17	e	14	on	8	a	8	as	3	35	fin, ĝoj, silent, hejm, rilat
25	o	15	is	11	as	11	on	7	i	7	68	mort, am, help, daŭr, tim
26	oj	45	o	17	ojn	15	on	5	aro	2	26	hom, jar, vort, okul, membr
27	oj	32	o	30	on	10	ojn	9	a	2	33	land, infan, man, person, pastr
28	oj	32	o	11	ojn	11	a	4	on	4	18	flor, genu, parenc, vers, frukt
29	oj	23	a	22	o	14	aj	11	ojn	5	15	scienc, detal, najbar, grek, individu
30	oj	17	o	17	on	7	ojn	7	aro	2	31	verk, arb, vest, kamp, paŝ

Tabelo 4: La grupoj de vastaj radikoj

En grupo 3, la "pluraj adjektivoj", la plej similaj radikparoj estas: `(dik, grand)`, `(dolĉ, gaj)`, `(dik, mol)`, `(gaj, larĝ)`, `(grav, pez)` La simileco inter tiuj paroj ĉirkaŭas 77%.

En grupo 5, la "malaj adjektivoj", la plej similaj radikparoj estas: `(amuz, interes)`, `(riĉ,`

`spirit)`, `(amuz, distr)`, `(financ, interes)` La simileco inter tiuj paroj ĉirkaŭas 55%.

En grupo 9, la "puraj verboj", la plej similaj radikparoj estas: `(est, vol)`, `(pov, vol)`, `(est, hav)`, `(ekzist, situ)`, `(ating, konstat)` La simileco inter tiuj paroj ĉirkaŭas 80%.

En grupo 11, la "ada verboj", la plej similaj radikparoj estas: `(kred, supoz)`, `(kompren, kred)`, `(detru, prepar)`, `(falĉ, plug)`, `(far, trov)`. La simileco inter tiuj paroj ĉirkaŭas 75%.

En grupo 13, la "ista verboj", la plej similaj radikparoj estas: `(pentr, skulpt)`, `(kurac, paŝt)`, `(instru, ĉas)`, `(mok, zorg)`, `(juĝ, ĉas)` La simileco inter tiuj paroj ĉirkaŭas 65%.

En grupo 17, la "adjektivaj substantivoj", la plej similaj radikparoj estas: `(mens, spirit)`, `(afrik, uson)`, `(printemp, vintr)`, `(nokt, vesper)`, `(printemp, turism)`. La simileco inter tiuj paroj ĉirkaŭas 85%.

En grupo 22, la "ino vortoj", la plej similaj radikparoj estas: `(knab, vir)`, `(scen, trezor)`, `(argument, reklam)`, `(farb, intervju)` La simileco inter tiuj paroj ĉirkaŭas 75%.

En grupo 26, la "pluraj substantivoj", la plej similaj radikparoj estas: `(soldat, jar)`, `(larm, okul)`, `(branĉ, poem)`, `(poem, vort)`, `(branĉ, foli)`, `(dent, okul)` La simileco inter tiuj paroj ĉirkaŭas 85%.

5 Konkludo

Ĉu ni lernis ion pri la sperto homa per tio analizo? Mi ne scias. Ni ja sukcesis trovi similecoj inter vortradikoj per la formoj de la vortoj kiu uzas ilin. Mi kredas, ke la strikta strukturo de Esperanto permesas nin malkovri tion. Mi dubegas, ke simila, tiel simpla analizo eblus angle, aŭ alia natura lingvo. Tiusence, ĝi montras al ni pluan aspekton de la beleco de Zamenhofa kreaĵo.

6 Teknikaj Rimarkoj

Tiu analizo estis farita uzanta la programlingvo "python". Se oni interesas plilabori la analizon, vidu `https://github.com/eichblatt/analyze_roots`. La plej grava parto de tio estas la vortkvanto, dosiero je 13 Mbajtoj, legebla uzante python.