Similarity Between Esperanto Roots, Based on the Frequency of Words Which Use Them

Steve Eichblatt

21–a de januaro, 2019

1 Introduction

Esperanto is a very regular language. Words are construction from roots and modifiers. If you imagine a huge matrix of every possible word contructed in this way, those words would all be valid Esperanto words. But the vast majority of them would never be used. Words like `dislegigon` or `interlegatiĝi`, with the root word `leg`, meaning "read", are senseless words in any language. But substitute the root `kon`, meaning "know", and the words translate to "to defamiliarize" and "to get to know one another".

So the subset of *used* words, from the set of all legal words, must say something about human experience. In this report I will explore this fact, and analyze the Esperanto corpus hoping to find some surprises.

2 The Data

I needed 2 data sources for this project, a corpus of used words and also a list of roots. I used the downloadable corpus at `http://tekstaro.com/`, and the online dictionary `http://reta-vortaro.de/tgz/inde` for the list of roots.

The corpus contans about 5 million words, with about 50,000 distinct words. Of course there are significantly fewer roots, but the corpus doesn't indicate the roots of the words.

3 Method

Because the analysis needs the roots of each word, I needed a program to decompose the words into roots and affixes. Fortunately due to the regularity of Esperanto, this was an easy program to write.

For example the sentence ‘‘malfeliĉe la tekstaro ne enhavas la radikojn de ĉiu vorto’’ becomes ‘‘mal,Feliĉ,e la tekst,ar,o ne en,Hav,as la radik,o,j,n de ĉiu vort,o.’’. Note the root is capitalized when it is not the beginning of the word.

The program which did this is relatively simple and not at all perfect. It's biggest shortcoming is that it incorrectly decomposes compound words. Thus the word ‘‘matenmangi’’ becomes ‘‘matenmanĝ,i’’, instead of ‘‘maten,Manĝ,i.’’. Fortunately, compounds words make up a small part of the entire corpus. If someone wanted to improve the program to handle these cases, it would be straightforward but slow.

The program considers only words which appear more than 3 times in the corpus. Also, if the program decomposes a word into roods and affixes, and the affixes appear less than 5 times in the corpus, that word is discarded.

From there the program finds that there are 5,300 valid roots, and 37,000 words in the corpus build from those roots.

Once you have the words decomposed, it is relatively easy to do the analysis to find the similarity between roods. Any use of a word is considered as a root and affix, and a large table is built.

4 Analysis

4.1 Quantities

One quickly sees which of the 5,300 roods occur in the most different words. Here are the highest 52 (see table 1)

Each root becomes 7 words on average. Around 1300 roots participate in only a single word. (eg. kvankam, korpulent, ...) The most often-used roots take part in around 100 words. Clearly the used words are very sparse in the matrix of possible words.

What are the 109 words build from the root kon? Here they are from most to least used:
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,j 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

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,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,`
`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 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

The words organized by roots enables us to estimate the similarity between the roots. Based on the affixes used with any 2 words, and their frequency, we can estimate empirically their similarity. Because the program is too slow to compare every pair of words, I compared only the pairs within the most frequent 300 roots.

Let's compare `ruĝ` and `blu`. They seem similar to our minds, no? The program estimates their similarity at around 80%. Why? Well, around 7% of the time, `ruĝ` is is a word with `iĝ` or `ig`, for example `ruĝiĝas, ruĝiĝante`, (=`reddens, reddenning` etc. The root `blu` doesn't "en", (we don't say `bluen`, although the concept exists). With a bigger corpus we might find words like `bluen` but it would not appear very often.

Similarity `blu` and `verd` are only 80% similary. Think `greenery`.

In fact `nigr` and `blank` are much more similar by this measure, about 90%.

So which are the most similar roots? By this analysis, it is the "atomic words" roots which are words in themselves. For example `sed`, `kvankam`, `baldaŭ`, `jes` (`but`, `however`, `soon`, `yes`), which always, or nearly always, appear without affixes.

When we exclude the atomic words from the table, and search for similar words we find that `nokt` (`night`) and `vesper` (`evening`) are the most similar of the roots. This is encouraging. Other very similar pairs are `(knab, vir)`, `(infan, person)`, `(mond, sun)`, `(hor, monat)`, `(hom, person)`, `(lingv, popol)`, `(famili, popol)`, `(oft, ĝust)`, `(hor, tag)`, `(est, hav)`, `(pov, vol)` `(histori, lingv)`. (that is: `(boy,man)`, `(child, person)`, `(world, sun)`, `(hour, month)`, `(man, person)`, `(language, people)`, `(family, people)`, `(often, correct)`, `(hour, day)`, `(be, have)`, `(can, want)`, `(history, language)`).

These pairs certainly seem to say something about the human experience.

We would really like to compare all the roots with all of the other roots, but the program is simply too slow. However, we can use a method called "spectral clustering" to quickly cluster the roots into groups. Then we can estimate the similarity within each cluster.

4.3 Root Clusters

I think that the most important criteria for separating roots, which we already saw, is if the root is an atomic word, that is, whether there are many different words made from it. So first I separate the roots into 3 groups: The "vast" the "narrow" and the "atomic". Vast roots take part in more than 7 words (which is more than the average), and never appear as a word without affixes. Narrow roots take part in 7 words or less, and also never appear as a word without affixes. Atomic roots sometimes (or always) are words themeselves. Atomic roots must also appear more than 100 times in the corpus.

4.3.1 Atomic Roots

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

In table 2, F indicates the root family. m_n indicates the n-th most common affix used with these roots, and f_n indicates its frequency. N_{tot} shows the total number of roots in the family. The 5 most often roots appear in the right column.

All of these groups have - as the most common affix. This means that the root is frequently a word. Group one has the roots `la`, `kaj`, (`the`, `and`) etc, which nearly always appear unmodified.

The next group in table 2 have more and more other modifications. We wee that `mi`, `li`, `vi`, `ŝi` (`I/me`, `he/him`, `you`, `she/her`) are together (being 80% atomic), but `ĝi` (`it`) is more often atomic. Note that the pronoun `si` (roughly `ones self`) is in the narrow roots, so rarely does it appear atomically (around 12%).

4.3.2 Narrow Roots

Table 3 shows the groups from the narrow roots by spectral clustering. The table shows that the most common affixes (columns m_1) are either "a" or "e" or "is" or "o" or "oj". So the roots cluster in word fragments, and most of the word fragments split into more groups. There are a great number of narrow roots and the table shows only its largest groups. Group 5 from the table `??` is the biggest, and the least interesting, containing only names.

The -o affix (basic nouns) is the most common. We see that there are some nouns whose second most common ending is -oj (plural), and others whose second most common ending is -on (accusative). We will see this again in the vast roots.

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 Vast Roots

Table 4 shows the groups from the vast roots by spectral clustering. The vast roots are the richest group to analyze. Its roots are part of more than 7 words so they have many word forms.

Now we can compare every root in these relatively small groups and see which are the most similar.

In group 2, the "adverbial adjectives", the most similare pairs are `(brav, naiv), (reciprok, simpl), (brav, strang), (malic, ĉef) (simpl, sincer), (intim, serioz)` (`(brave, naive), (reciprocal, simple), (brave, strange), (malicious, chief) (simple, sincere), (intimate, serious)`) Their similarities are around 85%.

In group 3, the "plural adjectives" the most similar pairs are `(dik, grand), (dolĉ, gaj), (dik, mol), (gaj, larĝ), (grav, pez)` (`(fat, bit), (sweet, gay), (fat, soft), (gay, large), (grave, heavy)`). The average similarity between these pairs is about 77%.

In group 5, the "opposites adjectives" the most similar are `(amuz, interes), (riĉ, spirit), (amuz, distr), (financ, interes)` (`(amuse, interest), (rich, spirit), (amuse, distract), (finance, interest)`). Their similarity is only about 55%.

Group 9, the "pure verbs" has the most similar pairs as `(est, vol), (pov, vol), (est, hav), (ekzist, situ), (ating, konstat)` (`(be, want), (can, want), (be, have), (exist, locate), (attain, state)`) Their similarity is about 80%.

Group 11, the "ongoing verbs", has the most similar pairs as: `(kred, supoz), (kompren, kred), (detru, prepar), (falĉ, plug), (far, trov).` (`(believe, suppose), (understand, believe), (destroy, prepare),(mow, plough), (do, find)`). Their average similarity is about 75%.

In grupo 13, la "ist verbs", the most similar pairs are: `(pentr, skulpt), (kurac, paŝt), (instru, ĉas), (mok, zorg), (juĝ, ĉas)` (`(paint, sculpt), (cure, feed), (instruct, chase), (mock, care for), (judge, chase)`). The similarity here is around 65%.

In group 17, the "adjectival nouns", the most similar pairs are `(mens, spirit), (afrik, uson), (printemp, vintr), (nokt, vesper), (printemp, turism)` (`(mind, spirit), (africa, USA), (spring (season), winter), (night, evening), (spring (season), tourism)`). Similarity withing this group is around 85%.

In group 22, the "feminizable words", the most similar pairs are: `(boy, man), (scene, treasure), (argument, reclaim), (paint, interview)` The similarity here is around 75%.

Group 26, the "plural nouns", the most similar pairs are: `(soldat, jar), (larm, okul), (branĉ, poem), (poem, vort), (branĉ, foli), (dent, okul)` (`(soldier, year), (tear, eye), (branch, poem), (poem, word), (branch, leaf), (tooth, eye)`). Similarity around 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

5 Conclusion

Did we learn something about the human experience from this analysis? I am not sure. We did succeed in finding similarity between word roots from the forms of the words which use them. I

believe that the strict structure of Esperanto allow us to discover that. I highly doubt that such a simple analysis would be possible in English or any other natural language. In this sense, it shows another aspect of the beauty of Zamehof's creation.

6 Technical Remarks

This analysis was done using the Python programming language. In you are intersted in extending the analysis, see `https://github.com/eichblatt/analyze_roots`. The most important part of this is the word frequency table, a file of about 13 MB, readable in python.