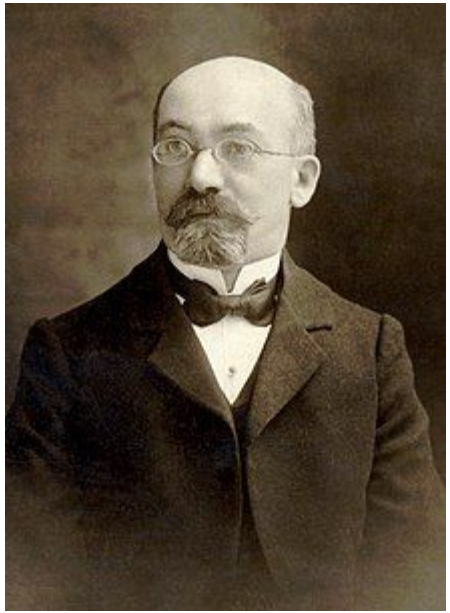
Root Word Analysis in Esperanto

Steve Eichblatt

Feb 22, 2019

About Esperanto



Ludvik Zmanhof
Or “Dr. Esperanto”

Esperanto, or the International Language, was invented by Ludwik Zamenhof, of Bialystok in 1887 as an easy-to-learn second language to bridge the language barrier.

The language is very regular with few rules and virtually no exceptions.

Esperanto words are constructed from root + affixes, and roots may be combined to form compound words.

All Esperanto roots and affixes are “atomic”, or indivisible.

Some Simple Esperanto

Prefixes: mal- = opposite (“un”); ne- = not; al- = toward; en- = in; eks- = former;

Suffices: -o = noun; -a = adjective; -e = adverb; -j = plural; -n = accusative;

-ig = to make into; -iĝ = to become;

-ist = one who does; -an = member of a group; -ar = collection of

-ec = trait of (“-ness”); -em = tendency to; -ind = worthy;

-ej = place of; -uj = container of; -id = child of; -in = feminine; etc.

Verb system: -i = infinitive; -as = present; -is = past; -os = future; -us = conditional;

-u = imperative; -ant = doing; -int = was doing; -ont = will do;

-at = being done ; -it = was done; -ot = will be done.

Note that some affixes modify meaning, and some indicate grammatical role.

The Idea

Esperanto words are constructed from a root and optional prefixes, and suffixes. Any root may take any set of prefixes or suffixes to make a word. All possible combinations are valid words.

You can imagine a huge table of every possible word, with a root in each row, and all possible affixes as columns (up to a finite number of them).

But only a subset of this table will be words that are actually used. That subset must tell us something about what humans experience.

So I wanted to build this table and explore the similarity of ideas between roots (rows in the table) which use the same affixes with similar frequency.

The Data

To build the table of all roots and their affixed I used 2 data sources available online.

A word list, or dictionary : http://reta-vortaro.de/tgz/index.html

A text corpus at http://tekstaro.com

The corpus contains about 5M words of Esperanto literature, which is about 50k different words, of which 37k are in the dictionary. These use about 5300 different roots. The authors of these works have many different native languages (not entirely European).

Each root is used in 7 words on average. About 1300 roots only participate in a single word (eg. “kvankam” (=although) appears only as an unmodified root, and “korpulent” appears only as the adjective, “korpulenta”).

The most widely-used roots participate in about 100 words.

Method

I wrote a program to divide the words into their pieces (“atoms”). The program atomized every word in the corpus, creating a data table with each word in a cell with the root as the row, and the set of affixes as the column.

Esperanto’s regularity makes this program surprisingly easy to write.

The atimization of a sentence:“**Mi verkis programon por disigi vortojn en iliajn pecojn**” Becomes: “**mi verk,is program,o,n por dis,Ig,i vort,o,j,n en ili,a,j,n pec,o,j,n**”

The root is identified as either the first, or the capitalized part of each word. This simple program cannot separate compound words.

The table of root frequencies forms the basis of the analysis.

A (Tiny) Subset of the Root Frequency Table

root	NENIU	o	o,j	o,n	o,j,n	a	a,j	ad,as	ad,i	ad,is	ad,o	ad,o,j
impon						50	7					
akr						103	47					
ordinar						174	109					
modern						205	90					
gist			4									
trakt		3		3						3	37	11
fuĝ		8										
freŝ		8				161	54					
ĵaluz		15		3		19						
aprob		26	4	23		5						
klopod		31	79	12	25						6	
vetur		43	5	12	3			3		5	29	
fid		50		30								
daŭr		158		5		50	5				8	
koler		159		53		75	16					
hotel		195	31	22	3	6						
ombr		264	67	70	20	9	6					
paf	15	18	11	6	5					5	12	
antaŭ	4523	4		3		233	148					
post	5949					110	57					
ne	38688	4				7						

Root Similarity Measure

From the root frequency table, I can estimate a measure of similarity of any pair of roots.

For example, comparing **ruĝ** (red) and **blu** (blue):

root_1	root_freq_1	mods_1	modstring	root_2	root_freq_2	mods_2
0	ruĝ	0.329235	[a]	a	blu	0.389831	[a]
1	ruĝ	0.233607	[a, j]	aj	blu	0.271186	[a, j]
2	ruĝ	0.109290	[iĝ, is]	iĝis	blu	0.008475	[iĝ, is]
3	ruĝ	0.094262	[a, n]	an	blu	0.129944	[a, n]
4	ruĝ	0.053279	[a, j, n]	ajn	blu	0.076271	[a, j, n]
5	ruĝ	0.043716	[o]	o	blu	0.036723	[o]
6	ruĝ	0.038251	[e]	e	blu	0.033898	[e]
12	ruĝ	0.006831	[et, a]	eta	blu	0.016949	[et, a]
13	ruĝ	0.005464	[o, n]	on	blu	0.011299	[o, n]
15	ruĝ	0.004098	[et, a, j]	etaj	blu	0.016949	[et, a, j],

root_1	root_freq_1	mods_1	modstring	root_2	root_freq_2	mods_2
7	ruĝ	0.024590	[iĝ, as]	iĝas	NaN	NaN	NaN
8	ruĝ	0.017760	[iĝ, ant, e]	iĝante	NaN	NaN	NaN
9	ruĝ	0.010929	[ul, o, j]	uloj	NaN	NaN	NaN
10	ruĝ	0.009563	[iĝ, i]	iĝi	NaN	NaN	NaN
11	ruĝ	0.006831	[ig, is]	igis	NaN	NaN	NaN
14	ruĝ	0.004098	[ig, as]	igas	NaN	NaN	NaN
16	ruĝ	0.004098	[et, a, n]	etan	NaN	NaN	NaN
17	ruĝ	0.004098	[is]	is	NaN	NaN	NaN
18	NaN	NaN	NaN	as	blu	0.008475	[as]

Similarities = 1-sum(abs(rf1-rf2))
= 0.81

Differences = sum abs(rf1-rf2)
= 0.05

Spectral Clustering

Since my computer isn't powerful enough to directly measure the similarity between every single pair of roots, I used "spectral clustering" to put the roots into clusters, and then compared the words within each cluster.

This naturally cut the set of root words into groups of fairly similar roots.

I noticed that the principal component that clustering was finding was the number of different words that played a part in similar numbers of different words.

Classes of Roots

I divided the roots into 3 classes: “Atomic”, “narrow” and “broad”.

Noble Roots: Sometimes (or always) form their own word. Eg: **la, kaj, de, en, mi, li, vi, ŝi.**

Narrow roots: take part in 7 words or less. Eg: **kelk, subit, manier.**

Wide Roots: take part in more than 7 words. Ekz: **bon, fort, mult, est, rajt, mond, voĉ.**

Within each class, I measured the similarity of each root with every other root in the class.

For comparing similarity, the wide roots are the most interesting class.

Most Widely Used Roots

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

Words Build from “kon”

Here is a list of many of the Esperanto words which have “kon”, (“to be familiar with”) as the root:

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 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 re,Kon,int,e ….

Noble 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	–	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 da atomaj radikoj

Narrow 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 da malvastaj radikoj

Wide 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	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

Most Similar Roots

Now we have classes that are small enough to compare each root within the class.

Similarities of (**brav**, **naiv**), (**simpl**, **sincer**), (**intim**, **serioz**), (**lingv**, **popol**)are around 85%.

(**dik**, **grand**), (**dolĉ**,**gaj**), (**grav**, **pez**) ĉirkaŭas 77%.

(**est**, **vol**) (**pov**, **vol**), (**est**, **hav**), (**ekzist**, **situ**), (**ating**, **konstat**) : 80%

(**kred**, **supoz**), (**kompren**, **kred**), (**detru**, **prepar**), (**falĉ**, **plug**), (**far**, **trov**): 75%

(**pentr**, **skulpt**), (**kurac**, **paŝt**), (**instru**, **ĉas**), (**mok**, **zorg**), (**juĝ**, **ĉas**): 65%

(**soldat**, **jar**), (**larm**, **okul**), (**poem**, **vort**), (**branĉ**, **foli**), (**dent**, **okul**): 85%

This is not surprising, considering that word representations are determined from frequencies of neighboring words.

The Human Experience?

I think that many of the root comparisons tell interesting stories which we, as humans, understand. Here are some differences between very similar roots:

(brav, naiv) -- naivulino is 2% of naive, but bravulino never appears.

(lingv, popol) -- Diffs: lingvaĵo (1.5%), popolaĉo (3%).

(dent, okul) -- Diffs: dentegoj (3%), okuletoj (.4%)

(vizaĝ, voĉ) -- Diffs: senvoĉe (voicelessly).

(hor, tag) -- Diffs: tagiĝo -- (daybreak).

Conclusion

The regularity and grammatical richness of Esperanto word construction makes Esperanto easy to analyze in terms of root words.

A much larger corpus would probably lead to more conclusive studies.

Similar roots do appear with the same affixes with similar frequencies. So some of the meaning of a root may be inferred from the affixes it appears with (if any).

Spectral Clusters show that the frequency of different grammatical roles often distinguish roots.

Without looking beyond each word, we can infer some meaning to many Esperanto roots.

https://github.com/eichblatt/analyze_roots