Take my Word for it:

a technique (software tool) for “averaging” words

Richard Forsyth,
www.richardsandesforsyth.net

Why am I here?

- To attempt to explain a technique that I believe has relevance to an idea important to Zamenhof & in interlinguistics generally
- Minor and technical in itself, but possibly useful to grander developments
  - [Early pioneers didn’t have today’s computational resources. Research question: Could that make a worthwhile difference?]

“Poles apart”?

An Approach to an Intercommunicative Vocabulary (Lingloss)

1. Identify a list of core semantic units (LESU) : conceptual slots to be filled (signiferaro ;)
2. Find expressions (alternatives?) for these items in a selection of existing languages
3. Find the most ‘typical’/’central’ among these by a computable procedure
4. Seek to post-optimize results by simple string manipulations

- [More on 3 & 4 than 1 & 2.]

Provisional attempt at LESU (working prototype)

- 6 identified sources (anglophone bias admitted!):
  - □ C.K.Ogden, 1937
  - □ Helen Eaton, 1940
  - □ Lancelot Hogben, 1963
  - □ Macmillian dictionary defining words, 2002
  - □ Longman dictionary defining words, 2003
  - □ Nerriére & Hon, 2009

Provisionally accept any word in 2 of 6 lists

- Present examples from words occurring in 5 of 5 (Eaton’s book didn’t arrive in time)
- 481 five-timers (English words)
    - Reduced to root forms
    - (very few inflected forms among these 481)
- 533 after using Esperanto for disambiguation
    - Basic/common words often polysemous
- Thus a sematic unit has a 2-part label, e.g.
    - (second, dua) ; (second, sekundo)

A Plenitude of Polysemy

- Many cases where Esperanto helps distinguish senses that need distinction
  - □ (fly, flugi); (fly, muŝo)
  - □ (light, lumo); (light, malpeza)
  - □ (right, dekstra); (right, prava) ; (right, rajto)
  - □ (watch, horloĝo); (watch, rigardi)
  - □ (wood, arbaro); (wood, ligno)
- Few cases of the reverse (but detectable by worksheet sort if required)
  - □ (among, inter) = (between, inter) [??]
  - □ (big, granda) = (great, granda) = (large, granda)

An Extract from the Trial List

wind	5 wind	vento
wind	5 wind	volvi
window	5 window	fenestro
wine	5 wine	vino
wing	5 wing	alo
wire	5 wire	drato
with	5 with	kun
woman	5 woman	virino
wood	5 wood	arbaro
wood	5 wood	ligno
wool	5 wool	lano
word	5 word	vorto
work	5 work	laboro
year	5 year	jaro
yellow	5 yellow	flava
yes	5 yes	jes
you	5 you	vi
young	5 young	juna

Next step, equivalents in other languages (e.g. Magyar)

term	freq	root	eo	paralogs
under	5	under	sub	alatt
unit	5	unit	unuo	egység
up	5	up	supre	fel
use	5	use	uzi	használ
value	5	value	valoro	érték
very	5	very	tre	nagyon
view	5	view	opinio	nézet
view	5	view	vidaĵo	látvány, kilátás
voice	5	voice	voĉo	hang
walk	5	walk	marŝi	sétál
wall	5	wall	muro	fal
war	5	war	milito	háború
warm	5	warm	varma	meleg
waste	5	waste	malŝparo	pocsékolás
waste	5	waste	rekremento	szemét, hulladék
water	5	water	akvo	víz

Next step, equivalents in other languages (e.g. Maori)

term	**freq**	**root**	**eo**	**paralogs**
under	5	under	sub	i raro
unit	5	unit	unuo	tētahi
up	5	up	supre	ki runga
use	5	use	uzi	tango
value	5	value	valoro	utu
very	5	very	tre	pū, rawa
view	5	view	opinio	whakaaro
view	5	view	vidaĵo	tirohanga
voice	5	voice	voĉo	reo
walk	5	walk	marŝi	wāke
wall	5	wall	muro	pa`tu`
war	5	war	milito	pakanga
warm	5	warm	varma	mahana
waste	5	waste	malŝparo	maumau, moumou
waste	5	waste	rekremento	otaota
water	5	water	akvo	wai

Next to the interesting bit

- How to find **average/centroid/median** of word-strings? (Representative?)
  - [Loglan (Brown, 1960) had similar idea, but details murky & words distorted by rigid phonetic preconditions.]
- ('young', 'juna') ==> ['juna', ‘jovem’, 'joven', 'giovane', 'iuvenis']
- Democratic??

Plenty of (numeric) distance measures in computing/stats

- Two to note:
  - Levenshtein distance/similarity
    - [consult Wikipedia et cetera.]
  - Czekanowksi distance/similarity (1909)
    - Statistical Anthropologist & Computational Linguist
      - Polish Professor at Poznań !
    - Czeksim(a,b) = 2 * ∑min(a[i],b[i]) / ∑(a[i]+b[i])
      - (where a and b are vectors of feature counts)

Czeksim for strings

- Mean of czeksim for 1,2,3,4-grams:

grande	**greatly**		
letters	**digrams**	**trigrams**	**tetragrams**
a 1 1	an 1 0	and 1 0	ande 1 0
d 1 0	at 0 1	atl 0 1	atly 0 1
e 1 1	de 1 0	eat 0 1	eatl 0 1
g 1 1	ea 0 1	gra 1 0	gran 1 0
l 0 1	gr 1 1	gre 0 1	grea 0 1
n 1 0	ly 0 1	nde 1 0	rand 1 0
r 1 1	nd 1 0	ran 1 0	reat 0 1
t 0 1	ra 1 0	rea 0 1	0 7
y 0 1	re 0 1	tly 0 1	
4 13	tl 0 1	0 9	
	1 11		

0.1993007 [0.6153846, 0.1818182, 0.0, 0.0]

czeksim = 0.1993007   levensim = 0.4615385

Dealing with diacritics

```
virino ['wahine'] ==>
virino wahine 0.1333
meansim = 0.1333
bestsim = 0.1333

flava ['kōwhai'] ==>
flava ko`whai 0.0417
flava kowhai 0.0455
meansim = 0.0436
bestsim = 0.0436
```

```
malseka ['ma`ku`'] ==>
malseka ma`ku` 0.1608
malseka maku 0.1919
meansim = 0.1764
bestsim = 0.1764

drato ['drót', 'huzal'] ==>
drato dro`t 0.2625
drato drot 0.2937
meansim = 0.2781
drato huzal 0.05
meansim = 0.05
bestsim = 0.2781
```

Dealing with alternatives (maximean matching method)

```
vidaĵo ['látvány', 'kilátás'] ==>
vidaj`o la`tva`ny 0.0938
vidaj`o latvany 0.0714
vidajo la`tva`ny 0.0667
vidajo latvany 0.0769
meansim = 0.0772
vidaj`o kila`ta`s 0.0938
vidaj`o kilatas 0.0714
vidajo kila`ta`s 0.0667
vidajo kilatas 0.0769
meansim = 0.0772
bestsim = 0.0772
```

```
rekremento ['szemét', 'hulladék'] ==>
rekremento szeme`t 0.2228
rekremento szemet 0.2381
meansim = 0.2304
rekremento hullade`k 0.0526
rekremento hulladek 0.0868
meansim = 0.0697
bestsim = 0.2304

rapida ['fast', 'quick', 'rapid', 'swift'] ==>
rapida fast 0.05
meansim = 0.05
rapida quick 0.0455
meansim = 0.0455
rapida rapid 0.8638
meansim = 0.8638
rapida swift 0.0455
meansim = 0.0455
bestsim = 0.8638
```

Finally, we can pick from several languages (eo, pt, es, it, la)

```
term : ('wall', 'muro') ==> ['muro', 'parede', 'pared', 'muro', 'murus']
*  0.3788 muro
   0.3788 muro
   0.2562 pared
   0.2523 parede
   0.2286 murus
['muro', ('parede', 0.05), ('pared', 0.0555556), ('muro', 1.0), ('murus', 0.4095238),
0.3787698]

term : ('war', 'milito') ==> ['milito', 'guerra', 'guerra', 'guerra', 'bellum']
*  0.5208 guerra
   0.5208 guerra
   0.5208 guerra
   0.0833 bellum
   0.0208 milito
['guerra', ('guerra', 1.0), ('guerra', 1.0), ('bellum', 0.0833333), ('milito', 0.0), 0.5208333]
```

Less clear-cut examples, czeksim

```
term : ('waste', 'malŝparo') ==> ['malŝparo', 'desperdício', 'derroche', 'spreco', 'effusio']
*  0.3157 desperdicio
   0.1685 desperdício
   0.1553 spreco
   0.1290 derroche
   0.1098 malŝparo
   0.0963 effusio
['desperdicio', ('desperdício', 0.8923427), ('derroche', 0.2167183), ('spreco', 0.2098039), ('effusio',
0.1423611), ('malŝparo', 0.1173375), 0.3157127]
```

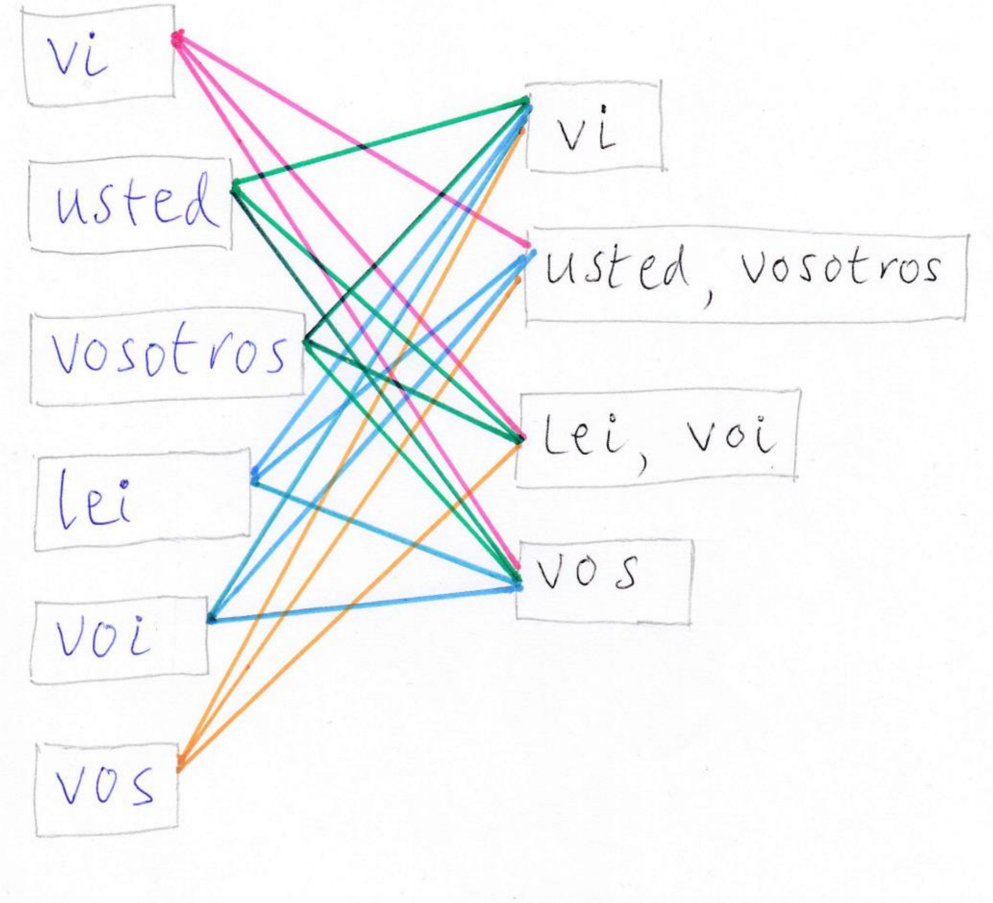
```
term : ('waste', 'rekremento') ==> ['rekremento', 'lixo', 'basura, desecho', 'spazzatura', 'quisquiliae']
*  0.0977 spazzatura
   0.0972 basura
   0.0722 quisquiliae
   0.0547 desecho
   0.0494 rekremento
   0.0466 lixo
['spazzatura', ('lixo', 0.0), ('basura, desecho', 0.2693452), ('quisquiliae', 0.0714286), ('rekremento',
0.05), 0.0976935]
```

Reassurance about multiple choices, e.g. lei, voi

term : ('you', 'vi') ==> ['vi', 'usted, vosotros', 'lei, voi', 'vos']

```
*  0.3380 vos
   0.3281 voi
   0.2402 vosotros
   0.2333 vi
   0.0944 lei
   0.0556 usted
```

['vos', ('usted, vosotros', 0.4252044), ('lei, voi', 0.3888889), ('vi', 0.2), 0.3380311]



What do we get? A provisional vocabulary (e.g. 12 high scores)

```
0.8070346 ('wind', 'vento')          :: vento
0.7395833 ('wave', 'ondo')           :: onda
0.7331845 ('wool', 'lano')           :: lana
0.7186508 ('wine', 'vino')           :: vino
0.6923070 ('value', 'valoro')        :: valor
0.6715090 ('way', 'metodo')          :: metodo
0.6558409 ('when', 'kiam')           :: quando
0.6527778 ('wing', 'alo')            :: ala
0.6139890 ('will', 'volo')           :: volontad
0.5625316 ('view', 'vidaĵo')         :: vista
0.5533333 ('yes', 'jes')             :: si
0.5490114 ('week', 'semajno')        :: semana
```

Provisional vocabulary (e.g. 12 low scores)

```
0.2497572 ('yellow', 'flava')        :: amarillo
0.2467314 ('word', 'vorto')          :: palavra
0.2404815 ('way', 'vojo')            :: camino
0.2323954 ('wire', 'drato')          :: filo
0.2255952 ('up', 'supre')            :: su
0.2194940 ('who', 'kiu')             :: quis
0.1915040 ('warm', 'varma')          :: calidus
0.1599303 ('wet', 'malseka')         :: molhado
0.1434028 ('very', 'tre')            :: muito
0.1387085 ('wood', 'arbaro')         :: bosque
0.1135101 ('woman', 'virino')        :: femina
0.0976935 ('waste', 'rekremento')    :: spazzatura
```

Proof of concept; still leaves much work to be done, e.g.

- List of 2/6 terms not fully compiled
  - 2300 roots leading to 3000-4000 items after disambiguation
- Will require serious attention to ‘paralogs’
  - collaboration
- Other similarity functions need to be tested
- Optimization by individual terms insufficient
  - Multi-objective optimization?!

Leaves many questions unanswered, e.g.

- Which source languages to use
  - Does ‘weighting’ make sense?
- What to do with low scorers
  - Perhaps drop lower-priority sources till score acceptable
- How well does string similarity correlate with learnability?
- How to merge ‘general vocabulary’ with semantic-domain subsystems?
- Phonetics / Orthography / Pronunciation
  - [IPA 23+3]

Thank you for your attention.

- Dankon por via atento
- Dziękuję za uwagę
- ☺

Refs.

- Brown, J.C. (1960). Loglan. *Scientific American*, 202(6).
- Eaton, H.S. (1940). An English French German Spanish Word Frequency Dictionary. New York: Dover [1961].
- Hogben, L. (1943). *Interglossa*. Harmondsworth: Penguin Books.
- Hogben, L. (1963). *Essential World English*. London: Michael Joseph Ltd.
- IALA (1951). *Interlingua-English Dictionary*. New York: Frederick Ungar Publishing Co.
- Longman (2003). *Dictionary of Contemporary English*. Harlow: Pearson Educational Ltd.
- Macmillan (2002). *MacMillan English Dictionary for Advanced Learners*. Oxford: MacMillan Education.
- Ogden, C.K. (1937). *The ABC of Basic English*. London: Kegan, Paul, Trench, Trubner & Co. Ltd.

Websites

- http://www.longmandictionariesusa.com/res/shared/vocab_definitions.pdf
- http://www.macmillandictionary.com/learn/clear-definitions.html
- http://ogden.basic-english.org/words.html
- http://www.oxfordlearnersdictionaries.com/wordlist/english/oxford3000/
- http://www.richardsandesforsyth.net
- http://www.slovio.com/
- http://www.typeit.org

26 Roman letters can be pronounced

Consonants:				
Roman Letter	**Anglo hint**	**IPAsign**	**features**	**Kyrillik**
b	b	b	bilabial voiced stop	б
c	sh	ʃ	unvoiced postalveolar fricative	ш
d	d	d	voiced alveolar stop	д
f	f	f	unvoiced nonsibilant fricative	ф
g	g	g	voiced velar stop	г
h	h	h	glottal approximant	[h]
k	k	k	unvoiced velar stop	к
l	l	l	coronal lateral approximant	л
m	m	m	bilabial nasal	м
n	n	n	bilabial alveolar	н
p	p	p	bilabial unvoiced stop	п
q	kh	x	velar fricative	х
r	rr	r	coronal trill	р
s	s	s	unvoiced sibilant fricative	с
t	t	t	unvoiced alveolar stop	т
v	v	v	voiced nonsibilant fricative	в
x	zh	ʒ	voiced velar nonsibilant fricative	ж
z	z	z	voiced sibilant fricative	з

26 Roman letters can be pronounced

Vowels & semivowels:

Roman Letter	**Anglo hint**	**IPAsign**	**features**	**Kyrillik**
a	ah	a	front open unrounded	а
e	e	e	front mid unrounded	э
i	ee	i	front close unrounded	и
j	y	j	palatal approximant	й
o	aw	o	back mid rounded	о
u	oo	u	back close rounded	у
w	w	w	voiced labio-velar approximant	[w]
y	ü	y	front close rounded	ю, ы

We can pick from multiple languages (eo, pt, es, it, la) [easy examples, levensim]

```
term : ('wall', 'muro') ==> ['muro', 'parede', 'pared', 'muro', 'murus']
*  0.5222 muro
   0.5222 muro
   0.4288 murus
   0.3884 pared
   0.3727 parede
['muro', ('parede', 0.2), ('pared', 0.2222222), ('muro', 1.0), ('murus', 0.6666667),
0.5222222]

term : ('war', 'milito') ==> ['milito', 'guerra', 'guerra', 'guerra', 'bellum']
*  0.5417 guerra
   0.5417 guerra
   0.5417 guerra
   0.1667 bellum
   0.0417 milito
['guerra', ('guerra', 1.0), ('guerra', 1.0), ('bellum', 0.1666667), ('milito', 0.0), 0.5416667]
```

Underlying Theme

- “central” / “common” / “typical” / “widespread” vocabulary => “acceptable” / “familiar” / “learnable” / “practical” vocabulary
- Not only Zamenhof’s idea, but also explicit or tacit in:
  - □ (May I mention the V-word?!)
  - □ Esperanto, Interlingua, Loglan / Lojban, Lingua Franca Nova, Slovio / Interslavic, et cetera

Magyar, Malay, Maori, Mandarin!

```
0.3210288 ('weather', 'vetero')      :: ahua o te
rangi
0.3126488 ('yes', 'jes')             :: ae
0.3039773 ('woman', 'virino')        :: no
0.3028291 ('will', 'volo')           :: pirangi
0.2995547 ('why', 'kial')            :: na te aha
0.2975385 ('wool', 'lano')           :: gyapju
0.2940146 ('wine', 'vino')           :: wain
0.2875000 ('way', 'vojo')            :: ut
0.2814453 ('wave', 'ondo')           :: bolang
0.2734666 ('waste', 'malŝparo')      :: langfei
0.2625319 ('unit', 'unuo')           :: tetahi
0.2612981 ('week', 'semajno')        :: xingqi
```