Covering diversity: some notes on sampling technique

Jan-Wouter Zwart
University of Groningen

TIN-dag, Utrecht, February 2 2008

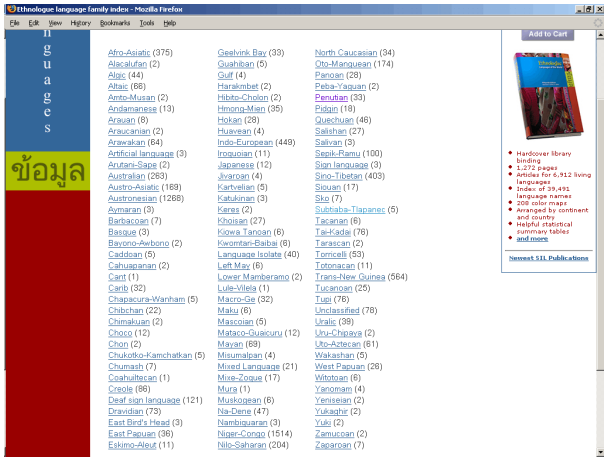
Sampling

- selection out of the world's languages (for survey/comparison)
- use some stratification (language families)
- avoid bias (genetic, geographic)
- **cover diversity** (leave nothing out)

Practical issues

- classification (splitting vs. lumping)
- existence of language descriptions
- availability of language descriptions
- ideal sample size (small for time, large for coverage)

Conservative (splitting) classification: Ethnologue



Example: Uralic

Ruhlen (1987) *A guide to the world's languages I: classification*. Stanford.

Ethnologue (splitting)
- Finnic (11)
- Finno-Ugric (1)
- Mari (2)
- Mordvin (2)
- Permian (3)
  - Komi
  - Udmurt
- Sami (11)
  - E
  - S
  - W
- Samoyed (7)
- Khanti
- Mansi

Ruhlen 1987 (lumping)
- Yukaghir (1)
- Uralic (23)
  - Samoyed (4)
    - N
    - S
  - Finno-Ugric (19)
    - Ugric
      - Hungarian
      - Ob-Ugric [Xanty, Mansi]
    - Finnic
      - Permic
      - Volgaic [Mari, Mordvin]
      - N Finnic
        - Saamic
        - Baltic Finnic

Previous work on diversity coverage: Rijkhoff et al 1993

- how many languages from each family should the sample contain?
- representative number (based on size) modulo **diversity value** (DV)
- DV calculated by inspecting the family tree
- classification: Ruhlen (1987)
- DV: average number of nodes per level in the family tree
- weighted for tree depth (higher levels count heavier)

Rijkhoff, Bakker, Hengeveld, Kahrel (1993) 'A method of language sampling.' *Studies in Language* 17, 169-203.

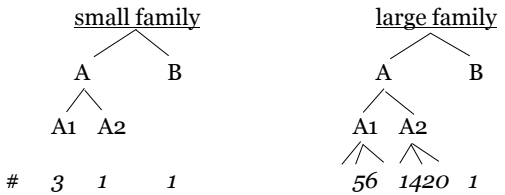
Questions left open

- what is the actual diversity coverage for a given sample ?
- how does addition/deletion of a language affect diversity coverage ?
- does size representativity adjusted for DV suffice for covering diversity ?

E.g. in a 250 language sample, Uralic-Yukaghir is represented by a single language (according to Rijkhoff et al.'s system).

Intuitively, we want small families to be overrepresented and large families to be underrepresented.

Size/representation



- Rijkhoff et al: 1 lg. from small family (regardless size of the sample)
- But diversity coverage requires that we include a language from A and B in both families, so at least 2 lgs. from the small family

Basic approach

- Rationale: every split (in the tree) represents an instance of variation
- splitting classification
- Rule 1: include a language from each family, including every isolated lg. (cf. Rijkhoff et al. 1993:179)
- Rule 2: within a family, include a language from each subfamily (recursive)

First pass

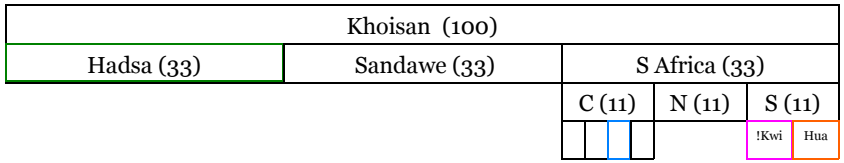
- Count the number of branches represented (again with weighting for depth)
- Problem: more deeply embedded languages yield more points, but not better diversity coverage

level 1	*level 2*	*level 3*	*level 4*	*level 5*
KHOISAN	Hadsa			
	Sandawe			
	S Africa	C	Hain//um	
			Kwadi	
			Nama	
			Tshu-Kwe	4 more
		N		
		S	!Kwi	
			Hua	

Adjustment: counting oppositions

- A branch is represented only if it represents an **instance of variation**
- In the Khoisan example, both Hadsa and Hua represent only one instance of variation: Khoisan vs. non-Khoisan (Level 1)
- If both Hadsa and Hua are present, there is an instance of variation at Level 2 (Hadsa vs. S Africa) as well as at Level 1 (Khoisan vs. non-Khoisan)
- If both Hua and !Kwi are present, there is no instance of variation at Level 2, but there is one at Level 1 and Level 4
- If both Hua and Nama are present, there is an instance of variation at Level 3 (CS Africa vs. SS Africa), but not at level 2 or 4

Scoring



- maximal score per level: 1. 100 2. 100 3. 33 4. 22
  divisor over 4 levels = 255 (*not* 400!)
- if the sample includes: the score is: and the diversity coverage:

Hadsa	100/0/0/0	100/255 = .39
Hua	100/0/0/0	100/255 = .39
Hadsa, Hua	100/66/0/0	166/255 = .65
Hua, !Kwi	100/0/0/11	111/255 = .44
Hua, Nama	100/0/22/0	122/255 = .48

Evaluating a sample

#	PHYLUM	LGS	%	SAMPLE 6			COVERAGE			
				lgs	/267	repr	opp	div	cov	
AFRICA										
1	Afro-Asiatic	375	5.43	12	4.49	.032	270	367	.71	
2	Khoisan	27	0.39	2	0.75	.074	122	255	.48	
3	Niger-Congo	1514	21.90	31	11.61	.020	270	384	.70	
4	Nilo-Saharan	204	2.95	7	2.62	.034	163	295	.55	
		2120	*30.67*	*52*	*19.48*	*.025*			*.61*	
TOTAL										
		6912	**100**	**267**	**100**	**.039**			**.64**	

- Khoisan is overrepresented, but has relatively poor coverage

Earlier sample

#	PHYLUM	LGS	%	SAMPLE 4			COVERAGE			
				lgs	/214	repr	opp	div	cov	
AFRICA										
1	Afro-Asiatic	375	5.43	8	3.74	.021	235	367	.64	
2	Khoisan	27	0.39	2	0.93	.074	122	255	.48	
3	Niger-Congo	1514	21.90	26	12.15	.017	254	384	.66	
4	Nilo-Saharan	204	2.95	5	2.50	.025	150	295	.51	
		2120	*30.67*	*41*	*19.16*	*.019*			*.57*	
TOTAL										
		6912	**100**	**214**	**100**	**.031**				

Comparison

	number of lgs		representation		coverage	
	S4	*S6*	*S4*	*S6*	*S4*	*S6*
afro-as.	8	12	.021	.032	.64	.71
khoisan	2	2	.074	.074	.48	.48
niger-cg	26	31	.017	.020	.66	.70
nilo-sah	5	7	.025	.034	.51	.55
total (sample)	*41 (214)*	*52 (267)*	*.019 (.031)*	*.025 (.039)*	*.57*	*.61 (.64)*

- sample growth: 11 lgs.
- effects on representation and coverage made visible

Conclusion

- diversity coverage may be calculated by scoring represented oppositions (sister pairs in a language family tree)
- the method
  - provides a useful tool for comparing (stages of) samples
  - makes it possible to evaluate the effects of adding/deleting languages
- view the sample used in the NWO-research program 'Dependency in Universal Grammar' at:

  www.let.rug.nl/zwart/diug/

Faculty of Arts, PO Box 716, 9700 AS Groningen
c.j.w.zwart@rug.nl ● www.let.rug.nl/zwart