On the typological distribution of rare characteristics

Michael Cysouw
Max Planck Institute for Evolutionary Anthropology, Leipzig
cysouw@eva.mpg.de

24 February 2005
27th annual meeting of the DGfS, Cologne
AG 4: Auf alles gefasst sein: Ausnahmen in der Grammatik

1. Using the *World Atlas of Language Structures*

The World Atlas of Language Structures (WALS) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of more than 40 authors (many of them the leading authorities on the subject). It will be published as a printed book in traditional atlas format, accompanied by a fully searchable electronic version that also allows various visualization effects.

The World Atlas of Language Structures consists of 142 maps with accompanying texts on diverse features (such as vowel inventory size, noun-genitive order, passive constructions, and 'hand'/'arm' polysemy), each of which is the responsibility of a single author (or team of authors). Each maps shows between 120 (35) and 1110 languages, each language being represented by a dot, and different dot colors showing different values of the features. Altogether more than 2,600 languages are shown on the maps, and more than 55,000 dots give information on features in particular languages.

The idea for the present investigation is to use the WALS data for 'holistic' typology. There are features coded from all areas of linguistic structure, so it is possible to look for combinations of different aspects of linguistic structure. For the analysis in this talk, I will not look at the content of the features, but at their relative ubiquity. I will ask question like: are there languages, families or areas that have more unusual characteristics than other? And which characteristics are they?

Haspelmath, Martin & Dryer, Matthew & Gil, David & Comrie, Bernard (eds.) 2005. *The World Atlas of Language Structures*. (Book with interactive CD-ROM) Oxford: Oxford University Press. (see http://www.eva.mpg.de/lingua/files/wals.html)

2. Computing the Rarity Index

Basic idea to compute a rarity index: take the chance of occurrence of a particular characteristic ('value') in the whole database, but normalise this by the number of characteristics distinguished in a particular parameter ('feature')

(1) $$R_{f_i} = n \cdot \frac{f_i}{f_{tot}}$$

n = number of values of a particular feature
f_i = frequency of value i
f_{tot} = total number of languages coded for this feature

I used the inverse of this index:

(2) $$R_i = \frac{f_{tot}}{n \cdot f_i}$$

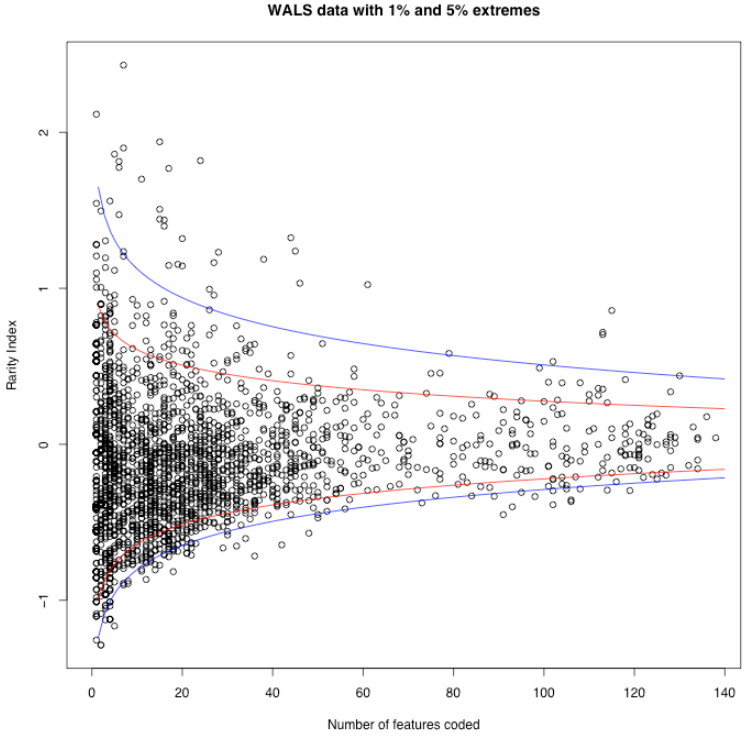
This reverse index is useful because the mean of all these indices over the whole database is 1:

(3) $$\frac{\sum_{i=1}^{n}(R_i \cdot f_i)}{f_{tot}} = 1$$

Simply taking the highest mean rarity over all features is not a good measure to evaluate which language has the most unusual characteristics. If only few features are coded in the WALS, there will be strong random effects. Languages with less code-points in the WALS will have more extreme values of the rarity index.

To normalize these effects, I computed the random distribution of rarity depending on the number of features coded (by a randomization technique) and ordered languages according to the chance level of occurrence of the rarity index. In the tables below, this rarity index-level is indicated by a percentage. For example, 99.9% means that this particular value is within the top 0.1% of all languages in the database. Note that this value is not a real significance value as given by many statistical analyses (although it is alike to it). This value indicates the relative unusualness of a particular data point within the present dataset only.

(4) Plot of rarity indices with some index-levels included



As a value for the rarity of a group of languages (genealogical groups or areal groups) I have used a weighted mean of the rarity index-levels of the individual languages. Basically, I take the mean of all rarity index-levels, but weight the languages according the logarithm of the number of features coded. In this way, the languages with more features coded have a stronger influence on the resulting value. The Group Rarity index GR_i is defined as follows:

(5) $$GR_i = \frac{\sum_{i=1}^{n}\log(L_i)\cdot(\%R)_i}{\sum_{i=1}^{n}\log(L_i)}$$

n = number of languages in a group
L_i = number of features coded for language i
$\%R_i$ = rarity index-level for language i

3. Some Surveys of Rarity

(6) Top 1% of languages according to Mean Rarity Index

Language	Family	Genus	Features Coded	Mean Rarity	%
Wari'	Chapacura-Wanhan	Chapacura-Wanhan	115	2.36	99.9
Dinka	Nilo-Saharan	Nilotic	45	3.45	99.9
Tiipay (Jamul)	Hokan	Yuman	44	3.76	99.9
Nuer	Nilo-Saharan	Nilotic	28	3.42	99.9
Karó (Arára)	Tupian	Tupi-Guarani	24	6.16	99.9
Winnebago	Siouan	Siouan	7	11.37	99.9
Mixtec (Chalcatongo)	Oto-Manguean	Mixtecan	113	2.05	99.8
Kutenai	Kutenai	Kutenai	113	2.02	99.8
Kombai	Trans-New Guinea	Awju-Dumut	38	3.27	99.8
Dahalo	Afro-Asiatic	Southern Cushitic	17	5.86	99.8
Maxakali	Macro-Ge	Maxakali	15	6.95	99.8
Warrwa	Australian	Nyulnyulan	20	3.74	99.7
Bunuba	Australian	Bunuban	16	4.21	99.7
Eyak	Na-Dene	Eyak	16	4.05	99.7
Yawuru	Australian	Nyulnyulan	15	4.51	99.7
Diegueño	Hokan	Yuman	61	2.78	99.6
Dagaare	Niger-Congo	Gur	27	3.20	99.6
Anindilyakwa	Australian	Anindilyakwa	15	4.23	99.6
Wik Ngathana	Australian	Pama-Nyungan	7	6.68	99.6
Puquina	Puquina	Puquina	5	6.42	99.6
Hakka	Sino-Tibetan	Chinese	26	2.70	99.5
Cree (Eastern)	Algic	Algonquian	20	3.14	99.5
Tobati	Austronesian	Oceanic	17	3.15	99.5
Kashaya	Hokan	Pomoan	6	6.13	99.5
Amis	Austronesian	Paiwanic	6	5.90	99.5
Gworok	Niger-Congo	Platoid	11	5.47	99.4
Kabardian	NW Caucasian	NW Caucasian	46	2.81	99.3
Nyulnyul	Australian	Nyulnyulan	27	2.60	99.3
Ngemba	Niger-Congo	Bantoid	1	8.30	99.3
Ngiti	Nilo-Saharan	Lendu	79	1.79	99.2
Nadeb	Vaupes-Japura	Vaupes-Japura	19	3.17	99.2
Asurini	Tupian	Tupi-GuaranÌ	6	4.36	99.1
Podopa	Teberan-Pawaian	Teberan	4	4.76	99.1
Mocovi	Guaicuruan	Guaicuruan	26	2.37	99.0

(7) Languages with more than 100 features coded in top 10% of Mean Rarity Index

Language	Family	Genus	Features Coded	Mean Rarity	%
Wari'	Chapacura-Wanhan	Chapacura-Wanhan	115	2.36	99.9
Mixtec (Chalcatongo)	Oto-Manguean	Mixtecan	113	2.05	99.8
Kutenai	Kutenai	Kutenai	113	2.02	99.8
Wichita	Caddoan	Caddoan	102	1.70	98.9
Mangarrayi	Australian	Mangarrayi	118	1.51	98.7
Mandarin	Sino-Tibetan	Chinese	130	1.55	98.2
German	Indo-European	Germanic	128	1.40	97.9
Iraqw	Afro-Asiatic	Southern Cushitic	104	1.48	97.9
Oneida	Iroquoian	Northern Iroquoian	101	1.49	97.5
Maricopa	Hokan	Yuman	112	1.43	97.3
Yagua	Peba-Yaguan	Peba-Yaguan	108	1.48	97.1
Warao	Warao	Warao	110	1.38	96.9
Gooniyandi	Australian	Bunuban	113	1.42	96.7
Latvian	Indo-European	Baltic	112	1.37	96.2
Apurina	Arawakan	Arawakan	110	1.32	95.5
Nunggubuyu	Australian	Nunggubuyu	103	1.33	95.5
Piraha	Mura	Mura	114	1.31	95.1
Kiowa	Kiowa-Tanoan	Kiowa-Tanoan	102	1.26	92.9
Ket	Yeniseian	Yeniseian	104	1.22	92.4
Malagasy	Austronesian	Borneo	124	1.20	91.0
Supyire	Niger-Congo	Gur	125	1.21	90.4
Ju|'hoan	Khoisan	Northern Khoisan	103	1.21	90.2
French	Indo-European	Romance	136	1.19	89.4

(8) Top 40% of weigthed rarity for Families (only Families with more than 3 languages)

Family	Languages	%
NorthwestCaucasian	7	87.8
Kartvelian	4	83.7
Caddoan	5	82.2
Wakashan	7	80.2
Iroquoian	8	76.3
Khoisan	11	74.5
Arauan	6	71.8
Salishan	24	71.2
NaDene	23	70.2
Algic	31	69.9
Hokan	21	65.4
Guaicuruan	5	65.1
EskimoAleut	19	64.4
KiowaTanoan	7	64.0
Penutian	26	63.0
Basque	12	60.3
Tucanoan	18	60.0

(9) Top 50% of weigthed rarity for Genera (only Genera with more than 3 languages)

Genus	Languages	%	Genus	Languages	%
NorthwestCaucasian	7	87.7	SouthernAtlantic	5	61.9
CentralSalish	13	85.8	Aslian	5	61.5
Yuman	10	85.5	Samoyedic	6	60.7
SouthernWakashan	4	85.2	Iranian	26	60.5
NorthernIroquoian	7	84.7	Basque	12	60.2
Kartvelian	4	83.6	Tucanoan	18	59.9
Caddoan	5	82.1	CrossRiver	8	59.6
Chinese	10	81.5	Otomian	6	59.5
Nyulnyulan	5	81.3	Numic	10	58.6
SouthernCushitic	6	80.2	Popolocan	7	56.8
Germanic	39	79.0	Mayan	34	56.3
MoruMadi	5	77.4	Cariban	19	55.8
CentralKhoisan	6	77.3	Tepiman	6	55.4
Arauan	6	71.7	PamaNyungan	105	55.2
Nupoid	4	71.6	Semitic	42	54.6
InteriorSalish	8	71.6	Kordofanian	8	54.2
Algonquian	29	69.6	Gur	32	54.0
Chinantecan	7	67.6	TimorAlorPantar	4	53.7
MesoPhilippine	11	67.3	LakesPlain	5	53.4
Kadai	4	66.4	AwjuDumut	5	53.1
Mek	4	65.8	SouthernPhilippines	5	52.9
NorthernPhilippines	16	65.6	Quechuan	11	52.7
Athapaskan	21	65.3	MixeZoque	11	52.6
Guaicuruan	5	65.0	Lezgic	10	52.0
EskimoAleut	19	64.3	EasternCushitic	13	51.4
KiowaTanoan	7	63.9	Paiwanic	4	51.0
Miwok	6	62.9	WesternMande	20	50.8
Nilotic	19	62.6	MarindProper	4	50.1
AvarAndicTsezic	13	62.0	Sundic	37	50.0
Kru	9	62.0			

4. Areal distribution of rarity

To evaluate whether there are geographical areas with a high preponderance of rare features, I investigated groups of languages that are geographically contiguous. For each language in the database, I took the 30 nearest languages (using a simple Euclidean distance, not taking account of natural barriers) and computed the rarity for all areal groups. The rarity index for each group is plotted on a map on the location of the centre of the group. Such an approach necessary will show some areal consistency. However, it is interesting to see what the centres of areally consistent groups are.

The seven areas with the most extreme areal level of 'rarity' are summarised here. These are the areas cantered around:

- Frisian (North-western Europe)
- Adyghe (Caucasus)
- Bikol (Philippines)
- Walmatjarri (Northern Australia)
- Lummi (Northwest Coast of North America)
- West Greenlandic (Northeast Coast of North America)
- Pirahã (Amazonia)

For each area, a table is given with the languages that are included. These are basically the 30 nearest languages to the centre-language.

- The languages included in the area were extracted from the WALS-database. The languages with no rare characteristics were removed
- The remaining languages are ordered to genealogical relationship.
- Within each genus, they are ordered to the number of code points these languages have in the WALS (there are maximally 140 code points per language; the numbers from the database are shown in the last column). This is important, because in the computation of the rarity for each area, the languages with many code points were valued more than the others.

Then follows a large table with various characteristics from the *World Atlas of Language Structures* (WALS). These characteristics were selected and ordered as follows:

- For each group of languages, the mean rarity index was computed for each feature. This mean is given in the first column. The higher this number, the more unusual this feature is in this area compared to the rest of the world (the whole-world mean is 1.0 for each feature).
- I have arbitrarily only included here the features with a mean rarity index higher than 1.5.
- For each feature, I have only collected those characteristics in the area with a high rarity index. I most cases, there are also some languages in the area that have 'normal' characteristics. Sometimes the high mean rarity is the result of only one language in the area having an extremely unusual characteristic, while all other languages in the area do not. **The lists of unusual characteristics should thus not be interpreted as claiming that all languages in the area have these characteristics, but only that the unusual characteristic in question occurs in this area; sometimes only in one language!**
- The unusual characteristics are grouped impressionistically into group of features that seem to have some coherence. The groups are ordered to the highest mean rarity (this ordering is used both within groups and between groups).
- A summary of the most outstanding unusual characteristics is extracted for each area and presented at the top of each list.