Language sample
for the typological study of caritives

Sofia Oskolskaya, Maksim Fedotov, Natalia Zaika
(Institute for Linguistic Studies, RAS, St. Petersburg)

Language Sampling: Issues and Challenges
Institute for Linguistic Studies, RAS, Saint Petersburg, 20 December 2019

Outline

- Introduction: the typological study of caritives
- A representative sample (main factors and biases)
- Size of the sample
- The sampling method
- Language selection
- Two samples instead of one

Introduction: The typological study of caritives[1]

CARITIVE as a comparative concept

CARITIVE describes non-involvement (including, but not limited to absence) of a participant (absentee) in a situation, with the non-involvement predication semantically modifying the situation or a participant of a different situation.

EXAMPLES:

1. *John came **without** his children*.
2. *John travelled **without** money*.
3. *A beard**less** man was sitting in the corner of the room.*

[1] The research is supported by the Russian Science Foundation, grant 18-78-10058.

Introduction: The typological study of caritives

What means are used to express caritive semantics?

Are there any correlations between linguistic parameters of these means?

- morphological parameters
- syntactic parameters
- semantic parameters

Introduction: The typological study of caritives

Morphological parameters

- affix vs. clitic vs. clause...
- derivational / inflectional markers on the caritive

Introduction: The typological study of caritives

Syntactic parameters

- Absentee: word classes (nouns, pronouns, adjectives, adverbs, (non-)finite verbs...)
- Syntactic function (attribute, predicate, depictive...)

Introduction: The typological study of caritives

Semantic parameters

- Animacy: human/animate/inanimate
- Referentiality
- Definiteness
- Meaning: companion, instrument, possessee, transport, circumstance...

A representative sample (main factors and biases)

<Quite traditional>

- A representative sample of world languages

  (If 90% of languages of the world lack a grammatical caritive marker, 90% of languages in the sample are supposed to lack it.)

- Genetically and areally balanced
- Bibliographically biased
- Typologically not balanced

Size and structure of the sample

- As big as possible
- Less than 200 due to natural limitations (very few good sources, see below)

- Proportion of languages from different families/genera and macroareas remains the same

The sampling method

Genus-Macroarea (GM) method described in [Miestamo et al. 2016]

Genus is a group of languages that have a common ancestor with a time-depth about 3500-4000 years before present, see (Dryer 1989).

Indo-European genera: Albanian, Armenian, Baltic, Slavic, Celtic etc.

Nivkh genus: Nivkh

521 genera in the World (from WALS?)

GM method in (Miestamo et al. 2016)

Table 2: Genera and languages by macroarea with different sample sizes.

	Genera	%	50	100	150	200	300	400	500	600
Africa	74	14.2	7	14	21	28	43	57	71	85
Eurasia	43	8.3	4	8	12	17	25	33	42	50
Southeast Asia & Oceania	66	12.7	6	13	19	25	38	51	64	76
Australia & New Guinea	140	26.9	13	27	40	54	81	108	135	161
North America	92	17.7	9	18	27	35	53	71	89	106
South America	106	20.3	10	20	30	41	61	81	102	122
Total	521	100.0	49	100	149	200	301	401	503	600

GM method in (Miestamo et al. 2016)

1. Select a proportional number of languages per Macroarea
2. 1 genus = 1 language
3. 1st round: 1 family = 1 language.
4. When all families are covered => 2nd round: +1 language from a different genus of families

Our sampling method

Close to the GM method in (Miestamo et al. 2016)

Some changes:

I. Number of genera

II. Language selection

I. Number of genera

- 521 genera in (Miestamo et al. 2016), based on WALS
- Problem: WALS includes only 2679 languages. Could it be that some genera were missed?
- It quite seems to be the case.
- Mostly because of isolates (any isolate comprises a genus by itself) or very small genera.
- We compared classifications from WALS and from Ethnologue and, as a result, added 100+ new (putative) genera.

I. Number of genera

- ➢ Classifications in WALS were based on Ethnologue => we compared them with Ethnologue rather than Glottolog.
- ➢ General idea: to add missing language groups = genera in terms of (Dryer 1989)
- ➢ Difficulty: Ethnologue does not use genera and does not indicate which taxons in its classification can count as genera.
- ➢ We followed a single algorithm while comparing WALS and Ethnologue and adding genera from the latter (we will share it, too).

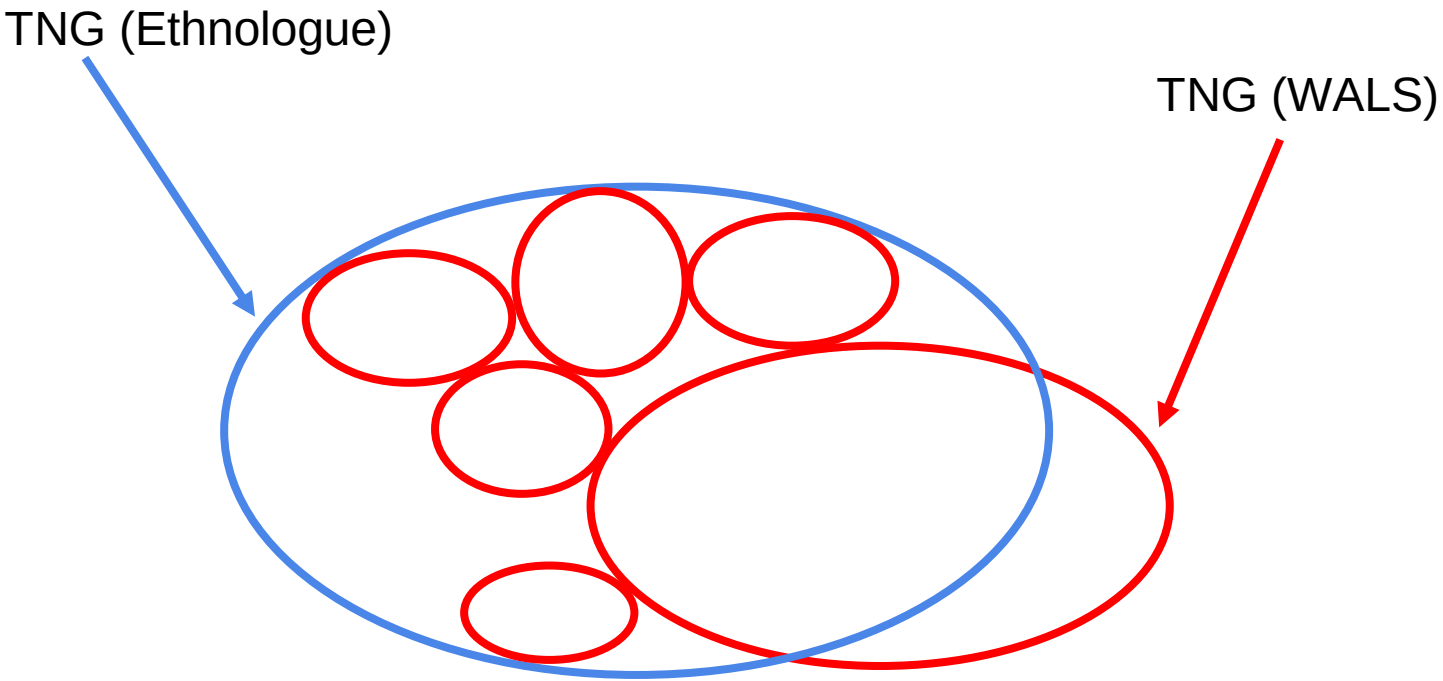
Simple case: Torricelli family (Papua New Guinea)

WALS	Ethnologue
Kombio-Arapesh	Kombio-Arapesh
Marienberg	Marienberg
Urim	Urim
Wapei-Palei	Wapei-Palei
West Wapei	West Wapei
	Maimai

More complicated case: Oto-Manguean languages

Ethnologue					WALS	
Otomangu	Eastern Ot	Amuzgo-N	Amuzgo		**Oto-Manguean**	Amuzgoan
			Mixtecan		Oto-Manguean	Mixtecan
		Popolocar	Popolocan		Oto-Manguean	Popolocan
			Zapotecan		Oto-Manguean	Zapotecan
	Western C	Oto-Pame	Oto-Pame	Chichimeco-Jonaz	Oto-Manguean	Chichimec
				Matlatzinca-Ocuilteco	Oto-Manguean	Matlatzincan
				Otomi, Mazahua	Oto-Manguean	Otomian
				Pame	Oto-Manguean	Pamean
			Chinantecan		Oto-Manguean	Chinantecan
		Tlapanec-	Subtiaba, Tlapanec		Oto-Manguean	Subtiaba-Tlapanec

Very complicated case: Trans-New Guinea languages



Total number of genera

➔ **660** genera (WALS + enrichments from Ethnologue)

Cf. 542 genera now in WALS

Cf. 521 genera in (Miestamo et al. 2016)

* Creoles, pidgins, sign languages were excluded.

* (Most?) extinct languages and some cases with doubtful status were excluded.

➔ We got (slightly) different proportions of genera in macroareas:

Macroarea	**Genera (660)**	**%**	Genera (521)	%
Africa	**93**	**14.09**	74	14.2
Eurasia	**57**	**8.64**	43	8.3
Southeast Asia & Oceania	**98**	**14.85**	66	12.7
Australia & New Guinea	**195**	**29.55**	140	26.9
North America	**94**	**14.24**	92	17.7
South America	**123**	**18.64**	106	20.3

Size and structure of the sample

- ➢ We (hope we) made a more complete list of genera of the world.
- ➢ We will be happy to share it, as well as the sample itself, so that it may be used in other typological projects.

- ➢ Why has nobody made such a useful list before?
  (based on WALS + Ethnologue / Glottolog)

- 1. We did something wrong. (What exactly?)
- 2. Somebody has done it, but we do not know. (Tell us!)
- 3. Indeed nobody has made such a list. (You are welcome to use ours, then).

II. Language selection

- GM method in (Miestamo et al. 2016): 1 family = 1 language
- Problem: big families are underrepresented (see also (Rijkhoff, Bakker 1993, 1998))

  Eurasia: 57 genera, including Indo-European (11 genera), Uralic (7 genera), Dravidian (10 genera), Nakh-Daghestanian (4 genera), 24 families with 1 or 2 genera (e.g. Japanese, Yukaghir etc.)

  8.6% in a sample => 2nd Indo-European language can appear only in a 400-languages sample

II. Language selection

- ➢ Aim of our project: to observe a general situation with caritives in the world
- ➢ E.g. if 90% of languages of the world lack a grammatical caritive marker, 90% of languages in the sample are supposed to lack it.

- ➢ We take into account the size of families:
- Eurasia 57 genera, 8.6% (~9%)
- Indo-European 11 genera => 1.7% (~2%)

=> We take 2 Indo-European languages from different genera for the 100-languages sample.

An example:

Family	Genus	Number (%)	Rounded (%)
Indo-European	Albanian	1.7	2
	Armenian		
	...		
Uralic	Finnic	1.1	1
	Mari		
	...		
Nakh-Daghestanian	...	0.63	5
Ainu	Ainu	0.16	
...			

II. Language selection

- A list of 660 genera.
- A proportion of genera/languages per each macroarea.
- A proportion of genera/languages per each family.
- 1 genus = 1 language
- 2nd round for big samples (more than 660 languages?)

Language selection: Technical factors

1. existence of collections of glossed texts
2. high-quality modern grammar descriptions (searchable grammars)
3. comprehensive dictionaries with sentential examples
4. existence of Bible translations
5. the areal (geographical) balance
6. WALS 200-languages sample

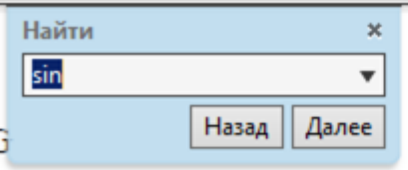
Language selection: Searchable grammars

(in the contents: “NP”, “negation”)
“cariti(ve)”
“abessi(ve)”
“privati(ve)”
“absen-”
“lack”
“without” / “без” / “sans” / “sin” / ...
“with no”
“negat-”
“-less”
“but for”(?) / “if not for” / “if it weren’t (for)”...

“except” / “кроме” / “sauf”
“comitative”
“(as)sociative”
“unitive”
“together”
“with” *(?)*
“accompany”
“cooperative”

“bald” (“hair + -less”?)
“blind” (“eye + -less”?)

Language selection: Searchable grammars



(551) *bɤ́-dǒu-sãn-dít*
otro-fecha-NEGV-haber.NEG
nunca

Morfema de negación adverbial *'ɯ̃̂* (NEG)

Con el morfema *'ɯ̃̂* se responde negativamente a una pregunta cerrada, se niegan constituyentes nominales y cláusulas y se construye una expresión de privativo adicionando el morfema *-jet* (COM) (552)b. Consideramos esta marca como una negación adverbial. Esta marca tiene tono propio, el cual conserva en todos los contextos, y puede estar como palabra independiente. La forma es homófona con el lexema *'ɯ̃́* 'morir'.

(552) a.	*ja-yúk-ni*	*kə̃̂n*	*'ɯ̃̂-*
	3SG-venir-PSVG	hamaca	NEG
	él vino sin su hamaca		

Language selection: Existence of Bible translations

“The most caritive” contexts:

- Mt 10:29; Mt 12:5; Mt 13:34; Mt 13:57; Mt 15:20; Mt 20:6; Mt 22:12; Mt 25:38; Mt 26:17; Mt 26:42
- Mk 4:34; Mk 6:4; Mk 7:2; Mk 7:5; Mk 7:18; Mk 12:20; Mk 12:21; Mk 14:1; Mk 14:12
- Lk 1:74; Lk 1:6; Lk 6:49; Lk 9:41; Lk 11:36; Lk 11:42; Lk 11:44; Lk 15:13; Lk 20:28; Lk 20:29; Lk 20:30; Lk 20:31; Lk 22:1; Lk 22:6; Lk 22:7; Lk 22:35
- Jn 1:3; Jn 7:15; Jn 8:7; Jn 13:22; Jn 15:5; Jn 15:25; 19:23

Language selection: Existence of Bible translations

- Mt 10:29
  Are not two sparrows sold for a farthing? and one of them shall not fall on the ground **without your Father**.
- Mt 13:34
  All these things spake Jesus unto the multitude in parables; and **without a parable** spake he not unto them.
- Lk 22:35
  And he said unto them, When I sent you **without purse, and scrip, and shoes**, lacked ye any thing? And they said, Nothing.

Language selection: The map

Language selection: Eurasia (8)

- Lithuanian < Indo-European
- Hindi < Indo-European
- Hill Mari < Uralic
- Telugu < Dravidian
- Khalkha < Mongolic
- Nanai < Tungusic
- Basque [isolate]
- Adyghe < Northwest Caucasian

Language selection: Africa (14)

- Hausa < Afro-Asiatic
- Oromo < Afro-Asiatic
- Modern Hebrew < Afro-Asiatic
- Furu = Bagiro < Central Sudanic
- Ik = Icé-tód = Ngulak < Kuliak
- Tirmaga Suri < Eastern Sudanic

- Lumun < Niger-Congo
- Zulu < Niger-Congo
- Yoruba < Niger-Congo
- Ƙoromfe < Niger-Congo
- Ewe < Niger-Congo
- Bambara < Mande
- Ju|’hoan < Kxa
- Sandawe [isolate]

Language selection: Southeast Asia & Oceania (15)

- Khmer < Austro-Asiatic
- Vietnamese < Austro-Asiatic
- Woi < Austronesian
- Malagasy < Austronesian
- Tagalog < Austronesian
- Malay < Austronesian
- Paiwan < Austronesian
- Chamorro < Austronesian
- Ladakhi < Sino-Tibetan
- Burmese < Sino-Tibetan
- Lepcha < Sino-Tibetan
- Mandarin Chinese < Sino-Tibetan
- Thai < Tai-Kadai
- ***(+ 2 more languages)***

Language selection: Australia & New Guinea (29)...

- Asmat < Trans-New Guinea
- Lower Grand Valley Dani < Trans-New Guinea
- Amele < Trans-New Guinea
- Una < Trans-New Guinea
- Maybrat < West Papuan
- Nunggubuyu < Gunwinyguan
- Alamblak < Sepik
- Yimas < Sepik-Ramu
- Urim < Torricelli
- Mangarrayi < Mangarrayi-Maran
- Martuthunira < Pama-Nyungan
- Guragone < Mangrida
- Klon < Timor-Alor-Pantar
- Taulil = Tulil < Baining-Taulil
- Imonda < Border
- Aimele < Bosavi
- Gooniyandi < Bunuban
- Wulna < Darwin Region

...

Language selection: Australia & New Guinea (29)

...

- Kaki Ae < Eleman
- Gija < Jarrakan
- Fas = Momu < Kwomtari-Baibai
- Marind < Marind
- Arammba < Morehead and Upper Maro Rivers
- Mullukmulluk = Malak-Malak < Northern Daly
- Bardi < Nyulnyulan
- Kayardild < Tangkic
- Ngarinyin = Ungarinjin < Worrorran
- ***(+ 2 more languages)***

Language selection: North America (14)

- Mazatec < Oto-Manguean
- Otomi < Oto-Manguean
- Southern Sierra Miwok < Penutian
- Coast Tsimshian < Penutian
- Nahuatl < Uto-Aztecan
- Seri < Hokan
- Arapaho < Algic
- Navajo < Na-Dene
- Aleut < Eskimo-Aleut
- Mohawk < Iroquoian
- Tzeltal < Mayan
- Miskito < Misumalpan
- Zoque < Mixe-Zoque
- Sioux < Siouan

Language selection: South America (19)

- Machiguenga < Arawakan
- Apurinã < Arawakan
- Kuna < Chibchan
- Pech < Chibchan
- Apinayé < Vacro-Ge
- Guarani < Tupian
- Qawasqar < Alacalufan
- Mapuche < Araucanian
- Aymara < Aymaran
- Ye’kuana = Maquiritari < Cariban
- Huambisa = Wampis < Jivaroan
- Wichí = Mataco < Matacoan
- Dâw < Nadahup
- Matsés < Panoan
- Puinave < Puinave
- Quechua < Quechuan
- Ese Ejja < Tacanan
- Wanano < Tucanoan
- Urarina [isolate]

Two samples instead of one

- While selecting languages for the sample, we understood that to some extent we are bringing in an additional factor for selection:
- If in the available materials (grammar, dictionary, texts) we find absolutely nothing about caritive contexts, we would replace this language in the sample by a closely-related one, for which more information on caritive contexts could be found.
- => An additional bias, since the described situation would occur more frequently in the languages without a specialized caritive marker than in languages with one.

Two samples instead of one

- At the same time, it obviously makes our study better, since this way we simply get more data on ways of expressing caritive semantics.
- Is there a solution to this dilemma?

- Two samples.
- We retain both versions of the sample (which largely, but not fully, intersect):

Two samples instead of one

1. The original — unbiased — one (without the replacements made because of lack of data on caritive contexts)

2. The modified one (with these replacements).

- We would mostly use the second one: for example when making claims about different types of specialized (or non-specialized) means of expressing caritive semantics.
- But we will use the first — unbiased — one when making claims about proportions and tendencies for which the bias in question is undesirable, e.g. about the general frequency of specialized caritive markers among all the languages.

References

- Dryer, Matthew S. 1989. Large linguistic areas and language sampling. Studies in Language 13: 257–292.
- Miestamo, Matti & Bakker, Dik & Arppe, Antti. 2016. Sampling for variety. Linguistic Typology 20(2): 233–296.
- Rijkhoff, Jan & Bakker, Dik & Hengeveld, Kees & Kahrel, Peter. 1993. A method of language sampling. Studies in Language 17.1: 169–203.
- Rijkhoff, Jan & Bakker, Dik. 1998. Language sampling. Linguistic Typology 2: 263–314.