LENGTH OF EAST CAUCASIAN SUBJECT INDEXES: A QUANTATIVE RESEARCH[1]

G. A. Moroz (HSE University, Russian Federation)

1 Introduction

Since Zipf's claim that word length is determined by frequency (Zipf 1936; Bentz, Ferrer-i-Cancho 2016; Ferrer-i-Cancho, Bentz, Seguin 2020), a lot of works appeared within the field of computational linguistics seeking to confirm or disconfirm this claim. It has been shown that frequency of linguistic objects (e.g. words, morphemes, syllables, phonemes) is connected to other cognitive processes like parsing and production. As a result, frequency of linguistic objects can correlate with other features, such as the information content of words (Piantadosi, Tily, Gibson 2011), affix ordering (Hay 2002), probability of phonological contrast loss (Wedel, Jackson, Kaplan 2013), phonological structure of words (Landauer, Streeter 1973), and many others. Language unit frequency also plays a huge role in usage-based linguistic approaches (see (Tummers, Heylen, Geeraerts 2005) and the references therein), since frequency is treated as an indication of prototypicality and usage tendencies.

Frequency itself is an ambiguous notion, since there are **type frequencies** and **token frequencies** (Berg 2014). Type frequency refers to frequency across e.g. a dictionary or any other collection of language units. Token frequency refers to frequency across some corpus of speech events. Not all linguistic problems are equally approachable using both types of frequency: some problems can be approached with both types of frequencies (e.g. obstruents vs. sonorants, transitive vs. intransitive verbs etc), some problems can be approached only with token frequency (e.g. past vs. present tense, singular vs. plural forms), and some problems can be approached only with type frequency (e.g. sex-based vs. non-sex-based gender systems).

In order to show how frequency can determine the length of grammatical morphemes, I decided to select a small subdomain and explore subject, person and number indexing on the verb. Although only a few East Caucasian languages show agreement with person, it has been found in all branches of East Caucasian languages except Khinalug, and it appears to be an independent development in all of them (Troubetzkoy 1929; Гасанова 1962; Helmbrecht 1996; Berg 1999; Муталов 2002; Schulze 2007; Сумбатова 2011; Foley 2020).

In this article I present a connection between frequency and length of person-number indexes via two independent researches: token frequency obtained from the Universal Dependencies' treebanks (section 2) and type frequency gathered within a typological study. After introducing the results of those two studies, I will present East Caucasian data (section 3). I show that the

[1] I would like to express my gratitude for Rasul's hospitality and willingness to help with research. Johannes Helmbrecht also thanked the honoree in his article, so I hope this topic will lure more researchers, who will be able to meet Rasul and learn to appreciate his passion and expertise in the future.

The article was prepared within the framework of the HSE University Basic Research Program.

unusual history of person-number indexes in these languages leads to violations of the tendencies that were shown in sections 2 and 3. The last section concludes the paper.

2 Token frequency of person-number indexes: data from Universal Dependencies' corpora

2.1 Data

In order to prove the claim that the relative frequency of person-number indexes is connected with marker length, one needs to obtain those frequencies. I used the Universal Dependencies' corpora (De Marneffe et al. 2014; Zeman et al. 2020) for this purpose. The Universal Dependencies project[2] provides a collection of treebanks (in version 2.7, which was used for this paper, there were 183 treebanks of 104 languages) that use the same strategy for marking syntactic dependency and morphological annotation. In order to prevent biases towards a particular language family I sampled languages with a big number of tokens (greater than 50,000), and only one language per branch. After deciding on the language sample, I automatically extracted all non-clitical verb forms that contain `Person` and `Number` tags. As a result, I obtained a list of language corpora that could be used in a frequency estimation: Arabic, Armenian, Bulgarian, Greek, Hindi, Persian, Romanian, Turkish, Wolof. This list is small, and 2/3 of the languages are Indo-Eropean (though from different branches). However, it can still be used to get a preliminary estimation. For the sake of comparability I decided to remove dual number forms from the Arabic corpus.

2.2 Results

The results of the analysis are presented in Figure 1. On the x-axis is the ratio of different forms colored by number and split by person on the y-axis. Results are presented not as separate values, but as a kernel density estimation curve, which is a smoothed version of the histogram. For example, the density peak in the third singular form around 0.5 signals that the majority of languages have a value around 0.5 for third person forms in the corpus. Overall grouping of this curve after the value 0.5 denotes that all languages have values around 0.5 or greater. This means that in all corpora third singular forms constitute half of all person forms. From Figure 1 one can see that third singular tokens form the majority of all forms. Peaks of the second and first singular forms are around 0 and do not differ much, except for a small subcluster for the second singular forms. In plural forms the difference between third and other persons is not so drastic, although it still shows a significant predominance of third person forms. It appears that in the plural forms there is no visible distinction between second and first persons that could be found in the singular forms. The overall effect of number is also visible on the graph: plural forms in first and second persons are a little bit less frequent, and in the third person it is even significantly rarer.

[2] All corpora can be accessed here: http://universaldependencies.org.

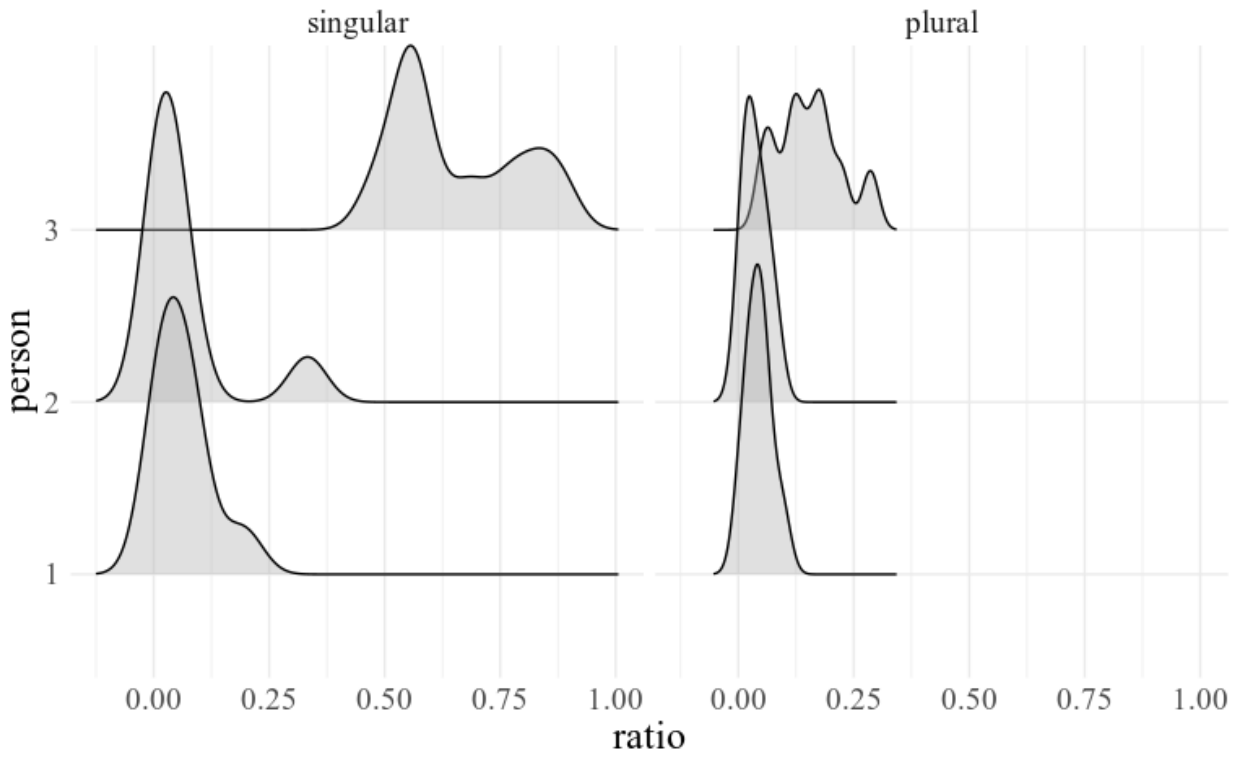

Figure 1. Kernel density function of the ratio of subject indexes in Universal Dependencies corpora (Arabic, Armenian, Bulgarian, Greek, Hindi, Persian, Romanian, Turkish, Wolof)[3]

To conclude this section I can formulate my expectations after collecting token frequencies of personal indexes. If the length of a personal index is determined by its frequency, then I expect that:

(1) third person forms are shorter than first and second person forms;
(2) first and second forms are more or less of the same length;
(3) singular forms within each person are shorter than plural forms.

For now we confirmed this with the token frequency analysis presented in this section. In order to strengthen those claims in the following section I will discuss a typological survey of personal indexes that analyzes type frequency (Seržant, Moroz in preparation). In this survey we also came to the same conclusions.

[3] All graphs in this paper were created with the R programming language (R Core Team 2020) using the following packages: `ggplot2` (Wickham 2016), `ggeffects` (Lüdecke 2018), and `lingtypology` (Moroz 2017).

3 Type frequency of person-number indexes: data from a typological survey

3.1 Data

This section is based on the results of a typological survey of personal indexes (Seržant, Moroz in preparation). One of the main ideas of this article is that the length of personal indexes is not random, but has some values that are more probable then others. Those preferences are described as attractors (Cooper 1999: 28). Attractors is a notion borrowed from the mathematical field of dynamic systems, in which they are defined as a subspace of possible values where the behavior of the system tends to be a fixed value. In order to discover those attractors, 383 languages from 53 families were analyzed (East Caucasian languages were not in the sample), covering all six macro-areas of the world: Eurasia, both Americas, Australia, Africa and Oceania (the languages are presented as dots on the map in Figure 2, see also the list in Appendix 1). The study is confined to intransitive verbs in the present tense. Dual forms were excluded. As a result each language contributed six markers (1, 2, 3 singular and 1, 2, 3 plural).

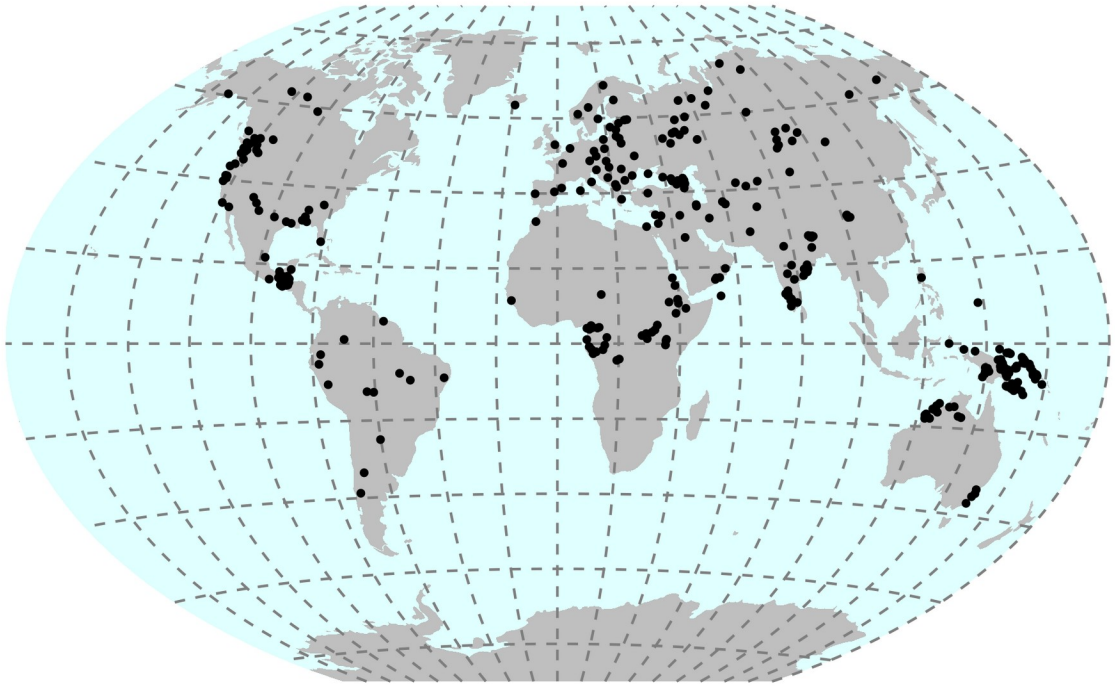

Figure 2. Distribution of languages from (Seržant, Moroz in preparation).[4]

3.2 Results

In order to test our attractor hypothesis we evaluated the Poisson regression model in order to model the observed relations between the length of indexes, person, and number. The first

[4] The plots in Figures 2 and 3 are provided after consulting and with agreement of the authors.

singular form was treated as a baseline for the regression. The `lme4` (Bates et al. 2015) formula used for this model is as follows:

(4) index length ~ person * number + (1|clade)

The overall predictions of our model are presented in Figure 3, with estimated values and a 95% confidence interval. The dots and triangles denote the estimated length of the corresponding personal index, e.g. the value of the first singular form is 1.65. This means that in general, this index tends to have a length of 1 or 2, but 2 is more likely. All variables in the model (person, number) except the interaction between second person and plural form are statistically significant. The non-significance of this interaction could be interpreted as a lack of any statistical difference between second singular and second plural forms. All other variables and their interactions differ significantly from zero.

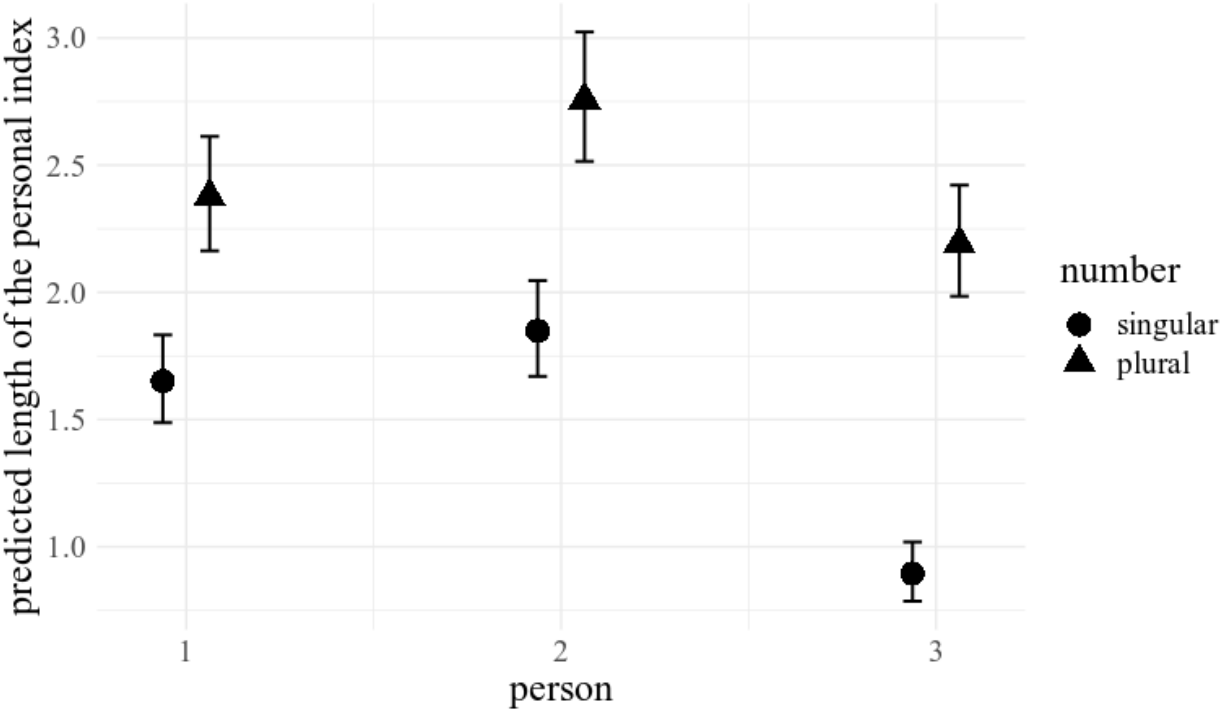

Figure 3. Poisson mixed effects model's predictions for the number of segments based on person and number (clade is used as a random effect). Triangles correspond to the average predicted value for plural forms. Dots correspond to the average predicted value for plural forms. The errorbar denotes the 95% confidence interval for the estimated mean.

In order to test whether there is a diachronic pressure towards the attractor lengths, we compared the lengths of each of the six person-number indexes in the respective proto-language reconstructed with the Historical-Comparative Method in authoritative literature. We tested this with the logistic regression model, which revealed that indeed in all persons there is a high probability to obey the attractor: all predicted probabilities for singular forms were higher than 90%, and higher then 50% for plural forms. We treated those results as evidence for the attractor hypothesis. From our token frequency analysis we can predict:

(5) significantly shorter forms for third person singular (compare with (1));
(6) not much difference in length between first and second persons (compare with (2));
(7) shorter indexes for singular forms (compare with (3)).

It appears that our token and type frequency analyses agree with each other, so in the following section I will show how the East Caucasian data fits to the attractor hypothesis.

4 East Caucasian data

4.1 Data

East Caucasian (Nakh–Dagestanian) is an indigenous language family spoken in a mountainous area in Dagestan, Chechnya, and Ingushetia (North Caucasian autonomous republics of the Russian Federation), as well as in the adjacent areas of Azerbaijan and Georgia. Traditionally it is divided into several branches (Avar-Andic, Dargwa, Lezgic, Nakh, Tsezic) and a few isolated languages (Lak, Khinalug) (Catford 1977; Hewitt 2004; Berg 2005; Daniel, Lander 2011; Ganenkov, Maisak 2020). Although in comparison to gender agreement, person agreement is rare across East Caucasian languages, although it is present in all branches of the East Caucasian family except Khinalug. In the current study I analyzed the following varieties:

- Nakh languages: Tsova-Tush (Bats), Kist dialect of Chechen;
- Andic languages: Akhvakh;
- Tsezic languages: Hunzib;
- All Dargwa idioms;
- All Lak dialects;
- Lezgic languages: Tabasaran, Udi;
- Zaqatala dialect of Avar.

Even though the dot to language representation is generally misleading (an alternative can be found in (Daniel et al. 2020)), I presented all data with the kind of map as in Figure 4. Figure 4 shows that the majority of languages with personal indexes are concentrated in central Dagestan, and not in the north or the south. Two dots located in Georgia (Tsova-Tush and the Kist dialect of Chechen) are colored black, since both lects have personal indexes, and its appearance could have been induced by the contacts with Georgians.

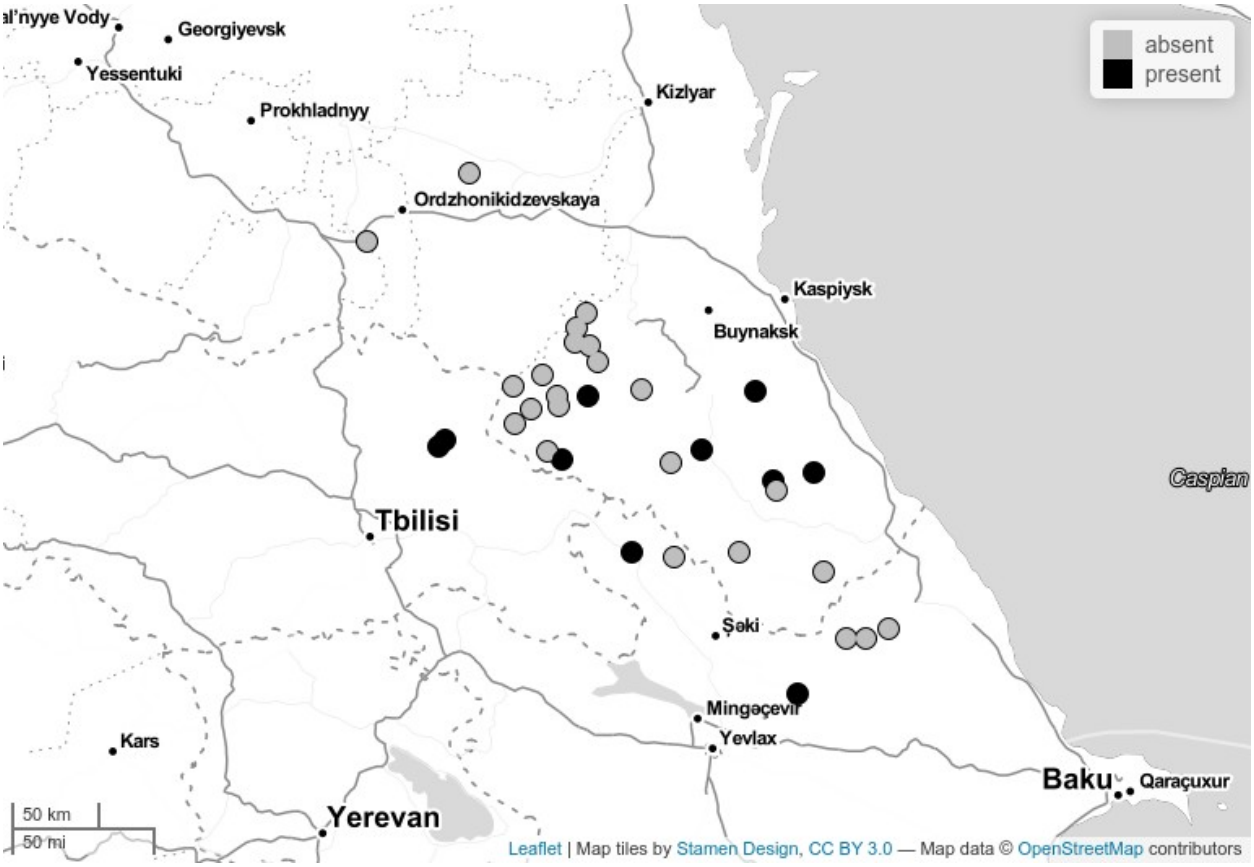

Figure 4. Distribution of East Caucasian languages with person-number indexes

Mimicking the data collection methodology from (Seržant, Moroz in preparation) I collected person-number indexes of intransitive verbs in the present tense from published sources. It is worth mentioning that markers found on other word forms in East Caucasian languages can have different markers (with different combinations of agreement categories, e.g. person and gender). At this point I decided to exclude Kist Chechen from the sample, since it has nonconcatenative person marking that is represented by an alternation of the initial vowel of verb stems (Алироев, Маргошвили 2006: 74–84). The result is summed up in Table 1, which contains nine East Cauacasian idioms (the whole list of languages is in Appendix 2):

Table 1. person-number indexes of East Caucasian languages.

language	1sg	2sg	3sg	1pl	2pl	3pl
Akhvakh	-do	-ri	-ri	-de	-ri	-ri
Avar (Zaqatala)	-ow, -ej, -eb	-a	-a	-al	-a	-a
North Dargwa (Sanzhi)	-d	-t:e	-	-d	-t:e	-

language	1sg	2sg	3sg	1pl	2pl	3pl
North Dargwa (Standard Dargwa)	-ra	-ri	-	-ra	-ri	-
Lak	-ra	-ra	-j	-ru	-ru	-j
Tabasaran	-za	-wa	-w	-tʃa	-tʃʷa	-w
Udi (Vartashen)	-zu	-nu	-ne	-jan	-nan	-q:un
Tsova-Tush	-s	-ħ	-	-txo	-aiʃ	-
Hunzib	-tʃo	-tʃo	-	-tʃo	-tʃo	-

It appears that person-number indexing in East Caucasian is a recent development (Schulze 2007 and others). For languages spoken in Georgia (Tsova-Tush, Kist Chechen) and Azerbaijan (Udi, Zakatala Avar) one can stipulate a contact-induced origin of person agreement. Person-number index development at least for some idioms located in Dagestan (Tabasaran and Hunzib) can also be explained through contact, since there has been extensive contact with Azerbijani and Georgian, respectively (Dobrushina 2016). It is more or less clear that the Tabasaran and Udi forms developed from personal pronouns, but this is not obvious for all other languages. For most of the cases the scenario was different: the markers present in Table 1 are drastically different from pronouns. It is also worth mentioning that there is a high level of number and person syncretism in East Caucasian languages, which makes it possible to analyze some person-number indexes systems as lacking the number category (e.g. all Dargwa and Hunzib). This is crucial for our analysis, since both token (Section 2) and type (Section 3) frequency analyses predicted longer indexes for plural forms.

4.2 Results

Let us now compare the length of East Caucasian person-number indexes with the predictions from Sections 2 and 3. In Figure 5 all data are presented together. This plot consists of six subplots for each combination of person and number. On the y-axis is the length of the personal index, and on the x-axis is the source of our data: either East Caucasian languages from Table 1 or (Seržant, Moroz in preparation). Each East Caucasian language is represented with one dot, which is randomly offset on the x-axis in order to avoid the overplotting of dots. The asterisk denotes the average length for a particular combination of person and number. The predictions from (Seržant, Moroz in preparation) are represented with errorbars (same as in Figure 3), which shows a 95% confidence interval for the estimated mean.

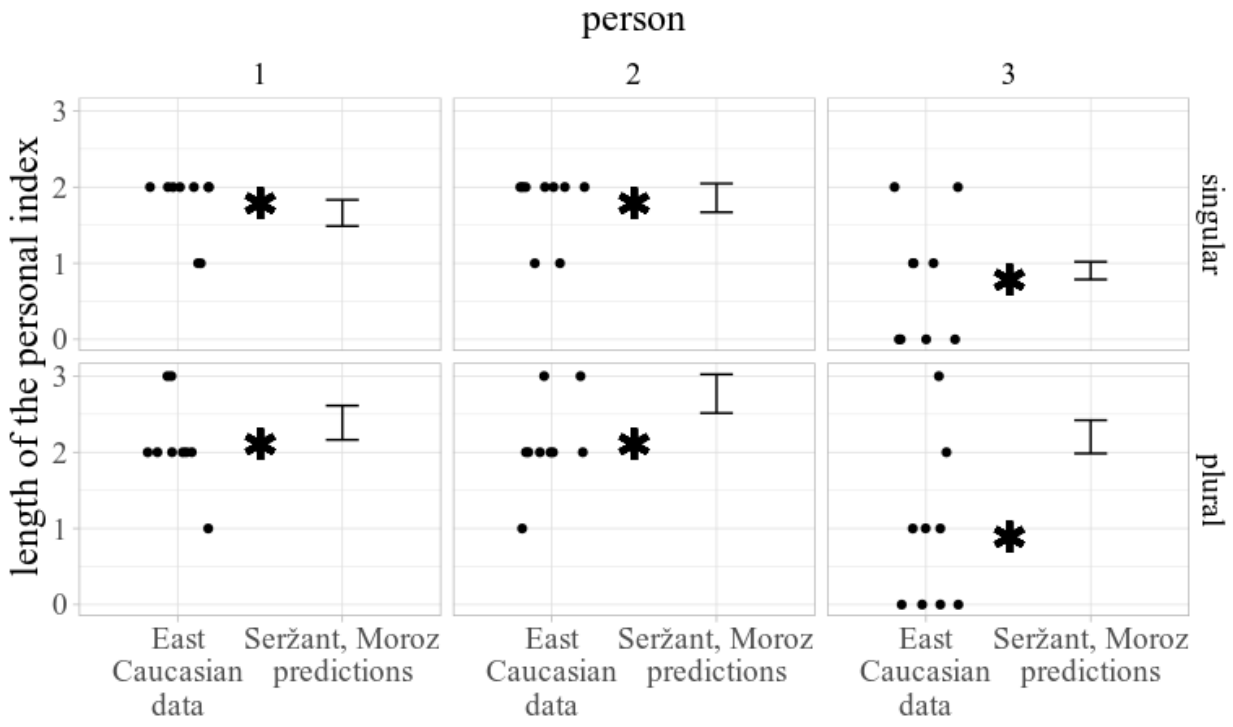

Figure 5. Comparison of East Caucasian person-number indexes' length (points) with predictions from (Seržant, Moroz in preparation) (errorbar). The asterisk denotes the mean length for the East Caucasian data.

Analyzing the difference between the asterisk value and the values predicted by the errorbars in Figure 5, we can see that only the average personal index length of the second and third plural forms deviate from the predictions. This not an unexpected deviation if we recall the observation from the previous section, that East Caucasian languages have a high level of number syncretism (from Table 1 it is obvious that only Udi distinguishes third singular and plural forms). As we can see, East Caucasian thus meets our expectations about the length of the third singular form (see (1) and (5)), and the first and second person forms (see (2) and (6)), while it violates our expectations about number (see (3) and (7)).

5 Conclusions

In this paper I analyzed the length of East Caucasian person-number indexes and checked whether they align with universal tendencies found in other language families. In order to present a broader context for the East Caucasian data and formulate some expectations about the length of person-number indexes, I provided a hypothesis that the length of the personal index correlates with its frequency, which is a reasonable hypothesis within a usage-based approach (see (Tummers, Heylen, Geeraerts 2005)). Since there are two types of frequencies (token frequencies and type frequencies, see (Berg 2014)), I provided results of two surveys of person-number indexes that cover each type of frequency. The first survey covered token frequencies of person-number indexes based on Universal Dependencies treebanks (see (De Marneffe et al. 2014; Zeman et al. 2020)). The second survey covered type frequencies of person-number

indexes based on the typological research of person-number indexes in the languages of the world (Seržant, Moroz in preparation). The two surveys agree with each other and make it possible to formulate some expectations for the length of person-number indexes:

- third person forms are shorter than first and second person forms;
- first and second person forms are more or less the same length;
- singular forms within each person are shorter than plural forms.

After these preparations I presented East Caucasian data. Even though there are only ten languages out of 33 that have person-number indexes, the languages are distributed across all branches of East Caucasian (except the family-level isolate Khinalug). Person-number indexes in East Caucasian languages are a quite recent and independent development. They show a high level of number and person syncretism and as a result, they violate only one of our expectations: singular and plural forms within each person have more or less the same length.

These results demonstrate that even though both surveys predicted more or less the same, there is always room for expecting some exceptions. East Caucasian languages are a particularly clear example of such an exception. This exceptionality could be explained by the history of East Caucasian languages, and it is possible that adding historical information about the grammaticalization path of person-number indexes will increase the accuracy of the model.

References

Алироев, И. Ю., Маргошвили, Л. Ю. 2006. *Кистины*. Москва: Книга и бизнес.

Гасанова, С. М. 1962. *Глагол в даргинском языке*. Махачкала: Изд-во АН СССР.

Муталов, Р. О. 2002. *Глагол даргинского языка*. Махачкала: ИПЦ ДГУ.

Сумбатова, Н. Р. 2011. Грамматические особенности и проблема происхождения личных показателей в даргинском языке. In: *Вопросы языкового родства 67*.

Bates, D. et al. 2015. Fitting linear mixed-effects models using lme4. In: *Journal of Statistical Software 67*, p. 1–48.

Bentz, C., Ferrer-i-Cancho, R. 2016. Zipf's law of abbreviation as a language universal. In: *Proceedings of the leiden workshop on capturing phylogenetic algorithms for linguistics*. University of Tübingen. p. 1–4.

Berg, H. van den 1999. Gender and person agreement in Akusha Dargi. In: *Folia linguistica 33*, p. 153–168.

Berg, H. van den 2005. The East Caucasian language family. In: *Lingua 115*, p. 147–190.

Berg, T. 2014. On the relationship between type and token frequency. In: *Journal of quantitative linguistics 21*, p. 199–222.

Catford, J. C. 1977. Mountain of tongues: The languages of the Caucasus. In: *Annual Review of Anthropology 6*, p. 283–314.

Cooper, D. L. 1999. *Linguistic attractors: The cognitive dynamics of language acquisition and change*. John Benjamins Publishing.

Daniel, Michael et al. 2020. *Typological atlas of the languages of daghestan (TALD)*. Moscow: Linguistic Convergence Laboratory, NRU HSE.

Daniel, M., Lander, Y. 2011. The Caucasian languages. In: Kortmann, B., Auwera, J. van der (Ed.): *The languages and linguistics of Europe: A comprehensive guide*. Berlin: De Gruyter Mouton. p. 125–158.

De Marneffe, M.-C. et al. 2014. Universal Stanford dependencies: A cross-linguistic typology. In: *LREC*. p. 4585–4592.
Dobrushina, N. 2016. Multilingualism in highland Daghestan throughout the 20-th century. In: Sloboda, M., Zabrodskaja, A., Laihonen, P. (Ed.): *Sociolinguistic transition in former eastern bloc countries: Two decades after the regime change*. p. 75–94.
Ferrer-i-Cancho, R., Bentz, C., Seguin, C. 2020. Optimal coding and the origins of zipfian laws. In: *Journal of Quantitative Linguistics*, p. 1–30.
Foley, S. 2020. Agreement in Languages of the Caucasus. In: Polinsky, M. (Ed.): *The Oxford Handbook of Languages of the Caucasus*. Oxford University Press.
Ganenkov, D., Maisak, T. 2020. Nakh-Dagestanian languages. In: Polinsky, M. (Ed.): *The Oxford Handbook of Languages of the Caucasus*. Oxford University Press.
Hammarström, H. et al. 2020. *Glottolog 4.3.* Jena: Max Planck Institute for the Science of Human History.
Hay, J. 2002. From speech perception to morphology: Affix ordering revisited. In: *Language 78*, p. 527–555.
Helmbrecht, J. 1996. The syntax of personal agreement in East Caucasian languages. In: *STUF-Language Typology and Universals 49*, p. 127–148.
Hewitt, G. 2004. *Introduction to the Study of the Languages of the Caucasus*. Lincom.
Landauer, T. K., Streeter, L. A. 1973. Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. In: *Journal of verbal learning and verbal behavior 12*, p. 119–131.
Lüdecke, D. 2018. Ggeffects: Tidy data frames of marginal effects from regression models. In: *Journal of Open Source Software 3*, p. 772.
Moroz, G. 2017. *Lingtypology: Easy mapping for linguistic typology.*
Piantadosi, S. T., Tily, H., Gibson, E. 2011. Word lengths are optimized for efficient communication. In: *Proceedings of the National Academy of Sciences 108*, p. 3526–3529.
R Core Team 2020. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.
Schulze, W. 2007. *Personalität in den ostkaukasischen Sprachen*.
Seržant, I., Moroz, G. in preparation. *Efficiency-driven attractor states of verbal person-number indexes.*
Troubetzkoy, N. S. 1929. Notes sur les désinences du verbe dans les langues tchétchénolesghiennes (caucasiques-orientales). In: *Bulletin de la Société de Linguistique de Paris 29*, p. 153–171.
Tummers, J., Heylen, K., Geeraerts, D. 2005. Usage-based approaches in cognitive linguistics: A technical state of the art. In: *Corpus Linguistics and Linguistic Theory 1*, p. 225–261.
Wedel, A., Jackson, S., Kaplan, A. 2013. Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. In: *Language and speech 56*, p. 395–417.
Wickham, Hadley 2016. *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York.
Zeman, D. et al. 2020. *Universal dependencies 2.7.* Online unter: http://hdl.handle.net/11234/1-3424.
Zipf, G. K. 1936. Word lengths are optimized for efficient communication. Oxon: Routledge.

Appendix 1

List of languages from (Seržant, Moroz in preparation). Glottocode is a universal language identifier from the Glottolog database (Hammarström et al. 2020).

language	glottocode	clade1	area
Lithuanian	lith1251	Indo-European	Eurasia
Latvian	latv1249	Indo-European	Eurasia
Mgreek	mode1248	Indo-European	Eurasia
Ukrainian	ukra1253	Indo-European	Eurasia
Belarusian	bela1254	Indo-European	Eurasia
Russian	russ1263	Indo-European	Eurasia
Polish	poli1260	Indo-European	Eurasia
Kashubian	kash1274	Indo-European	Eurasia
Serbian-Croatian	sout1528	Indo-European	Eurasia
Czech	czec1258	Indo-European	Eurasia
Slovak	slov1269	Indo-European	Eurasia
Bulgarian	bulg1262	Indo-European	Eurasia
Macedonian	mace1250	Indo-European	Eurasia
Slovenian	slov1268	Indo-European	Eurasia
UpperSorbian	uppe1395	Indo-European	Eurasia
German	stan1295	Indo-European	Eurasia

language	glottocode	clade1	area
Dutch	dutc1256	Indo-European	Eurasia
English	stan1293	Indo-European	Eurasia
Swedish	swed1254	Indo-European	Eurasia
Norwegian	norw1259	Indo-European	Eurasia
Icelandic	icel1247	Indo-European	Eurasia
Ossetian	iron1242	Indo-European	Eurasia
Persian	west2369	Indo-European	Eurasia
Marathi	mara1378	Indo-European	Eurasia
Hindi	hind1269	Indo-European	Eurasia
Italian	ital1282	Indo-European	Eurasia
Portuguese	port1283	Indo-European	Eurasia
Spanish	stan1288	Indo-European	Eurasia
Catalan	stan1289	Indo-European	Eurasia
Romanche	fran1269	Indo-European	Eurasia
French	stan1290	Indo-European	Eurasia
Sardinian	sass1235	Indo-European	Eurasia
Rumanian	roma1327	Indo-European	Eurasia
Albanian	gheg1238	Indo-European	Eurasia

language	glottocode	clade1	area
Seimat	seim1238	Oceanic	Papua
Penchal	penc1239	Oceanic	Papua
Lele (Papua New Guinea)	lele1270	Oceanic	Papua
Ere	eree1241	Oceanic	Papua
Nyindrou	nyin1250	Oceanic	Papua
Kove	kove1237	Oceanic	Papua
Gitua	gitu1237	Oceanic	Papua
Sio	sioo1240	Oceanic	Papua
Tami	tami1290	Oceanic	Papua
Mindiri	mind1255	Oceanic	Papua
Gedaged	geda1237	Oceanic	Papua
Medebur	mede1237	Oceanic	Papua
Manam	mana1295	Oceanic	Papua
Biem	biem1237	Oceanic	Papua
Kaiep	kaie1237	Oceanic	Papua
Numbami	numb1247	Oceanic	Papua
Iwal	iwal1237	Oceanic	Papua
Yabem	yabe1254	Oceanic	Papua
Labu	labu1248	Oceanic	Papua
Wampar	wamp1247	Oceanic	Papua
Arifama-Miniafia	arif1239	Oceanic	Papua
Are	aree1239	Oceanic	Papua

language	glottocode	clade1	area
Wedau	weda1241	Oceanic	Papua
Iamalele	iama1237	Oceanic	Papua
Dobu	dobu1241	Oceanic	Papua
Duau	duau1237	Oceanic	Papua
Suau	suau1242	Oceanic	Papua
Kilivila	kili1267	Oceanic	Papua
Misima-Paneati	misi1243	Oceanic	Papua
Sudest	sude1239	Oceanic	Papua
Magori	mago1248	Oceanic	Papua
Sinaugoro	sina1266	Oceanic	Papua
Motu	motu1246	Oceanic	Papua
Bulu (Papua New Guinea)	bulu1253	Oceanic	Papua
Bola	bola1250	Oceanic	Papua
Notsi	nots1237	Oceanic	Papua
Barok	baro1253	Oceanic	Papua
Konomala	kono1269	Oceanic	Papua
Patpatar	patp1243	Oceanic	Papua
Label	labe1239	Oceanic	Papua
Minigir	mini1251	Oceanic	Papua
Kandas	kand1301	Oceanic	Papua
Nehan	neha1247	Oceanic	Papua
Solos	solo1257	Oceanic	Papua

language	glottocode	clade1	area
Petats	peta1245	Oceanic	Papua
Halia	hali1244	Oceanic	Papua
Torau	tora1259	Oceanic	Papua
Mono-Alu	mono1273	Oceanic	Papua
Varisi	vari1239	Oceanic	Papua
Ririo	riri1237	Oceanic	Papua
Babatana	baba1268	Oceanic	Papua
Yapese	yape1248	Oceanic	Papua
Liko	lika1243	Bantu	Africa
Nzadi	nzad1234	Bantu	Africa
Bonkeng	lund1274	Bantu	Africa
Duala	dual1243	Bantu	Africa
Benga	beng1282	Bantu	Africa
Basa (Cameroon)	basa1284	Bantu	Africa
Dimbong	dimb1238	Bantu	Africa
Nugunu (Cameroon)	nugu1242	Bantu	Africa
Eton-Mengisa	eton1253	Bantu	Africa
Makaa	maka1304	Bantu	Africa
Kwakum	kwak1266	Bantu	Africa
Mpongwe	mpon1255	Bantu	Africa
Kélé	kele1257	Bantu	Africa
Tsogo	tsog1243	Bantu	Africa

language	glottocode	clade1	area
Punu	punu1239	Bantu	Africa
Njebi	njeb1242	Bantu	Africa
Mbere	mber1262	Bantu	Africa
Teke-Tege	teke1275	Bantu	Africa
Ding	ding1239	Bantu	Africa
Ngundi	ngun1270	Bantu	Africa
Akwa	akwa1248	Bantu	Africa
Mand	atem1241	Sogeram	Papua
Apali	apal1256	Sogeram	Papua
Gants	gant1244	Sogeram	Papua
Nend	nend1239	Sogeram	Papua
Manat	payn1244	Sogeram	Papua
Mum	mumm1238	Sogeram	Papua
Sirva	sile1255	Sogeram	Papua
Aisian	aisi1234	Sogeram	Papua
Kulsab	fait1240	Sogeram	Papua
Mandobo Atas	mand1444	Awyu-Dumut	Papua
Yonggom	yong1280	Awyu-Dumut	Papua
Ketum-Wambon	ketu1239	Awyu-Dumut	Papua
Edera Awyu	eder1237	Awyu-Dumut	Papua
Aghu	aghu1255	Awyu-Dumut	Papua
Kombai	komb1274	Awyu-Dumut	Papua

language	glottocode	clade1	area
Japhug	jiar1240	Rgyalrongic-Kiranti	Eurasia
Situ	situ1238	Rgyalrongic-Kiranti	Eurasia
Khroskyabs	guan1266	Rgyalrongic-Kiranti	Eurasia
Northern Gyalrong	sida1238	Rgyalrongic-Kiranti	Eurasia
Northern Gyalrong	zbua1234	Rgyalrongic-Kiranti	Eurasia
Tshobdun	tsho1240	Rgyalrongic-Kiranti	Eurasia
Camling	caml1239	Rgyalrongic-Kiranti	Eurasia
Bantawa	bant1281	Rgyalrongic-Kiranti	Eurasia
Puma	puma1239	Rgyalrongic-Kiranti	Eurasia
Limbu	limb1266	Rgyalrongic-Kiranti	Eurasia
Belhariya	belh1239	Rgyalrongic-Kiranti	Eurasia
Sunwar	sunw1242	Rgyalrongic-Kiranti	Eurasia
Khaling	khal1275	Rgyalrongic-Kiranti	Eurasia

language	glottocode	clade1	area
Tamil	tami1289	Dravidian	Eurasia
Irula-Muduga	kada1242	Dravidian	Eurasia
Malayalam	mala1464	Dravidian	Eurasia
Irula of the Nilgiri	nort2701	Dravidian	Eurasia
Yerukula	yeru1240	Dravidian	Eurasia
Kota (India)	kota1263	Dravidian	Eurasia
Toda	toda1252	Dravidian	Eurasia
Kodava	koda1255	Dravidian	Eurasia
Kannada	nucl1305	Dravidian	Eurasia
Tulu	tulu1258	Dravidian	Eurasia
Koraga	kora1289	Dravidian	Eurasia
Bellari	bell1261	Dravidian	Eurasia
Telugu	telu1262	Dravidian	Eurasia
Northwestern Kolami	nort2699	Dravidian	Eurasia
Southeastern Kolami	naik1250	Dravidian	Eurasia
Duruwa	duru1236	Dravidian	Eurasia
Mudhili Gadaba	mudh1235	Dravidian	Eurasia
Maria (India)	gond1265	Dravidian	Eurasia
Konda	kond1295	Dravidian	Eurasia
Pengo	peng1244	Dravidian	Eurasia
Manda (India)	mand1413	Dravidian	Eurasia

language	glottocode	clade1	area
Kui (India)	kuii1252	Dravidian	Eurasia
Kurukh	kuru1302	Dravidian	Eurasia
Malto	malt1248	Dravidian	Eurasia
Brahui	brah1256	Dravidian	Eurasia
Finnish	finn1318	Uralic	Eurasia
Estonian	esto1258	Uralic	Eurasia
Liv	livv1244	Uralic	Eurasia
Votic	voti1245	Uralic	Eurasia
Ingrian	ingr1248	Uralic	Eurasia
North Saami	nort2671	Uralic	Eurasia
South Saami	sout2674	Uralic	Eurasia
Erzya	erzy1239	Uralic	Eurasia
Moksha	moks1248	Uralic	Eurasia
Mari (East Sepik Province)	east2328	Uralic	Eurasia
Selkup	selk1253	Uralic	Eurasia
Tundra Nenets	nene1249	Uralic	Eurasia
Forest Enets	enet1250	Uralic	Eurasia
Nganasan	ngan1291	Uralic	Eurasia
Kamas-Koibal	kama1351	Uralic	Eurasia
Komi-Zyrian	komi1268	Uralic	Eurasia
Komi-Permyak	komi1269	Uralic	Eurasia

language	glottocode	clade1	area
Udmurt	udmu1245	Uralic	Eurasia
Hungarian	hung1274	Uralic	Eurasia
Central Mansi	mans1258	Uralic	Eurasia
Surgut Khanty	khan1273	Uralic	Eurasia
North Azerbaijani	nort2697	Turkic	Eurasia
Tatar	tata1255	Turkic	Eurasia
Southern Altai	sout2694	Turkic	Eurasia
Nogai	noga1249	Turkic	Eurasia
Bashkir	bash1264	Turkic	Eurasia
Northern Altai	nort2686	Turkic	Eurasia
Gagauz	gaga1249	Turkic	Eurasia
Dolgan	dolg1241	Turkic	Eurasia
Kazakh	kaza1248	Turkic	Eurasia
Karaim	kara1464	Turkic	Eurasia
Karachay-Balkar	kara1465	Turkic	Eurasia
Kirghiz	kirg1245	Turkic	Eurasia
Crimean Tatar	crim1257	Turkic	Eurasia
Krymchak	krym1236	Turkic	Eurasia
Kumyk	kumy1244	Turkic	Eurasia
Taiga Sayan Turkic	kara1462	Turkic	Eurasia

language	glottocode	clade1	area
Tuvinian	tuvi1240	Turkic	Eurasia
Turkish	nucl1301	Turkic	Eurasia
Turkmen	turk1304	Turkic	Eurasia
Northern Uzbek	nort2690	Turkic	Eurasia
Uighur	uigh1240	Turkic	Eurasia
Urum	urum1249	Turkic	Eurasia
Khakas	khak1248	Turkic	Eurasia
Khalaj	khal1270	Turkic	Eurasia
Khorasani Turkish	khor1269	Turkic	Eurasia
Chuvash	chuv1255	Turkic	Eurasia
Chulym Turkic	chul1246	Turkic	Eurasia
Shor	shor1247	Turkic	Eurasia
Sakha	yaku1245	Turkic	Eurasia
Tigre	tigr1270	Semitic	Eurasia
Tigrinya	tigr1271	Semitic	Eurasia
Amharic	amha1245	Semitic	Eurasia
Argobba	argo1244	Semitic	Eurasia
Harari	hara1271	Semitic	Eurasia
Zay	zayy1238	Semitic	Eurasia
Gafat	gafa1240	Semitic	Eurasia
Sebat Bet Gurage	chah1248	Semitic	Eurasia
Classical Mandaic	clas1253	Semitic	Eurasia
Ancient Hebrew	anci1244	Semitic	Eurasia

language	glottocode	clade1	area
Western Neo-Aramaic	west2763	Semitic	Eurasia
Mehri	mehr1241	Semitic	Eurasia
Soqotri	soqo1240	Semitic	Eurasia
Harsusi	hars1241	Semitic	Eurasia
Jibbali	sheh1240	Semitic	Eurasia
Hobyót	hoby1242	Semitic	Eurasia
Egyptian Arabic	egyp1253	Semitic	Eurasia
Gilit Mesopotamian Arabic	meso1252	Semitic	Eurasia
Standard Arabic	stan1318	Semitic	Eurasia
Moroccan Arabic	moro1292	Semitic	Eurasia
Huastec	huas1242	Mayan	North America
Tzotzil	tzot1259	Mayan	North America
Chuj	chuj1250	Mayan	North America
Popti'	popt1235	Mayan	North America
Q'anjob'al	qanj1241	Mayan	North America
Mam	mamm1241	Mayan	North America
Aguacateco	agua1252	Mayan	North America
Poqomchi'	poqo1254	Mayan	North America

language	glottocode	clade1	area
Poqomam	poqo1253	Mayan	North America
Kekchí	kekc1242	Mayan	North America
K'iche'	kich1262	Mayan	North America
Achi	achi1256	Mayan	North America
Kaqchikel	kaqc1270	Mayan	North America
Uspanteco	uspa1245	Mayan	North America
Tz'utujil	tzut1248	Mayan	North America
Yucatec Maya	yuca1254	Mayan	North America
Itzá	itza1241	Mayan	North America
Mopán Maya	mopa1243	Mayan	North America
Lacandon	laca1243	Mayan	North America
Cholti	chol1283	Mayan	North America
Chortí	chor1273	Mayan	North America
Chol	chol1282	Mayan	North America
Tabasco Chontal	taba1266	Mayan	North America
Ixil	ixil1251	Mayan	North America
Tzeltal	tzel1254	Mayan	North America
Ngarinyin	ngar1284	Worrorran	Australia

language	glottocode	clade1	area
Worora	woro1258	Worrorran	Australia
Kwini	kwin1241	Worrorran	Australia
Wunambal	wuna1249	Worrorran	Australia
Unggumi	ungg1243	Worrorran	Australia
Choctaw	choc1276	Muskogean	North America
Chickasaw	chic1270	Muskogean	North America
Alabama	alab1237	Muskogean	North America
Koasati	koas1236	Muskogean	North America
Mikasuki	mika1239	Muskogean	North America
Mikasuki	hitc1239	Muskogean	North America
Afro-Seminole Creole	afro1254	Muskogean	North America
Creek	cree1270	Muskogean	North America
Navajo	nava1243	Athabaskan	North America
Mescalero-Chiricahua Apache	mesc1238	Athabaskan	North America
Hupa	hupa1239	Athabaskan	North America
Mattole	matt1238	Athabaskan	North America

language	glottocode	clade1	area
Kato	kato1244	Athabaskan	North America
Galice	gali1261	Athabaskan	North America
Chipewyan	chip1261	Athabaskan	North America
Sarsi	sars1236	Athabaskan	North America
North Slavey	nort2942	Athabaskan	North America
Dogrib	dogr1252	Athabaskan	North America
Tanaina	tana1289	Athabaskan	North America
Lillooet	lill1248	Salishan	North America
Shuswap	shus1248	Salishan	North America
Thompson	thom1243	Salishan	North America
Okanagan	okan1243	Salishan	North America
Okanagan	colv1241	Salishan	North America
Kalispel-Pend d'Oreille	kali1309	Salishan	North America
Spokane	spok1245	Salishan	North America
Coeur d'Alene	coeu1236	Salishan	North America
Columbia-Wenatchi	colu1250	Salishan	North America
Tillamook	till1254	Salishan	North America

language	glottocode	clade1	area
Upper Chehalis	uppe1439	Salishan	North America
Twana	twan1247	Salishan	North America
Southern Puget Sound Salish	sout2965	Salishan	North America
Northern Lushootseed	lush1252	Salishan	North America
Clallam	clal1241	Salishan	North America
Northern Straits Salish	clal1241	Salishan	North America
Northern Straits Salish	song1308	Salishan	North America
Halkomelem	halk1245	Salishan	North America
Squamish	squa1248	Salishan	North America
Sechelt	sech1246	Salishan	North America
Bella Coola	bell1243	Salishan	North America
Mauwake	mauw1238	Nuclear Trans New Guinea	Papua
Lavukaleve	lavu1241	isolate	Papua
Maybrat	maib1239	Maybrat-Karon	Papua
Tidore	tido1248	NorthHalmahera	Papua

language	glottocode	clade1	area
Yawa	nucl1454	Yawa-Saweru	Papua
Bilua	bilu1245	isolate	Papua
Terei	tere1278	South Bougainville	Papua
Iloko	ilok1237	Austronesian	Papua
Kuman	kuma1280	Nuclear Trans New Guinea	Papua
Korafe	kora1294	Nuclear Trans New Guinea	Papua
Marind	nucl1621	Anim	Papua
Maklew	makl1246	Bulaka River	Papua
Jaminjung-Ngaliwuru	djam1255	Mirndi	Australia
Nungali	nung1291	Mirndi	Australia
Jingulu	djin1251	Mirndi	Australia
Ngarnka	ngar1283	Mirndi	Australia
Wambaya	nucl1328	Mirndi	Australia
Gudanji	guda1243	Mirndi	Australia
Binbinka	binb1242	Mirndi	Australia
Dyaberdyaber	dyab1238	Nyulnyulan	Australia
Nyulnyul	nyul1247	Nyulnyulan	Australia
Bardi	bard1254	Nyulnyulan	Australia

language	glottocode	clade1	area
Nimanbur	nima1245	Nyulnyulan	Australia
Yawuru	yawu1244	Nyulnyulan	Australia
Nyigina	nyig1240	Nyulnyulan	Australia
Thurawal	thur1254	Southeastern Pama-Nyungan	Australia
Dhurga	dhur1239	Southeastern Pama-Nyungan	Australia
Gundungurra	gund1248	Southeastern Pama-Nyungan	Australia
Southern Coastal Yuin	sout2771	Southeastern Pama-Nyungan	Australia
Mapudungun	mapu1245	Araucanian	South_America
Pilaga	pila1245	Guaicuruan	South_America
Achuar	achu1248	Chicham	South_America
Aguaruna	agua1253	Chicham	South_America
Aikana	aika1237	isolate	South_America
Allentiac	alle1238	Huarpean	South_America
Yanesha'	yane1238	Arawakan	South_America
Wayana	gali1262	Cariban	South_America

language	glottocode	clade1	area
Fulnio	fuln1247	isolate	South_America
Kotiria	guan1269	Tucanoan	South_America
Arikapu	arik1265	Nuclear-Macro-Je	South_America
Xerente	xere1240	Nuclear-Macro-Je	South_America
Kayapo	kaya1330	Nuclear-Macro-Je	South_America
Jarawara	jara1276	Arawan	South_America
Nandi	nand1266	Nilotic	Africa
Kipsigis	kips1239	Nilotic	Africa
Markweeta	mark1255	Nilotic	Africa
Markweeta	endo1242	Nilotic	Africa
Pökoot	east2420	Nilotic	Africa
Biafada	biaf1240	Northern Atlantic	Africa
Tennet	tenn1246	Surmic	Africa
Awngi	awng1244	Afroasiatic	Africa
Buduma	budu1265	Afroasiatic	Africa
Fiyadikka	fiya1238	Nubian	Africa
Miza	miza1238	Central Sudanic	Africa
Ma'di	madi1260	Central Sudanic	Africa
Lugbara	lugb1240	Central Sudanic	Africa

language	glottocode	clade1	area
Mamvu	mamv1243	Central Sudanic	Africa
Lese	lese1243	Central Sudanic	Africa
Svan	svan1243	Kartvelian	Eurasia
Lahnda	lahn1241	Indo-European	Eurasia
Adyge	adyg1241	Abkhaz-Adyge	Eurasia
Lak	lakk1252	Nakh-Daghestanian	Eurasia
Kati	kati1270	Indo-European	Eurasia
Bondo	bond1245	Austro-Asiatic	Eurasia
Yukaghir	sout2750	Yukaghir	Eurasia
Acoma	west2632	Keresan	North America
Biloxi	bilo1248	Siouan	North America
Cahuilla	cahu1264	Uto-Aztecan	North America
Chumash	barb1263	isolate	North America
Cherokee	cher1273	Iroquoian	North America
Chimariko	chim1301	isolate	North America
Highland Oaxaca Chontal	high1242	Tequistlatecan	North America
Coosan	coos1249	Coosan	North America

Appendix 2

List of East Caucasian languages used in the survey. Glottocode is a universal language identifier from the Glottolog database (Hammarström et al. 2020).

lang	idiom	glottocode
Akhvakh		akhv1239
Andi		andi1255
Avar		avar1256
Avar	Zaqatala	zaqa1242
Bagvalal		bagv1239
Botlikh		botl1242
Chamalal		cham1309
Godoberi		ghod1238
Karata		kara1474
Karata	Tukita	kara1474
Tindi		tind1238
North Dargwa	Sanzhi	sanz1248
North Dargwa	Standard Dargwa	darg1241
Khinalug		khin1240
Lak		lakk1252
Agul		aghu1253
Archi		arch1244
Budukh		budu1248
Kryz		kryt1240
Lezgian		lezg1247
Rutul		rutu1240
Tabasaran		taba1259
Tsakhur		tsak1249

lang	idiom	glottocode
Udi		udii1243
Tsova Tush		bats1242
Chechen		chec1245
Chechen	Kist	chec1245
Ingush		ingu1240
Bezhta		bezh1248
Hinuq		hinu1240
Hunzib		hunz1247
Khwarshi - Inkhoqwari	Khwarshi	khva1239
Tsez		dido1241