The AUTOTYP Genealogy and Geography Database 2013 Release[*]

Johanna Nichols[a], Alena Witzlack-Makarevich[b], and Balthasar Bickel[b]

[a]University of California, Berkeley
[b]University of Zürich

1 Introduction

The AUTOTYP genealogy and geography database contains information on the names, genealogy and geography of about 2700 languages and language varieties worldwide. The database is available for download as a comma- or tab-delimited file (`autotyp.csv`, `autotyp.tab`) that can be loaded into spreadsheet, database or statistics applications, and as a set of color-coded KML files (`autotyp.kml`, `autotyp.areas.kml`, `autotyp.continents.kml`) that can be opened and explored in Google Earth (`http://earth.google.com/`).[1] Each language entry has a unique ID in a field named `LID` and is matched, wherever possible, to `ISO639.3` codes. The following explains the rationale behind the genealogical and geographical classifications and the content of the relevant database fields.

2 Genealogy

Our genealogy database does not rely on geographical groupings, widely mentioned but unproven groupings, or similar hypothetical groupings, and in this respect differs from such classifications as Lewis et al. (2013) and Dryer & Haspelmath (2011). (In the few cases where we felt it necessary to include a label for a residual grouping that is not a clade, such as Western Malayo-Polynesian, we have included "non-clade" in the name of the group.) It is similar to the classifications of Campbell (1997, 2007),

[*] Cite this resource as:
Nichols, Johanna, Alena Witzlack-Makarevich & Balthasar Bickel. 2013. The AUTOTYP genealogy and geography database: 2013 release. `http://www.spw.uzh.ch/autotyp/`.
Comments and references are always welcome at `johanna@berkeley.edu`.

[1] All files are encoded as `utf-8` with `LF` line breaks and without a byte order marker (`BOM`).

Campbell & Poser (2008), Campbell (2012), Hammarström (2012), but does not attempt to include languages or families for which there is classificatory information but little or no typological information (e.g. Beothuk, Cayuse) and does not include languages with typological information available that happen not to be in our database. It is constantly updated based on critical assessment of recent publications, and the number of languages in the database continues to grow. The genealogical database uses two classificatory levels (and corresponding database fields) that are cross-linguistically comparable:

`language`: In this database a language is actually a dialect or variety, in that individual dialects are entered as languages (with alternative names, if known). That is, each language has its own ID number (`LID` 'language ID') distinct from that of the pan-dialectal language: e.g. `LID` 87 German, 1227 Upper Austrian German, 1295 Berlin German, 1310 Zürich German, 2845 Old High German, etc. (The pan-dialectal language is identified as the lowest taxonomic level, the 'dialect group', but this information is incomplete in many areas and has not been included in the current release.) The general criterion for a separate language record in the database is that it has a distinct property in at least one of the typological variables that we survey; this may or may not coincide with sociolinguistic criteria.

`stock`: This is the highest-level language family that satisfies the two criteria of demonstrability and reconstructability. A grouping is demonstrable if there is evidence showing that it is a family, or its resemblances significantly exceed what can be expected by chance or from typological or universal principles. It is reconstructable if it exhibits recurrent sound correspondences so that reconstruction is possible in principle. (For this definition of the stock see Nichols 1997.) Stocks do not all have the same internal age: for instance, Indo-European is about 6000 years old (since its dispersal), Uralic probably similar in age, Nakh-Daghestanian older; Mayan and Muskogean are probably around 4000 years old; and Chumashan, Berber, Japanese-Ryukyuan, and Quechuan are all about 2000 years old or less. All of these count as stocks because they are, within their own genealogical lines, the maximal clades that are both demonstrable and reconstructable. If firm external kin for any of these are ever found (such that the higher families are both reconstructable and demonstrable), then the present stocks will be demoted to branches and the new groupings set up as stocks.

In defining the stock with a determinate upper limit and separately demanding both demonstrability and reconstructability, we differ from all other classification databases.

We have currently 328 stocks that we consider valid and use in classification. Note that every isolate and every unclassified language counts as a separate stock. Most

> of our research effort on classification has gone into identifying valid stocks. The list of stocks changes as descriptive and classificatory work proceeds; for instance, the former Reefs-Santa Cruz stock has recently been demoted to a branch of Oceanic (cf. Ross & Næss 2007). We anticipate that the number will eventually stabilize at around 300. Campbell (2007) and Hammarström (2012) give figures similar to ours.
>
> We also count every creole ($N = 32$) and every sign language ($N = 39$) in the database as a separate stock (in addition to the 328 other stocks).

Between these two levels, AUTOTYP uses convenience levels that are not comparable and not necessarily complete. We attempt to include the results of current classificatory work where that is more or less definitive, but the higher-level subgrouping of old families such as Indo-European, Sino-Tibetan and Austronesian is a perennially difficult issue for which a typological database may never be able to cite a stable received view. The subgroupings we use have the following arbitrary labels and field names:

`mbranch`: Major branch, e.g. for Indo-European the major branches are Germanic, Balto-Slavic, Indo-Iranian, etc. For Austronesian they are the three Formosan groups Atayalic, Tsouic, and Paiwanic as well as Malayo-Polynesian.

`sbranch`: Subbranch

`ssbranch`: Sub-subbranch

`lsbranch`: Lowest subbranch

An additional taxonomic level and field name is `dialect.group` but this information needs further reviewing before it may be included in a future release.

The actual AUTOTYP database has fields for discussion notes and references on classification, not included here. It also provides for one genealogical level above the stock, for families that are demonstrable but not (yet) reconstructable and for very likely older groupings. These include Afroasiatic (securely demonstrated but not uncontroversially reconstructable), Niger-Congo (likely), California-Plateau Penutian (which we regard as close to demonstrated), Dene-Yeniseian (statistically demonstrated in our view but raising many questions), Indo-Uralic (likely), and a very few others. These are not included here but may be included in later releases. For evaluation of those we know about see Nichols (2010).

Note that many languages in the database are not fully classified: these languages belong to a stock for which some internal classification is given in the database, but there is no information in the database about the position of these particular languages within that classification. This either means that the current state of the art does not allow further classification or that we have not yet reviewed the relevant literature. Wherever possible, full classifications are gradually being added.

3 Geography

The geography database contains information on the geographical location of languages and a small-scale and a large-scale classification of languages into areas. The geographical location is given by coordinates in decimal degrees, with negative numbers for longitudes west of Greenwich and latitudes south of the equator. The precision varies. When two languages are spoken in the same location, the coordinates are entered as slightly separated to allow for convenient plotting. When a language is spoken over a large territory, the coordinates approximate what is generally considered to be the sociolinguistic center of the language.

The area classifications are based on our assumptions about contact events in history, informed by current knowledge of the historical, genetic, anthropological, and archeological record. We try to keep our definition of areas free of linguistic information in order to avoid circularity in areal linguistics research (cf. Bickel & Nichols 2006 for discussion). The following lists the definition of continent-sized areas or 'macro-areas' (registered in the field `continent` in the database). A rough impression of these areas is provided in Map 1, but for detailed inspection we recommend loading the file `autotyp.continents.kml` into Google Earth.

Africa: All of Africa.

W and SW Eurasia: Western and Southwestern Eurasia. Including the Caucasus and following the northern shore of the Black Sea and the Carpathians to the Baltic sea. Also includes the languages of Ancient Anatolia and Mesopotamia east to the Hindu Kush (including the Iranian plateau).

N-C Asia: Northern and Central Asia. Siberia, the Tibetan plateau, Mongolia, Korea, and Japan.

S/SE Asia: South and Southeast Asia. The Indic subcontinent (including the Himalayas, but not the Tibetan plateau), mainland Southeast Asia, and insular Southeast Asia up to the Wallace Line (between Sumbawa and Flores).

NG and Oceania: New Guinea and Oceania (all islands east of the Wallace Line).

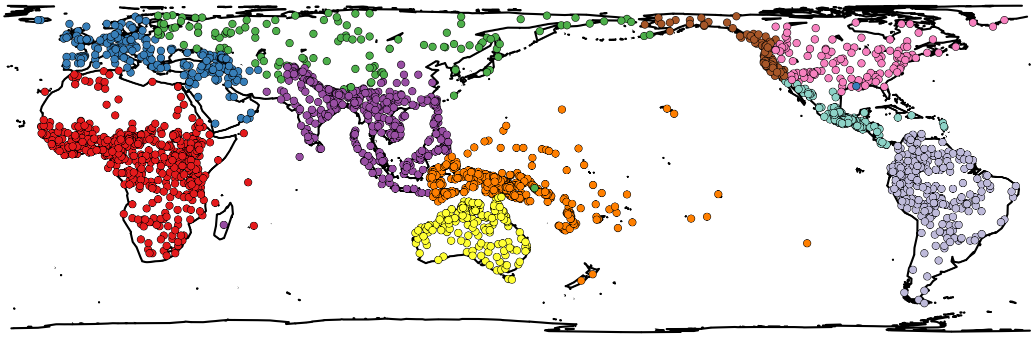
Australia: All of Australia.

W N America: Western North America. From the Pacific coast (excluding the Aleutian islands, including only mainland Alaska) to the lower eastern slope of the major coastal range (Rockies, Cascades, Sierras), thus comprising the coastal and intermontane regions. The southern boundary of North America runs along the US/Mexican border.

E N America: Eastern North America. From the lower eastern slope of the major coastal range to the Atlantic coast (thus comprising the Great Plains and eastern US and Canada). The southern boundary runs along US/Mexican border.

C America: Central America. The southern boundary is the Panama Canal; the northern boundary is the U.S.-Mexico border. Perhaps this boundary should eventually be placed at the northern limit of the Mesoamerican cultural area.

S America: All of South America.



Map 1: Continent-sized areal classification ($N = 10$)

Smaller-scale areas (registered in the field `area` in the database) are listed in the following. Map 2 again gives a rough impression of these areas, but for any closer inspection, we refer to the Google Earth file `autotyp.areas.kml`.

N Africa: Northern Africa. North of about 14 N (the latitude of the northern end of Lake Chad and north of the Niger, just north of Khartoum, without Eritrea), but excluding the Ethiopian Plateau (i.e. no farther east than about 35).

Greater Abyssinia: Between about 5 N and about 17 N, and east of about 35 E (this line slants up to about 38 E at 17 N). Includes the Ethiopian plateau and northern Somalia (and possibly also the Yemen tip of the Arabian peninsula).

African Savannah: The savannah north of the African rain forest. Between about 15 N and about 5 N, and east to about 35 E.

S Africa: Southern Africa. South of about 4.5 N and east of about 10 E. Includes the rain forest and everything to the south of it. (The northwest corner is just southeast of Cameroon Mountain, or at about Douala, Cameroon. The northeast border is from Mogadishu to the Ethiopian border.)

Europe: Europe up to the Eurasian steppe. Starting from the Black Sea, the eastern border of Europe follows the eastern slopes of the Carpathian mountains up to about the 22 E and then follows the Wisla up to the Baltic see; and from there we draw a straight line north. (This puts Poland inside Europe, based on its tight historical interaction with the rest of Europe; but one could as well draw the limit following the Oder-Elbe line.)

Greater Mesopotamia: Anatolia, Caucasus, Near East, Mesopotamia, and Iran. From the northern limits of the Black and Caspian seas south to ancient Anatolia (modern Turkey) and Mesopotamia and the Iranian Plateau. Includes the Arabian peninsula.

Indic: The Indic subcontinent. Extends to the eastern border of the Iranian Plateau in the west, to the south-oriented valleys of the Himalayas, the Karakorum, and the Pamirs and also those of the high-altitude Hindu Kush (but excluding Afghanistan and the Tibetan plateau) in the north, and to the Brahmaputra in the east.

Inner Asia: Inner Siberia and Central Asia: western Turkestan and Afghanistan but not the Iranian plateau, and also the Eastern Eurasian steppe up to the Carpathians; the Tibetan Plateau; Mongolia (inner and outer); most of Siberia.

Southeast Asia: From the Brahmaputra to near insular Southeast Asia (up to the Wallace Line, between Sumbawa and Flores, thus including Sumatra, Java, and Borneo) and most of China (but not Qinghai, Tibet, Xinjiang or Inner Mongolia).

N Coast Asia: Coastal Northern Asia. Manchuria, Korea, Japan, and Siberia up to about 100 miles or 160 km inland from the coast. Also including the Aleutian Islands.

N Coast New Guinea: Northern coastal New Guinea (up to within about 20 miles of the coast) including the Bird's Head peninsula; small offshore islands (e.g. Karkar) and the Bismarcks (New Britain, New Ireland) but not the Solomons.

Interior New Guinea: The highlands and most of their north slope.

S New Guinea: Southern New Guinea. The southern slope and coast.

Oceania: Pacific islands, including the Solomons. Western boundary is the Wallace Line (about 120 E), so the Philippines, Sulawesi, Flores, Sumba, Timor, etc. are in Oceania.

N Australia: Northern Australia. North of 18 S. (This is an artificial boundary and could probably be improved.)

S Australia: Southern Australia. South of 18 S. (This is an artificial boundary and could probably be improved.)

Alaska-Oregon: Western North America (up to the crest of the major coastal range) from mainland Alaska (not including the Aleutian Islands in the Bering Sea) through the Pacific Northwest down to southern Coos Bay (43 N).

California: Defined culturally as the area where acorns were the main basis of the economy. Eastern boundary: the crest of the Sierras. Northern boundary: Coos Bay (latitude 43 N). Otherwise congruent with the state of California.

Basin and Plains: The southwestern desert and semidesert and the prairie, extending into interior northern Canada. The southern extent needs to be defined; Coahuiltec (which straddles the Texas-Mexico border) has for now been put in this area. Approximate eastern boundary: line from Matamoros/Brownsville or Houston to Duluth (western tip of Lake Superior). Approximate northern boundary: from Duluth to the eastern Rocky Mountains foothills at about 58 N. The western boundary runs along the crest of the major coast range (Sierra Nevada, Cascades), so the area includes the old east slope riverine and lacustrine environment and its immediate periphery.

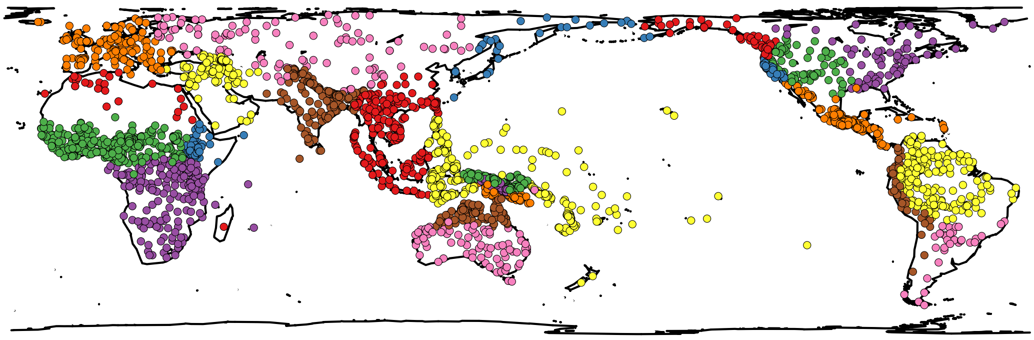
E North America: Southeast and Eastern North America. The Gulf Coast and eastern woodlands. East of a line from Brownsville/Matamoros at the Mexico/US border to Duluth on the western tip of Lake Superior to about Fort Nelson (58 N, 122 W).

Mesoamerica: The southern boundary is the Panama Canal; the northern boundary is the US/Mexico border. Perhaps this boundary should eventually be placed at the northern limit of the Mesoamerican cultural area.

NE South America: The northeast of South America. From the eastern slope of the Andes to the Caribbean and Atlantic coasts (between Caracas and Port of Spain), south to 15 S (this is an arbitrary demarcation of the Amazon and La Plata drainages).

Andean: Andean South America. The western coast, the Andes, and their eastern slope, south of the border of Panama (about 5 N). Southern boundary: about 50 S.

SE South America: The southeast of South America. From the eastern slopes of the Andes to the Atlantic coast, south of 15 S. Includes all of Patagonia and the Gran Chaco, and all of Tierra del Fuego.



Map 2: Smaller-scale areal classification ($N = 24$)

Areas larger than continents can be derived by composition. The Pacific Rim macroarea for example (as described in Bickel & Nichols 2006) consists of several smaller areas spanning parts of several macroareas:

- Oceania
- N Coast New Guinea
- N Australia
- N Coast Asia
- Alaska-Oregon
- California
- Mesoamerica
- Andean

and optionally also Southeast Asia, which may or may not actually belong in the area. Another example is the Trans-Pacific area (called “Circum-Pacific” in Bickel & Nichols 2006). This area is composed of the continents “NG and Oceania”, “Australia”, and all areas in the Americas.

References

Bickel, Balthasar & Johanna Nichols. 2006. Oceania, the Pacific Rim, and the theory of linguistic areas. *Proc. Berkeley Linguistics Society* 32.

Campbell, Lyle. 1997. *American Indian languages: the historical linguistics of Native America.* New York: Oxford University Press.

Campbell, Lyle. 2007. How many language families are there in the world? Presented at International Congress of Historical Linguists, Montreal.

Campbell, Lyle. 2012. The classification of South American languages. In Verónica Grondona & Lyle Campbell (eds.), *The indigenous languages of South-America: a comprehensive guide*, Berlin: De Gruyter Mouton.

Campbell, Lyle & William J. Poser. 2008. *Language classification: history and method.* Cambridge: Cambridge University Press.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2011. *The World Atlas of Language Structures Online.* Munich: Max Planck Digital Library [`http://wals.info/`] 2011th edn.

Hammarström, Harald. 2012. *The language families of the world.* Ms. Max-Planck-Institute for Evolutionary Anthropology, Leipzig.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2013. *Ethnologue: languages of the world, 17th edition.* Dallas: SIL International.

Nichols, Johanna. 1997. Modeling ancient population structures and population movement in linguistics and archeology. *Annual Review of Anthropology* 26. 359–384.

Nichols, Johanna. 2010. Language families, macroareas, and contact. In Raymond Hickey (ed.), *The handbook of language contact*, 361–379. London: Blackwell.

Ross, Malcolm & Åshild Næss. 2007. An Oceanic origin for Äiwoo, the language of the Reef Islands? *Oceanic Linguistics* 46. 456–498.