The Glottolog Data Explorer: Mapping the world's languages

Andrew Caines[1], **Christian Bentz**[2], **Dimitrios Alikaniotis**[1],
Fridah Katushemererwe[3], **Paula Buttery**[1]
[1]Department of Theoretical & Applied Linguistics, University of Cambridge, U.K.
[2]Institute of Linguistics, Universität Tübingen, Germany
[3]Department of Linguistics, Makerere University, Uganda
apc38@cam.ac.uk, chris@christianbentz.de, da352@cam.ac.uk, katu@chuss.mak.ac.ug, pjb48@cam.ac.uk

Abstract

We present THE GLOTTOLOG DATA EXPLORER, an interactive web application in which the world's languages are mapped using a JavaScript library in the 'Shiny' framework for R (Chang et al., 2016). The world's languages and major dialects are mapped using coordinates from the Glottolog database (Hammarström et al., 2016). The application is primarily intended to portray the endangerment status of the world's languages, and hence the default map shows the languages colour-coded for this factor. Subsequently, the user may opt to hide (or re-introduce) data subsets by endangerment status, and to resize the datapoints by speaker counts. Tooltips allow the user to view language family classification and links the user to the relevant Glottolog webpage for each entry. We provide a data table for exploration of the languages by various factors. The web application is freely available at `http://cainesap.shinyapps.io/langmap`

Keywords: Glottolog, world languages, endangered languages

1 Introduction

When it comes to the cartographic visualization of language resources, one type of resource that lends itself well to such an exercise is the typological database, as demonstrated, for example, by *The World Atlas of Language Structures* (Dryer and Haspelmath, 2013). Here, we present THE GLOTTOLOG DATA EXPLORER (GDE), an interactive visualization of the *Glottolog* database (Hammarström et al., 2016).

It is primarily intended to draw attention to the endangered status of many of the world's languages. Our aim is to illustrate the huge number of distinct languages around the world in the present day, and to imply that there is much to be lost, socially and culturally, if the languages classified as currently vulnerable or endangered were allowed to shrink away into extinction. Hence we allow the user to easily remove the vulnerable, endangered and extinct languages from the map and immediately visualize the projected scenario (see section 3).

The intended audience includes both linguistics specialists and the wider public. For the latter audience the GDE has to be visually appealing, interactive and easy to understand. At the same time we wish the GDE to be useful to the former group, at least to give an insight to certain research questions if not to answer them. We discuss our plans in this regard in section 5, and welcome feedback as to how we can make the GDE a useful research tool.

2 Glottolog

GLOTTOLOG is a catalogue of the world's languages, curated by members of the Max Planck Institute for Evolutionary Anthropology. It contains typological, geographic and bibliographic information for each so-called *languoid*. The creators of Glottolog chose this term to indicate that they are not just cataloguing *languages*, but any dialect, language or language family "that linguists need to be able to identify"[1]. Every languoid is curated into the catalogue with a list of any relevant linguistic works (grammars, dictionaries, *etc*), its classification in 'the Glottolog tree' (a linguistic genealogy), and – crucially for our purposes – latitude and longitude coordinates which allow its location to be mapped. Glottolog is freely available, regularly updated, and welcomes contributions from the linguistic community.

2.1 Data collection

We obtained Glottolog languoid data using a two-step process. First, we retrieved a full list of languoids from their resource map in JSON format[2]. At the time of writing there were a total of 22,924 languoids in the Glottolog catalogue. We subsequently looped through each languoid code, downloading the data made available by Glottolog, again in JSON format[3]. For each languoid we checked for the presence of longitude and latitude values, retaining only those languoids with these variables populated – that which would allow us to map their location in the GDE. This step excluded 15,295 unsituated languoids, the bulk of which are 10,414 dialects taken from the Multitree project (The LINGUIST List, 2014), not yet systematically cleaned of errors, geo-located, and properly included in the Glottolog catalogue.

The remainder are made up of 4112 language families, which would be even trickier than languages to geo-locate, and 769 languages without longitude or latitude values, many of which are so-called 'bookkeeping' languoids retained in the catalogue for the sake of completeness, even though they have for some reason been withdrawn from the 'live' language ontology. For instance ACHI′, CUBULCO

[1]Source: Glottolog website, accessed 2016-03-24

[2]`http://glottolog.org/resourcemap.json?rsc=language`

[3]For example, the URL for LATIN, which has the Glottolog identifier `lati1261`: `http://glottolog.org/resource/languoid/id/lati1261.json`

was merged with ACHI in 2008, as it is "a dialect or dialect group name, and therefore incorrect for a language designation"[4].

Thus we are left with 7629 languoids of more certain status and associated with geo-coordinates. Of these 221 are 'bookkeeping' languoids, and so we exclude them from the GDE. Our final dataset therefore contains 7407 entries, for which we present some high-level descriptive statistics regarding language type in Table 1 in the manner of the Glottolog information page[5]. The user may view the languoids for each language type listed in Table 1 by searching for that type in the GDE TABLE tab (see section 3).

Type	Count
Spoken L1 language	7183
Unattested	46
Unclassifiable	18
Pidgin	18
Mixed language	15
Artificial language	3
Speech register	3
Sign language	121
Total	*7407*

Table 1: Geo-located languoids extracted from Glottolog

For each languoid in our set of 7407, we collected the values needed for our web application: its Glottolog alphanumeric identifier, its assigned latitude and longitude coordinates, its higher-level genealogical classification, and its endangerment status on the UNESCO scale (Moseley, 2010). UNESCO's six degrees of vitality/endangerment are: extinct, critically endangered, severely endangered, definitely endangered, vulnerable, and safe. The Glottolog curators added 'unknown' for those languoids not featured in UNESCO's database, and replaced 'safe' with 'living', quite understandably given that 'safe' implies a certain stability for these languages, even those which may be on the borderline with the 'vulnerable' category.

Regarding genealogical classification, we selected at most the three highest classes for each languoid, where by 'highest' we mean the super-groupings such as Indo-European, Afro-Asiatic, Sino-Tibetan, Austronesian, and so on. Many languoids are classified to greater than three levels, maximally seventeen in fact, but the decision was taken to limit our data collection in this way so that the resulting user experience was not too unwieldy.

The information we hold for each languoid is of the following form:

```
name:  Senara Sénoufo
id:  sena1262
latitude:  10.4987
longitude:  -5.28216
family1:  Atlantic-Congo
family2:  Volta-Congo
family3:  North Volta-Congo
status:  Living
```

Our collected languoid information is presented to the user in two forms: as pop-up 'tooltips' for any selected languoid on the MAP tab, and alternatively in list format in on the TABLE tab (see section 3). Every tooltip contains a link to the languoid's Glottolog webpage, on which all associated information, including full family classification, is given. We last accessed the Glottolog catalogue on 2016-02-09, and will continue to regularly download the latest Glottolog data and ensure GDE contains up-to-date information.

3 Glottolog Data Explorer

The GDE application was written in R (R Core Team, 2015) using an interface to the LEAFLET JavaScript library (Cheng and Xie, 2015) and developed as a web application in the SHINY framework (Chang et al., 2016). It is hosted on SHINYAPPS servers and is freely available at `http://cainesap.shinyapps.io/langmap`.

The MAP itself is a layered widget starting with a Stamen basemap, lines and labels[6]. The decision to use Stamen is purely aesthetic: we found that our data points showed up best on these map tiles. Other basemaps we considered were OpenStreetMap, Esri's National Geographic map, and NASA's 'Earth at night' (Figure 1).

Figure 1: Considered basemap tiles, clockwise from bottom-left: NASA's 'Earth at night', OpenStreetMap, Esri National Geographic, Stamen watercolour.

We add the languoids to the basemap as geo-located markers coloured and layered according to endangerment status. Colour choices were quite straightforwardly grey for languoids of unknown status, shades of red to purple for the

[4]Source: `http://glottolog.org/resource/languoid/id/achi1258`; accessed 2016-02-11

[5]Source: `http://glottolog.org/glottolog/glottologinformation`; accessed on 2016-02-11.

[6]Made available by Stamen Design under a Creative Commons Attribution (CC BY 3.0) licence, with data from OpenStreetMap under a Creative Commons Attribution-ShareAlike (CC BY-SA 3.0) licence. See `http://maps.stamen.com`

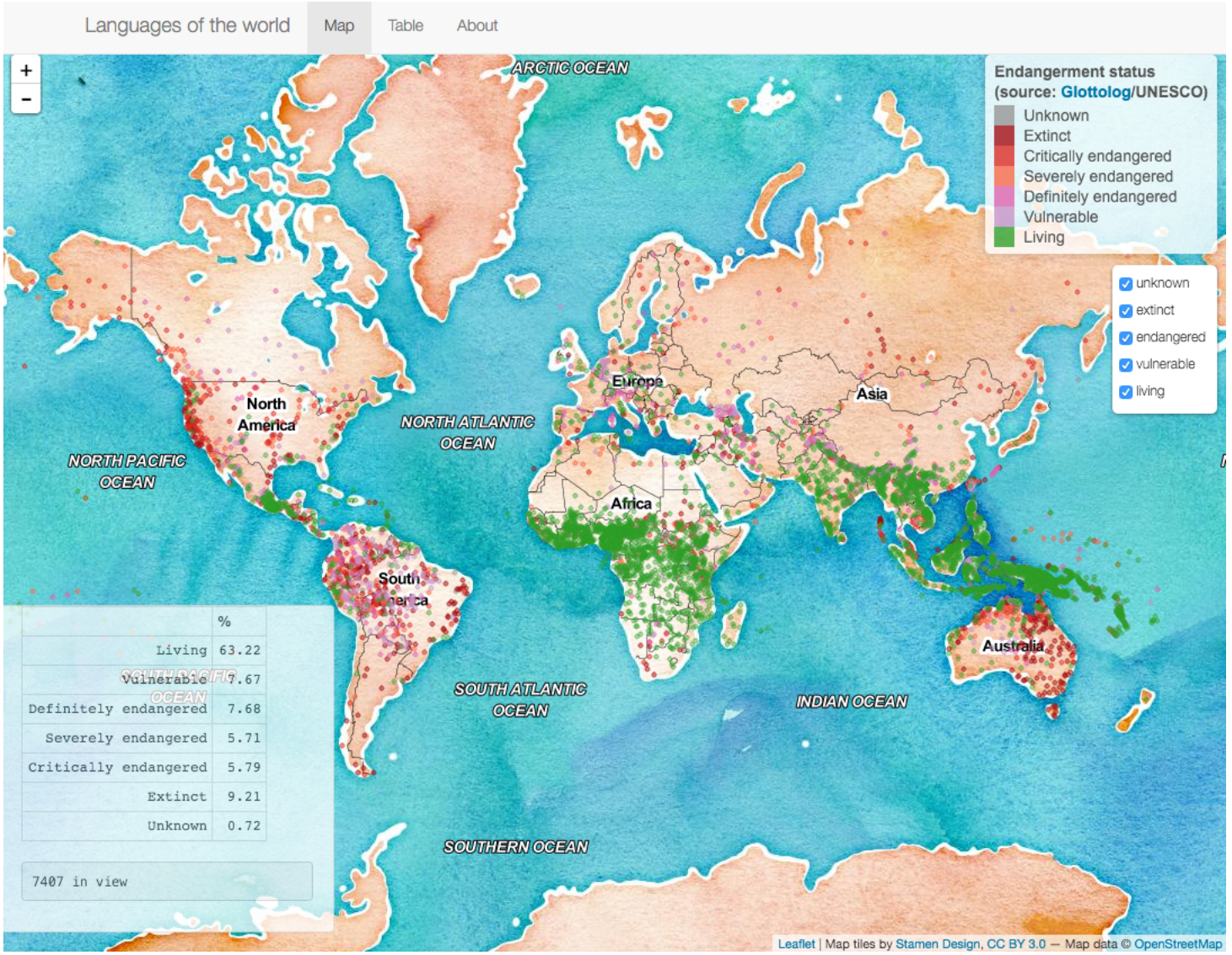

Figure 2: The GDE MAP tab, whole world view.

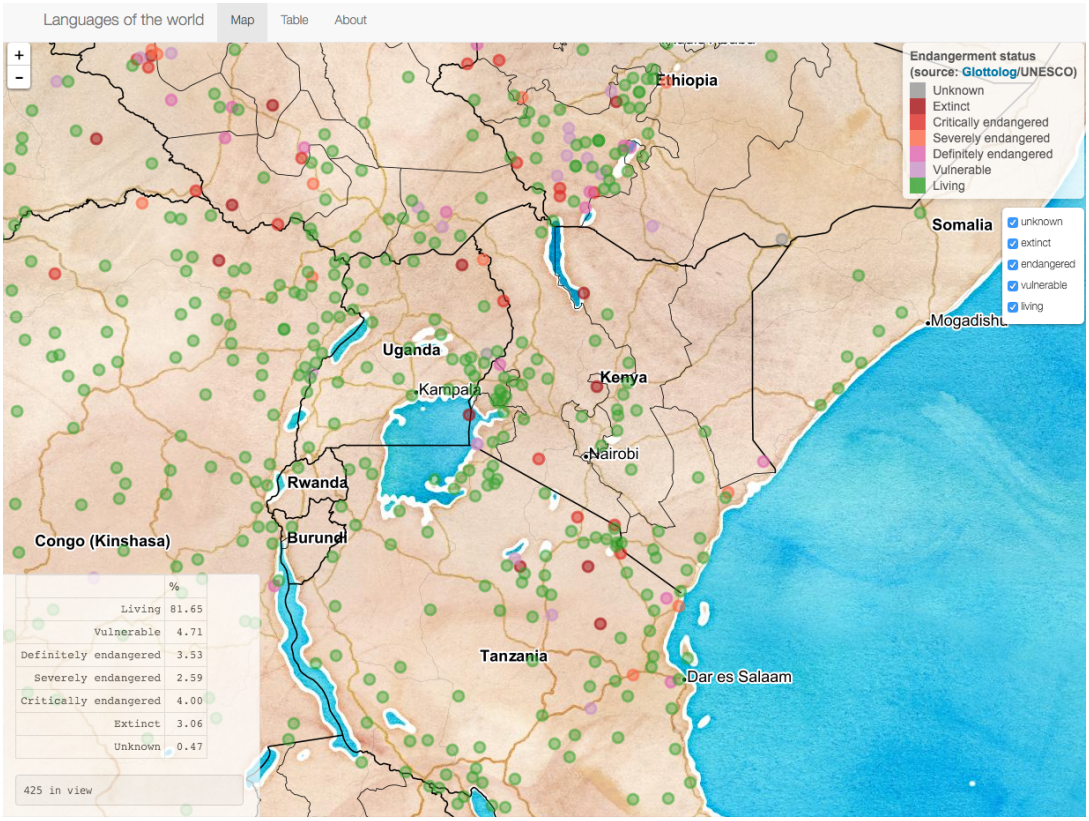

Figure 3: The GDE MAP tab, zoomed into East Africa; *n.b.* adjusted table counts, bottom-left.

endangered and vulnerable languoids, and green for 'living' (Figure 2). We provide a legend to this effect, along with a control panel to exclude (and include) languoids according to their status (top-right of map).

We chose to layer the languoids from extinct to living to draw attention to how many living languages there are and where there are noticeably high quantities (*e.g.* West Africa and Papua New Guinea above all). To put the extinct languoids on top did not make sense to us: this would be arresting but in some sense futile. It's already too late for these languages, and what's more, many of them are distant to the present day – ancient languages such as Latin and Egyptian. However, many are not, and we have made the living languoid points less opaque than the other languoids, so that the reds and purples of extinct and endangered languages lurk ominously in the background.

On the left of the window there are zoom controls, along with a count of the number of languoids currently in view, which at the most is 7407 (Figure 2) but which, if we zoom in on the east African region for example, reduces to 293 (Figure 3). Note that the table of endangerment proportions also adjusts to the current view (*cf.* Figure 2, Figure 3), and that the table may be 'picked up' and dragged to another point on the map if so desired. Tooltips mean that if the user selects a languoid datapoint, a textbox pops up on screen with more information about that languoid: its higher level family classifications, its endangerment status, speaker count, and a link to its Glottolog webpage (*e.g.* Figure 4).

The user may also opt to display the languoid markers sized by their speaker counts from Ethnologue (Lewis et al., 2015). We use a logarithmic scale for this visualization function (Figure 5). Speaker counts were collected by the second author (Bentz, 2016); note that we make them visible as part of the tooltip but they are not available to download in the TABLE. We would need an Ethnologue licence (see also section 5) to redistribute these data. However, we envisage that it would be most valuable for other researchers to test hypotheses relating to population size and language features.

With 7407 data points at most, overplotting is evidently a danger in the GDE, and something that's hard to avoid when there are so many dense geographic clusters. Apart from the partial transparency mentioned above, we resize the data points so that they are smallest at the outermost map zoom, increase as the user zooms in, and vice versa. As the points have to be redrawn as the user moves between zoom levels this leads to a slowdown in performance. With funding, we could upgrade our Shiny Apps subscription to a paid one with its accompanying performance boost involving increased memory and multi-threading.

The TABLE tab allows the user to view the data table underlying the map. Aside from browsing page by page, there are several table-filtering methods available (highlighted in the figures with red boxes; these boxes do not appear in GDE): a search textbox, endangerment status tick-boxes, and language family selection (Figure 6). The user may also download the table rows currently in view in one of several formats, thanks to the DT package which provides an R interface to the JavaScript library DataTables (Xie,

Figure 4: The GDE MAP tab, tooltips example.

2015).

Finally, the ABOUT tab provides information about the GDE web application: the motivation for creating it, a brief commentary about present-day language endangerment, credits and acknowledgements as to where the data come from and the tools used to make the application (Section 6).

4 Language endangerment

Estimates vary as to the severity of the prognosis at this point: UNESCO refers to a 'widely accepted' endangerment ratio of 50% (Moseley, 2010), whilst one heavily cited study predicts that "the coming century will see either the death or doom of 90% of mankind's languages" (Krauss, 1992), and the unattributed claim that a language dies every two weeks is in wide circulation (see the recent *Ethnos Project* blogpost by Mark Oppenneer for examples[7]). Others paint a less alarmist (though still alarming) picture (*e.g.* Campbell et al. (2013)). In any case, in the words of the linguist Lyle Campbell, the present predicament and its immediate consequences are "tragic, with its irreparable damage and loss" (Campbell et al., 2013).

As stated at the outset, language endangerment was our primary focus in creating GDE. For the 7407 geo-located languoids we extracted from the Glottolog catalogue we find that only 63% were classified by UNESCO as 'safe' (or 'living', as Glottolog relabelled it; Table 2).

Endangerment	Count	%
Living	4683	63.22
Vulnerable	568	7.67
Definitely endangered	569	7.68
Severely endangered	423	5.71
Critically endangered	429	5.79
Extinct	682	9.21
Unknown	53	0.72
Total	*7407*	*100*

Table 2: Endangerment status statistics for the languoids extracted from Glottolog

If one excludes already extinct and status-unknown languoids from the proportional calculation, then the safe percentage rises to 70% $(4683/(7407-(682+53))$.

If one assumes that the path from 'safe' to 'extinct' is monotonic and inevitable, then we are faced with the loss of three-in-ten existing languages, a prospect less alarming than the oft-repeated 90% figure that comes from Krauss Krauss (1992). Nonetheless it would be a horrendous loss of cultural heritage and diversity, especially if one considers regional endangerment. For example, by zooming the

[7] http://www.ethnosproject.org/status-of-the-ethnosphere

Figure 5: The GDE MAP tab, datapoints sized by speaker counts with, for instance, Italian the green circle at centre (55 million speakers), Judeo-Italian the small green circle to its left (200 speakers), and Sicilian the purple circle at bottom (4.7 million speakers).

Languages of the world | Map | Table | About

Source: Glottolog

Endangerment status
- [ ] Living
- [ ] Vulnerable
- [x] Definitely endangered
- [x] Severely endangered
- [x] Critically endangered
- [ ] Extinct
- [ ] Unknown

Language family: level 1: Indo-European

Language family: level 2: Germanic

Language family: level 3: Northwest Germanic

Show 20 entries

Search:

name	status	family1	family2	family3
Berbice Creole Dutch	Critically endangered	Indo-European	Germanic	Northwest Germanic
Cimbrian	Definitely endangered	Indo-European	Germanic	Northwest Germanic
Eastern Yiddish	Definitely endangered	Indo-European	Germanic	Northwest Germanic
Jutish	Definitely endangered	Indo-European	Germanic	Northwest Germanic
Mócheno	Definitely endangered	Indo-European	Germanic	Northwest Germanic
Northern Frisian	Severely endangered	Indo-European	Germanic	Northwest Germanic
Petjo	Critically endangered	Indo-European	Germanic	Northwest Germanic
Pitcairn-Norfolk	Definitely endangered	Indo-European	Germanic	Northwest Germanic
Plautdietsch	Definitely endangered	Indo-European	Germanic	Northwest Germanic
Saterfriesisch	Severely endangered	Indo-European	Germanic	Northwest Germanic
Swedish	Definitely endangered	Indo-European	Germanic	Northwest Germanic

Showing 1 to 11 of 11 entries

Previous 1 Next

Figure 6: The GDE TABLE tab, endangered Northwest Germanic languoids, 20 per page.

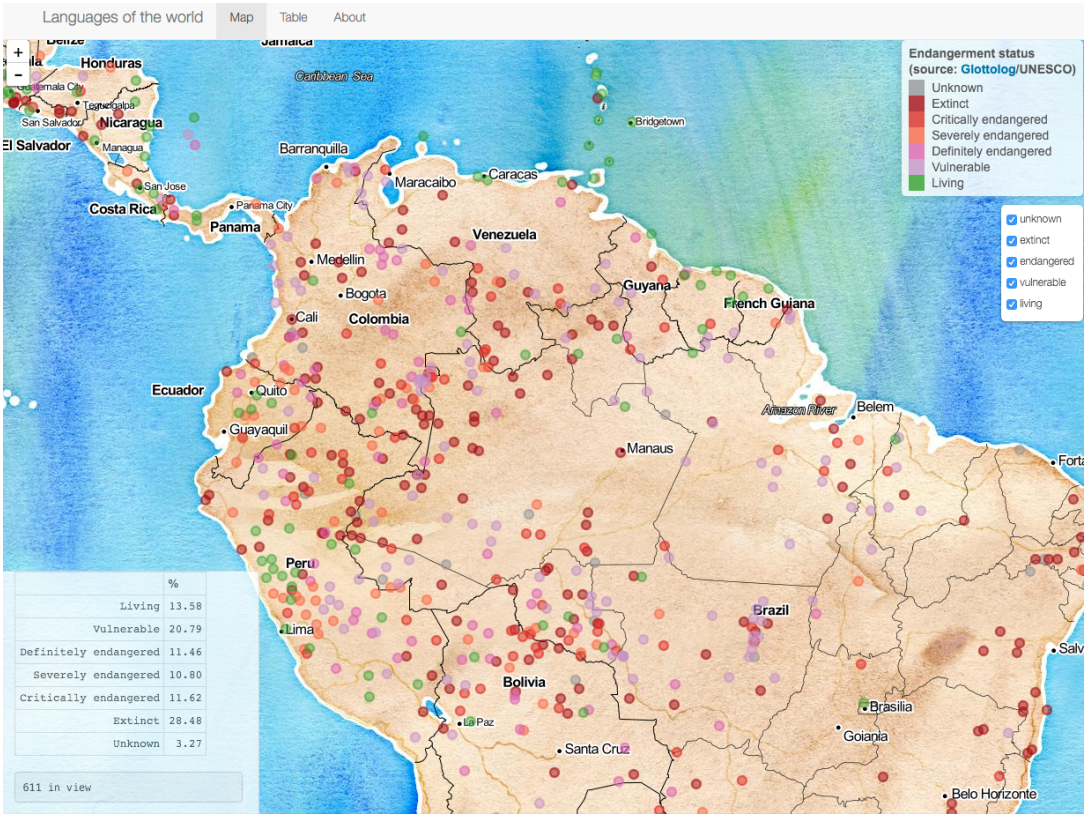

Figure 7: The GDE MAP tab, the northern region of South America including the Amazon rainforest.

GDE map in to the approximate region of the Amazon rainforest and its surroundings in northern South America, and by considering extant languoids only, we see that the 'safe' proportion falls to just 13.5% (Figure 7). More precise regional analyses may be performed in future once we associate languoid geo-coordinates with information by country and region (section 5).

How do the endangerment counts presented here, that come from UNESCO via Glottolog, compare to others? Ethnologue's most recent report states that 63% (4719/7480) of the languages identified in 1950 "are still being passed on to the next generation in a sustainable way", whilst 32% "are currently at some stage in the process of language loss", and 377 (5%) "have been identified as having lost all living speakers and ceasing to serve as a language of identity for an ethnic community in the last six decades" (Simons and Lewis, 2013).

Meanwhile the ENDANGERED LANGUAGES CATALOGUE project (ELCat) puts the endangerment statistic at 43% of 7102 existing languages, and states that 457 (9.2%) have "fewer than ten speakers and are very likely to die out soon, if no revitalization efforts are made", while 634 of all known languages have already become extinct, 141 of which in the last forty years (Campbell et al., 2013).

More worrying yet is ELCat's observation regarding language families, an issue not represented in the GDE[8]:

> We know of a hundred language families that have gone extinct over the course of history – 24% of the world's linguistic diversity. But the fact that 28 of them have gone extinct over the relatively short time span of the last 50 years is symptomatic of the accelerated rate of language loss we are experiencing in recent times.

Analogies with biodiversity loss are clear: indeed, direct parallels between language and species extinction have been drawn, with both linked to economic development (Amano et al., 2014). ELCat is an ongoing project that continues to focus on endangered languages and add to its catalogue with crowdsourced contributions. They have also produced a map visualization of language endangerment which the reader may find at `http://www.endangeredlanguages.com`, though note that it only maps endangered languages rather than all languages as in the GDE.

The above, somewhat bleak, assessments of course may be improved by government or community efforts toward language revitalization. This is no easy endeavour, but we point to notable successful efforts in recent times, such as Hebrew and Scottish Gaelic (McEwan-Fujita, 2011; Kaufman, 2005). We also draw attention to the efforts in computational linguistics to spread the kind of natural language processing technology that is so prevalent and which has so benefitted major languages (above all English) to 'low resource' languages – those languages for which the tools and databases that underpin, for instance, web search, grammar checkers and teaching apps do not yet exist. Notable ex-

[8] Source: `http://rosettaproject.org/blog/02013/mar/28/new-estimates-on-rate-of-language-loss`; accessed 2016-12-11.

amples include the HUMAN LANGUAGE PROJECT (Abney and Bird, 2010; Emerson et al., 2014), LOWLANDS (*e.g.* Agic et al. (2015)) and RU_CALL – computer-assisted language learning for the revitalization of a Ugandan language, Runyakitara (Katushemererwe and Nerbonne, 2015).

5 Future development

The GDE currently visualizes endangerment status for each languoid, indicated by datapoint colours on the map. We envisage further development to either offer endangerment data source options to the user – for instance offering a choice of UNESCO, Ethnologue, or ELCat classification – and/or to offer alternative visualizations with the Glottolog languoid set: *e.g.* number of L2 speakers, altitude and complexity measures Bentz (2016). Another improvement would be to offer the speaker counts for download via the TABLE. For this we would need funding to purchase an Ethnologue licence (Lewis et al., 2015) that allows for redistribution of these data.

As for the GDE itself, we intend to add further functionality: first, the option to minimize the data table in the MAP tab; second, the facility to download the filtered data (or, all of it) in the TABLE tab; thirdly, visualization of languoids grouped by region, country and family. We welcome further suggestions toward the improvement of GDE usability, data presentation, and content.

6 Acknowledgements

We thank Harald Hammarström of the Max Planck Institute for Psycholinguistics, and Robert Forkel of the Max Planck Institute for the Science of Human History for their help in accessing and explaining the Glottolog database. We are grateful for the many helpful comments and questions from two anonymous reviewers. We also thank Dr Anne Alexander and participants for their feedback at the 'Graphical Display: Challenges for Humanists' workshop organised by the Cambridge Digital Humanities Network in 2015, and Jane Walsh for her ongoing support of linguistics research at Cambridge as coordinator of the Language Sciences Strategic Research Initiative. The first and last authors are funded by Cambridge English Language Assessment as part of the Automated Language Teaching & Assessment Institute. The third author is supported by the Onassis Foundation. The fourth author is in receipt of a CAPREx award from the Cambridge-Africa Programme.

7 Bibliographical References

Abney, S. and Bird, S. (2010). The Human Language Project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Agic, Z., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Amano, T., Sandel, B., Eager, H., Bulteau, E., Svenning, J.-C., Dalsgaard, B., Rahbek, C., Davies, R. G., and Sutherland, W. J. (2014). Global distribution and drivers of language extinction risk. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1793).

Bentz, C. (2016). *Adaptive Languages: An Information-Theoretic Account of Linguistic Diversity*. Phd thesis, University of Cambridge.

Campbell, L., Lee, N. H., Okura, E., Simpson, S., and Ueki, K. (2013). New Knowledge: Findings from the Catalogue of Endangered Languages (ELCat). In *3rd International Conference on Language Documentation & Conservation*.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J., (2016). *Shiny: web application framework for R*. R package version 0.13.0.

Cheng, J. and Xie, Y., (2015). *Leaflet: create interactive web maps with the JavaScript 'Leaflet' library*. R package version 1.0.0.

Matthew S. Dryer et al., editors. (2013). *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Emerson, G., Tan, L., Fertmann, S., Palmer, A., and Regneri, M. (2014). SeedLing: Building and using a seed corpus for the Human Language Project. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S., (2016). *Glottolog 2.7*. Jena: Max Planck Institute for the Science of Human History.

Katushemererwe, F. and Nerbonne, J. (2015). Computer-Assisted Language Learning in support of (re-)learning native languages: The case of Runyakitara. *Computer Assisted Language Learning*, 28(2):112–129.

Kaufman, D. (2005). Acquisition, Attrition, and Revitalization of Hebrew in Immigrant Children. In Dorit Diskin Ravid et al., editors, *Perspectives on Language and Language Development: Essays in Honor of Ruth A. Berman*, pages 407–418. Springer, Boston.

Krauss, M. (1992). The world's languages in crisis. *Language*, 68(1):4–10.

M. Paul Lewis, et al., editors. (2015). *Ethnologue: Languages of the World*. SIL International, Dallas, 18th edition. Online version: http://www.ethnologue.com.

McEwan-Fujita, E. (2011). Language revitalization discourses as metaculture: Gaelic in Scotland from the 18th to 20th centuries. *Language and Communication*, 31(1):48 – 62.

Christopher Moseley, editor. (2010). *Atlas of the World's Languages in Danger, 3rd edn*. UNESCO Publishing, Paris.

R Core Team, (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Simons, G. F. and Lewis, M. P. (2013). The world's languages in crisis: a 20-year update. In Elena Mihas, et al., editors, *Responses to language endangerment. In honor of Mickey Noonan*. John Benjamins, Amsterdam.

The LINGUIST List. (2014). Multitree: A digital library of language relationships.

Xie, Y., (2015). *DT: a wrapper of the JavaScript library 'DataTables'*. R package version 0.1.48.