LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 12 (2018), pp. 359–392
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24792
Revised Version Received: 5 July 2018

Simultaneous Visualization of Language Endangerment and Language Description

Harald Hammarström
Uppsala University

Thom Castermans
TU Eindhoven

Robert Forkel
Max Planck Institute for the Science of Human History

Kevin Verbeek
TU Eindhoven

Michel A. Westenberg
TU Eindhoven

Bettina Speckmann
TU Eindhoven

The world harbors a diversity of some 6,500 mutually unintelligible languages. As has been increasingly observed by linguists, many minority languages are becoming endangered and will be lost forever if not documented. Urgently indeed, many efforts are being launched to document and describe languages. This undertaking naturally has the priority toward the most endangered and least described languages. For the first time, we combine world-wide databases on language description (Glottolog) and language endangerment (ElCat, Ethnologue, UNESCO) and provide two online interfaces, GlottoScope and GlottoVis, to visualize these together. The interfaces are capable of browsing, filtering, zooming, basic statistics, and different ways of combining the two measures on a world map background. GlottoVis provides advanced techniques for combining cluttered dots on a map. With the tools and databases described we seek to increase the overall knowledge of the actual state language endangerment and description worldwide.

Licensed under Creative Commons
Attribution-NonCommercial 4.0 International

E-ISSN 1934-5275

1. Introduction[1] There are approximately 6,500 mutually unintelligible languages spoken in the world at present (Hammarström 2015:733). The diversity of these languages is an abundant resource for understanding the unique communication system of our species and for tracing the history of the populations that speak them (Evans & Levinson 2009). As has been increasingly observed by linguists (Wurm 1956; 1991; Zaborski 1970; Capell 1962; Becker-Donner 1962; Stone 1962; Kibrik 1991; Adelaar 1991), and especially since the seminal article by Krauss (1992), many minority languages are becoming endangered and will be lost forever if not documented. There is now a range of books describing these processes and their consequences in detail (Evans 2009; Harrison 2007; Thomason 2015; Grenoble & Whaley 1998; Abley 2003; Dalby 2003; Crystal 2000; Nettle & Romaine 2000). As for the actual inventory and vitality status of languages currently endangered, a number of surveys of are available, such as the Ethnologue (Simons & Fennig 2017), the UNESCO Atlas (Moseley 2010), and the *Catalogue of Endangered Languages* (ELCat) (http://www.endangeredlanguages.com). As several authors have stressed (e.g., Swadesh 1960; Krauss 2007; Sands 2017; Campbell & Rehg 2018) it is now a "race against time" to document them before it is too late.

Such an endeavor naturally involves prioritization, in that the most urgent are the most endangered and least described. A ranking in terms of endangerment and description may inform strategies in documentation programs such as the Endangered Languages Documentation Programme (http://www.eldp.net) or the NSF Documenting Endangered Languages program (https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12816). We say "inform" rather than "dictate" because the general question is more complex. Any individual or joint documentation effort may achieve their objectives in the absence of such a ranking, or may wish to incorporate additional factors into it, e.g., language family (Hammarström 2010), region, access, or certainty, while nevertheless being informed by such a ranking (Chelliah & Reuse 2011:79–92; Hauk & Heaton 2018). In addition, language documentation itself is multifaceted. In the present paper we are able to survey language description in the sense of Himmelmann (1998; 2012) – more specifically, grammatical description – and measure it by academic publications. This is quite an idealization. There are stakeholders in language documentation who are not primarily interested in grammar books (cf. Rhodes & Campbell 2018; Himmelmann 2006), and who are best served by adding other aspects and sources of information to that which we provide here.

Given that the number of languages is vast and the activity of language documentation and description is a decentralized task (Hammarström & Nordhoff 2011; Chelliah & Reuse 2011:33–78), until now there has been no global database or publication that allows the user to look at the combination of these two features.[2] In the present paper, we describe two web-based applications that provide the functionality to explore language endangerment and description in a variety of ways. The

[1]A preliminary version of some material on GlottoVis in this paper has appeared in Castermans et al. (2017).

[2]Though there exist several web applications that allow browsing of either language endangerment data, e.g., the UNESCO Atlas of the World's Languages in Danger (http://www.unesco.org/languages-atlas/), or descriptive data, e.g., the Glottolog Data Explorer (Caines et al. 2016), but not both.

interfaces are built on two (replaceable) databases: Glottolog 3.1 (Hammarström et al. 2017) for the descriptive status of the languages of the world, and a combination of the three existing databases – ELCat, UNESCO and Ethnologue (see below) – for endangerment statuses. There are two versions of the web interface:

1. **GlottoScope:**
   A classic interface where each language is represented by a dot. This interface can be accessed at http://glottolog.org/langdoc/status.

2. **GlottoVis:**
   An innovative interface where languages are dynamically aggregated into sunburst charts to avoid clutter. This interface can be accessed at http://glammap.win.tue.nl/glottovis.

Both interfaces allow filtering, zooming, browsing, basic statistics, and links back to the original data sources. Tasks realizable by the interfaces include:

- determine the endangerment and descriptive status of a specific language;
- find languages with a specific descriptive or endangerment status;
- find the languages with a certain endangerment or descriptive status closest to specific geographic locations;
- see the geographical distribution of endangerment status;
- see the geographical distribution of descriptive status;
- see the geographical distribution of endangerment and descriptive status;
- see statistical distribution of endangerment status;
- see statistical distribution of descriptive status;
- see statistical distribution of endangerment and descriptive status;
- identify a region with many endangered/undescribed languages; and
- compare different regions in terms of endangerment/description.

The underlying data sources and management are described in §2, and the two web interfaces are described in §3.1 (GlottoScope) and §3.2 (GlottoVis).

2. Data

2.1 The language inventory As the language inventory we use Glottolog 3.1 (Hammarström et al. 2017), which is currently the most comprehensive inventory of extinct and living languages. This inventory is very similar to that which one obtains by merging the 20th edition of the Ethnologue (Simons & Fennig 2017) with the extinct languages of the ISO-639-3 inventory (see http://www-01.sil.org/iso639-3/). The (diminishing number of) differences between Glottolog 3.1 and the ISO-639-3 inventory are stated in each corresponding entry in Glottolog along with a reason (see also Hammarström 2015:Appendix).

We restrict our attention to the attested classifiable (spoken or signed) L1 languages (see http://glottolog.org/glottolog/glottologinformation for an explanation of these criteria) since this set of languages is what typical users expect. This amounts to 7568 languages.[3] The coordinates for languages are also taken from Glottolog 3.1 (Hammarström et al. 2017).

2.2 Data on language description For the purpose of GlottoScope and GlottoVis, each language is given a single *description level*. The description level is calculated as follows. We start from the bibliographic reference collection of Glottolog 3.1 (Hammarström et al. 2017). A sufficient subset of these bibliographic references are annotated as to language and type of description. Considering all the references for a given language, its description level is the maximally extensive type of its references. These steps are described in more detail below.

Glottolog 3.1 collects 320,181 bibliographic references for the languages of the world, and as such constitutes the base for assessing the description level of each language. References are annotated as to the target language (the language being described) and description type (grammar, dictionary, etc.) – see Tables 1 and 2. A subset of the Glottolog collection has been annotated manually by experts and thus has highly accurate annotations, whereas the remaining references have been annotated automatically (see Hammarström 2011), which is reliable only on average. Since high precision is required for the purpose of GlottoScope and GlottoVis, we use the subset of references which has been annotated manually. This subset, which amounts to over 44,000 references, is nevertheless complete in that it includes the most extensive description (see below) for every language (Hammarström & Nordhoff 2011:32–34). This subset does not count ongoing work (unless published) and does not count unpublished manuscript work (unless available on the web or in a publicly accessible archive). It does, however, include the important class of masters' and doctoral theses, since these are in principle available at the host institution.

Some examples of references and their description types are given in Table 2. A few comments are in order. A single publication may realize more than one category, e.g., the dictionary of San Felipe Usila Chinantec is a dictionary but also contains an extensive grammar sketch (Skinner & Skinner 2000:469–587). The human judgment

[3]If one were to go strictly by mutual intelligibility, the number would, with high probability, be close to 6500 (Hammarström 2015:733).

of the actual contents – not the number of pages – governs the assignment of category. One difficulty is in drawing the line between the different categories consistently, so it is likely that the database contains some inconsistencies. Although having a long grammar is the top category, this does not imply that a language is fully described once in that category. In fact, we know of no language which is so extensively described that it could be called "fully described".

Table 1. Detailed list of grammatical descriptions in Glottolog.

Score	Most extensive grammatical description type		# languages	
5	long grammar	extensive description of most elements of the grammar ≈300+ pages	1,625	21.5%
4	grammar	a description of most elements of the grammar (≈150 pages)	898	11.8%
3	grammar sketch	a less extensive description of many elements of the grammar (≈50 pages)	1,973	26.0%
2	specific feature	description of some element of grammar (i.e., noun class system, verb morphology, etc.)	395	5.2%
2	phonology	a description of the sound inventory utilizing minimal pairs	275	3.6%
2	dictionary	≈75 pages and beyond	158	2.0%
2	text	text material	86	1.1%
1	wordlist	≈100–200 words	1,580	20.8%
0	minimal	a small number of morphemes	441	5.8%
0	overview	document with meta-information about the language (i.e., where spoken, non-intelligibility to other languages, etc.)	137	1.8%
			7,568	

Table 2. Examples of description types and associated languages in Glottolog.

Language	Description Type	Bibliographic Reference
Tauya [tya]	long grammar	MacDonald, Lorna. (1990) *A Grammar of Tauya* (Mouton Grammar Library 6). Berlin: Mouton de Gruyter. xiii+385pp.

Continued from previous page

Bolon [bof]	grammar	Zoungrana, Ambroise. (1987) *Esquisse phonologique et grammaticale du Bolon (Burkina-Faso) – contribution à la dialectologie mandé.* Université de la Sorbonne Nouvelle (Paris 3) doctoral dissertation. 336pp.
Usila Chinantec [cuc]	grammar sketch, dictionary	Skinner, Leonard E. & Marlene B. Skinner. (2000) *Diccionario Chinanteco de San Felipe Usila, Oaxaca* (Serie de vocabularios y diccionarios indígenas Mariano Silva y Aceves 43). Coyoacán, México: Instituto Lingüístico de Verano. xxix+602pp.
Norwegian Sign Language [nsl]	specific feature	Slowikowska Schröder, Bogumila. (2010) *Imperativ i norsk tegnspråk – en eksplorerende studie av et fenomen innen et visuelt-gestuelt språk [Imperative in Norwegian sign language an explorating study of a phenomenon in a visual-gestural language].* University of Oslo MA thesis. 119pp.
Sobei [sob]	phonology	Sterner, Joyce K. (1975) Sobei phonology. Oceanic Linguistics 14. 146–167.
Northern Tujia [tji]	dictionary	Zhang, Weiquan. (2006) Hàn yǔ tǔjiā yǔ cídiǎn 汉语土家语词典 [Chinese-Tujia dictionary]. Guiyang Shi: Guizhou Minzu Chubanshe. 6+20+3+436pp.
Nisga'a [ncg]	text	Boas, Franz. (1902) *Tsimshian Texts* (Bulletin of American Ethnology 27). Washington: Government Printing Office. 254pp.
Asháninka [cni], Yine [pib], Shipibo-Conibo [shp]	wordlist	Carrasco, Francisco. (1901) *Principales palabras del idioma de las tribus de infieles antis, piros, conibos, sipibos.* Boletín de la Sociedad Geográfica de Lima 11. 204–211.
Dizin [mdx]	minimal	Conti Rossini, Carlo. (1937) *Il Popolo dei Magi nell'Etiopia Meridionale e il suo linguaggio.* In V Sezione: Etnografica-Filologica-Sociologica (Atti del Terzo congresso di Studi Coloniali VI), 108–118. Firenze: Centro di Studi Coloniali, Instituto Coloniale Fascista.

Continued from previous page

Busuu [bju], Bishuo [bwh], Bikya [byb] Kutep [kub], Yukuben [ybl], Akum [aku], Beezen [bnz], Naki [mff]	overview	Breton, Roland. (1995) *Les Furu et leurs Voisins: DÉcouverte et essai de classification d'un groupe de langues en voie d'extinction au Cameroun.* Cahiers des Sciences Humaines 31(1). 17–48.

The reference database gives a series of bibliographical items for each language. The *most extensive description* (MED) is the longest description according to the hierarchy in Table 1. For example, if there is a wordlist, grammar sketch, dictionary, and grammar for a language, its MED will be (the longest) grammar. The number of items in each category does not make a difference, i.e., there may be one, two, five or a hundred grammar sketches for a language, but as long as there is no grammar, the MED type will still be grammar sketch.[4]

It must be noted that the hierarchy in question centers on *grammatical* description, rather than lexical or textual documentation (cf. Woodbury 2011), in that grammatical description is ranked higher than lexicon and text regardless of the size. A more ambitious project than the current one could target more dimensions of documentation than grammatical description. Unfortunately, this is not possible with the high-precision subset of the Glottolog references, since it is not known to be complete and uniformly annotated with respect to, for example, lexicon and text. In addition to academically published materials, there is also a legacy of audio and audiovisual materials constituting language documentation. Modern technology (see, for example, Bird et al. 2014; Bird 2010) promises to produce such data in much larger volumes in the future. While it is regrettable that we cannot cover lexical, textual, and audiovisual documentation in the present project, all is not lost. Until recently, grammatical description and language documentation was mostly done in a dialectic manner (cf. Rhodes & Campbell 2018; Himmelmann 2012), so the existence of one can be expected to be strongly correlated with the other. Hopefully, future projects and databases may serve to extend the coverage, especially targeting the modern outlet of non-academic materials.

While the description level as calculated above gives an approximate idea of the state of grammatical description of a language, the discretization into categories introduces hard boundaries which are not there "in nature". Moreover, it is oblivious to a number of more or less relevant aspects. For example, no quality assessments

[4]There are a few cases where separate descriptions of specific features, e.g., phonology, noun class system, verb phrase, clause combining, etc. "add up" to a grammar sketch or even a grammar, yet no single publication encompasses all of them. These cases are manually patched so that the description level of the language as a whole is rendered as grammar sketch or grammar.

requiring specialized knowledge are executed, and no differentiation is made between old and modern grammars or the theoretical framework used for description. Similarly, no record is made of which grammars have morpheme-by-morpheme glossing and which do not, even though this is known to greatly enhance reader accessibility.

2.3 Data on language endangerment Three global data sets on language endangerment are available:

2.3.1 The UNESCO Atlas of the World's Languages in Danger It builds on a series of earlier surveys by area, written by experts and published in book form (van Driem 2007b; Dimmendaal & Voeltz 2007; Crevels 2007; Wurm 2007a; Salminen 2007b; Golla 2007; Adelaar 2007a; Adelaar 2007b; Bradley 2007a; Yamamoto 2007; Blench 2007; Brenzinger 2007a; Brenzinger 2007b; Grinevald 2007; Connell 2007; Tryon 2007; Bradley 2007b; Owens 2007; Moore 2007; Evans 2007; Kazakevich & Kibrik 2007; Wurm 2007b; Salminen 2007a; Adelaar 2007a; Adelaar 2007b; van Driem 2007a).[5] Endangerment is assessed as per the UNESCO Language Vitality and Endangerment framework (Brenzinger et al. 2003) reproduced in Table 3. The version of the data set used for the present paper was downloaded on July 27, 2017 and contains 2,724 entries. 2,082 (76.4%) of the entries have a source, though, in many cases, the source indicated is an overview which does not give individual sources for its data points, so the information can often not be traced down to the underlying observation, even when there is an explicit source. An entry in the UNESCO Atlas of the World's Languages in Danger may either undermatch, i.e., correspond to a dialect, or overmatch, i.e., correspond to more than one language in the inventory used for the present paper. All in all, the entries in the UNESCO Atlas correspond to 2,414 spoken or signed L1 languages in the present inventory.[6] (Whenever multiple conflicting endangerment assessments are attached in the UNESCO Atlas to what is considered a single language in the present inventory, the endangerment status of that language is counted as per its least endangered dialect.)

2.3.2 Ethnologue, 20th edition It builds on the earlier editions which, since the 17th edition (Lewis et al. 2013), have included information on language endangerment using the Expanded Graded Intergenerational Disruption Scale (EGIDS) scale (Lewis & Simons 2010:Table 5). Except for a large class of extinct languages, the 20th edition of the Ethnologue (Simons & Fennig 2017) provides a vitality status for nearly all languages of the present inventory. For comparability, the extinct languages have been added according to Hammarström et al. (2017), making a total of 7,434 spoken or signed L1 language (non-)endangerment assessments. Of the 7,434 assessments, 4,354 are counted as not endangered and 3,080 as endangered. Ethnologue does not provide a source for each individual status judgment. The omission of

[5]Available at http://www.unesco.org/languages-atlas/.

[6]The matchings were updated, corrected, and expanded by the present authors starting from the ISO 639-3 assignment (if any) in the downloaded database.

Table 3. Categories of endangerment in the UNESCO Atlas of the World's Languages in Danger (Moseley 2010).

Endangerment status	Description
safe	language is spoken by all generations; intergenerational transmission is uninterrupted
vulnerable	most children speak the language, but it may be restricted to certain domains (e.g., home)
definitely endangered	children no longer learn the language as a "mother tongue" in the home
severely endangered	language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves
critically endangered	the youngest speakers are grandparents and older, and they speak the language partially and infrequently
extinct	there are no speakers left

individual sources is difficult to explain since, ultimately, all information comes from somewhere (Simons & Fennig 2017).

2.3.3 The Catalogue of Endangered Languages (ELCat) Judging from the sources cited, ELCat draws on any information available, including other databases mentioned in the present paper. Endangerment is assessed as per the Language Endangerment Index (Lee & Van Way 2018; Campbell 2017:4–5, Table 6). The version of the dataset used for the present paper was downloaded from http://endangeredlanguages.com/userquery/download/ on July 27, 2017 and contains 3,407 entries. Though not included in the downloadable dataset, individual sources are cited on the website for entries (Campbell & Belew 2018:9–10). Similarly to the UNESCO Atlas, in many cases, the source indicated is an overview which does not give individual sources for its datapoints, so the information can often not be traced down to the underlying observation. An entry in ELCat may either undermatch, i.e., correspond to a dialect, or overmatch, i.e., correspond to more than one language in the inventory used for the present paper. All in all, the entries in ELCat correspond to 3,197 spoken or signed L1 languages in the present inventory.[7] (Whenever multiple conflicting endangerment assessments are attached in ELCat to what is one and the same language in the present inventory, the endangerment status of that language is counted as its least endangered dialect.)

The three databases have design differences. In particular, three different scales of endangerment are used: the UNESCO, EGIDS, and LEI. Fortunately for comparison, the scales are sufficiently commensurate (the formula of ELCat poses some challenges; see Lee & Van Way 2018:68–72; Campbell 2017:4–6). A mapping between the scales and the translation into the Agglomerated Endangerment Status (see below) is given

[7]See footnote 6.

in Table 4. The mapping was elaborated by Frank Seifart (see Seifart et al. in press; Lewis & Simons 2010:110) based on the definitions in the respective scales.

Once the scales are made comparable, we can observe significant differences in the data delivered. There are a total of 3,870 languages considered endangered in *at least one* of the datasets. That is, 3,870 languages are not considered already extinct or not-endangered by all three datasets. The alluvial diagram in Figure 1 shows the differences across the three datasets. The three major discrepancies are the following:

1. Over 1,000 languages considered either "not endangered" (742) or "threatened" (473) in E20 are "shifting" in ELCat.
2. Over 400 languages considered "threatened" (416) in E20 are absent from ELCat.
3. Over 700 languages are absent from the smallest catalogue, UNESCO, but considered "threatened" (555) or "shifting" (148) in E20, or alternatively "threatened" (275) or "shifting" (595) in ELCat.

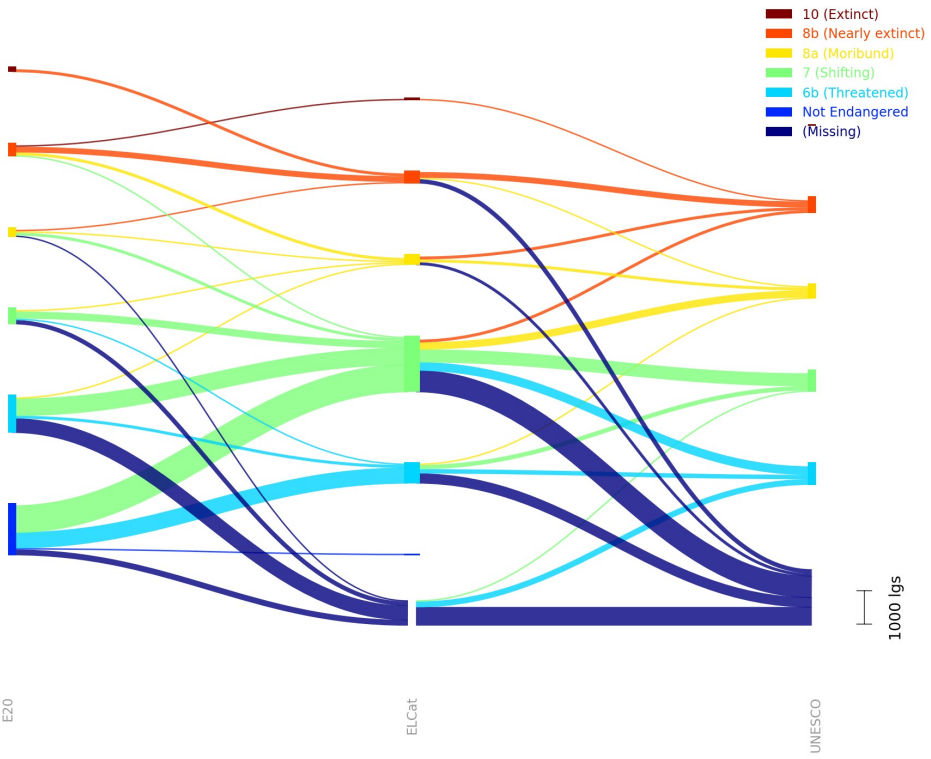


Figure 1. An alluvial diagram of the differences across the UNESCO, ELCat and E20 dataset.

Since individual sources are missing in E20, one cannot resolve these differences in a systematic manner. In the absence of a strictly scientific repertoire choice for GlottoScope/GlottoVis we have opted for the following pragmatic solution:

1. If the language is present in ELCat, the language endangerment status is taken from ELCat. This is justified on the grounds that ELCat has explicit sources for its datapoints. It also runs less risk of underestimating endangerment (with its potentially fatal consequences) since ELCat in general rates languages at a higher degree of endangerment than the alternatives.
2. If the language is not present in ELCat, the language endangerment status is taken from UNESCO, since sources are present there more often than in E20.
3. If the language is present neither in ELCat nor UNESCO, the language endangerment status is taken from E20.
4. If the language is present neither in ELCat, UNESCO, nor E20 it is taken from Glottolog, if present there.
5. Otherwise, its endangerment status is unknown.

In other words, the order of preference is ELCat > UNESCO > E20 > Glottolog > Unknown. The endangerment database computed in this way will be referred to as the Agglomerated Endangerment Status (AES). Table 4 summarizes the number of languages from each source and endangerment category. Agglomerated Endangerment Status (along with provenance and date of adoption) is included in Glottolog 3.1 and can be downloaded from there.[8]

There were other options to combine the existing information on language endangerment towards a global indicator. For example, one could take the majority view whenever the databases give conflicting information. If the voting datapoints are independent observations, such a strategy is very likely to yield a better approximation of the true state than any of the single datapoints. However, there is reason to believe than most datapoints spanning several databases are not independent (one being directly adopted from the other) and differences are frequently tied to different years, i.e., older vs. updated information. Another possibility would be to take the pessimistic bound, i.e., to choose the most endangered status in the case of multiple differing assessments. This would make sense if shifts in language endangerment are always unidirectional (towards more endangerment) and if endangerment assessments are always based on observation. Fortunately, shifts in endangerment status are not always for the worse, and unfortunately, many endangerment assessments are based on presumption rather than observation, i.e., ideas on what would be (un-) expected of a speaker population of a certain size in a certain region. For these reasons, we have chosen to be guided rather by the existence of a source so that the quality of the individual datapoints can be scrutinized and improved on in a systematic manner.

3. Visualization of language endangerment and description data We now turn to visualization of the description and endangerment data just described. The problem at hand resembles that of dynamic map labeling (Been et al. 2006). Two interfaces

[8]The endangerment information is found in the corresponding field in each languoid entry in the repository on https://github.com/clld/glottolog.

Table 4. Agglomerated Endangerment Status (AES) sources and endangerment statistics.

Priority	Source	# languages	AES	# languages
1	ELCat	3,197	7 (Shifting)	1,566
			6b (Threatened)	613
			8b (Nearly extinct)	384
			8a (Moribund)	335
			10 (Extinct)	208
			Not endangered	91
2	UNESCO	408	6b (Threatened)	176
			7 (Shifting)	77
			8a (Moribund)	27
			8b (Nearly extinct)	26
			10 (Extinct)	102
3	E20	3,435	Not endangered	2,849
			6b (Threatened)	349
			7 (Shifting)	83
			8a (Moribund)	28
			8b (Nearly extinct)	14
			10 (Extinct)	112
4	Glottolog	403	(Not extinct)	58
			10 (Extinct)	345
5	None	125	–	125
		7,568		7,568

are described: GlottoScope and GlottoVis. GlottoScope is a straightforward visualization with one glyph on the map for each language. This is a natural visualization form but renders the map difficult for the eye to appreciate when there are too many glyphs close to each other, i.e., in language-dense regions. GlottoVis is a visualization which dynamically combines what would be too many glyphs in the same place to one composite larger glyph. The juxtaposition of sample screenshots in Figure X gives a comparative view. We present two interfaces rather than one for various reasons. GlottoScope is straightforward and is naturally understood by novice users, but carries the limitation that too many glyphs cannot be faithfully shown in too little space. Also, both description and endangerment status are ordinal but one of the two has to be indicated by shape rather than colour, despite the fact that shape is not ordinal (Bertin 1967; Tufte & Graves-Morris 1983). The solution in GlottoVis for both the density and shape problem is to aggregate nearby items into disjoint glyphs (Ward 2008). Similar techniques have been used before (see, for example, Scheepens et al. 2014, which includes a good overview), and related approaches have also been applied to network exploration (see, for example, Vehlow et al. 2013).

Table 5. The EGIDS scale of language endangerment (Lewis & Simons 2010).

Endangerment status	Description
0 International	The language is used internationally for a broad range of functions.
1 National	The language is used in education, work, mass media, and government at the nationwide level.
2 Regional	The language is used for local and regional mass media and governmental services.
3 Trade	The language is used for local and regional work by both insiders and outsiders.
4 Educational	Literacy in the language is being transmitted through a system of public education.
5 Written	The language is used orally by all generations and is effectively used in written form in parts of the community.
6a Vigorous	The language is used orally by all generations and is being learned by children as their first language.
6b Threatened	The language is used orally by all generations but only some of the child-bearing generation are transmitting it to their children.
7 Shifting	The child-bearing generation knows the language well enough to use it among themselves but none are transmitting it to their children.
8a Moribund	The only remaining active speakers of the language are members of the grandparent generation.
8b Nearly extinct	The only remaining speakers of the language are members of the grandparent generation or older who have little opportunity to use the language.
9 Dormant	The language serves as a reminder of heritage identity for an ethnic community. No one has more than symbolic proficiency.
10 Extinct	No one retains a sense of ethnic identity associated with the language, even for symbolic purposes.

3.1 GlottoScope GlottoScope (http://glottolog.org/langdoc/status) is built in as a component of Glottolog and provides a straightforward interface combining the Agglomerated Endangerment and Descriptive Status. Figure 3 shows a screenshot, showing the selection table (top right), map (mid), and statistics table (bottom).

3.1.1 GlottoScope Design and Functionality The map shows a glyph at the coordinate of each language on a world map background. The color of the glyph reflects the endangerment status and its shape reflects the descriptive status. The color scale is, iconically, from green to red, reflecting endangerment level, leaving black for extinct languages (Pravossoudovitch et al. 2014). The roles of the shape/color may be reversed, such that documentation is from red (wordlist or less) to green (grammar), by clicking the focus toggle button in the selection table (see Figure 4). The legend

Level of Endangerment	Intergenerational Transmission	Absolute Number of Speakers	Speaker Number Trends	Domains of use of the language
0 Safe	All members of the community, including children, speak the language.	> 100 000	Almost all community members or members of the ethnic group speak the language, and speaker numbers are stable or increasing.	Used in most domains, including official ones such as government, mass media, education, etc.
1 Vulnerable	Most adults and some children are speakers.	10 000–99 999	Most members of the community or ethnic group speak the language. Speaker numbers may be decreasing, but very slowly.	Used in most domains except for official ones such as government, mass media, education etc.
2 Threatened	Most adults in the community are speakers, but children generally are not.	1 000–9 999	A majority of community members speak the language. Speaker numbers are gradually decreasing.	Used in some nonofficial domains along with other languages, and remains the primary language used in the home for many community members.
3 Endangered	Some adults in the community are speakers, but the language is not spoken by children.	100–999	Only about half of community members speak the language. Speaker numbers are decreasing steadily, but not at an accelerated pace.	Used mainly just in the home and/or with family, but remains the primary language of these domains for many community members.
4 Severely Endangered	Many of the grandparent generation speak the language, but younger people generally do not.	10–99	Less than half of the community speaks the language, and speaker numbers are decreasing at an accelerated pace.	Used mainly just in the home and/or with family, and may not be the primary language even in these domains for many community members.
5 Critically Endangered	There are only a few elderly speakers.	1–9	A small percentage of the community speaks the language, and speaker numbers are decreasing very rapidly.	Used only in a few very specific domains, such as in ceremonies, songs, prayer, proverbs, or certain limited domestic activities.

Table 6. The Language Endangerment Index (LEI-ELCat) used by ELCat (Lee and Way 2018:68–72, Campbell 2017:4–6).

Table 7. Mappings between the endangerment categories in the source databases and the Agglomerated Endangerment Scale (AES).

UNESCO	LEI (ELCat)	EGIDS	AES
safe	at risk	1 (National) 2 (Regional) 3 (Trade) 4 (Educational) 5 (Written) 6a (Vigorous)	Not endangered
vulnerable	vulnerable	6b (Threatened)	6b (Threatened)
definitely end.	threatened endangered	7 (Shifting)	7 (Shifting)
severely end.	severely end.	8a (Moribund)	8a (Moribund)
critically end.	critically end.	8b (Nearly extinct)	8b (Nearly extinct)
extinct	dormant awakening	9 (Dormant) 9 (Reawakening) 9 (Second language only) 10 (Extinct)	10 (Extinct)

tab in the top left corner of the map allows (un-)selecting any of the categories for display.

The selection table allows the user to restrict the display to a certain macro-region (as defined in Hammarström & Donohue 2014) and/or to a certain set of language families (with the list of families from Glottolog). This feature can be of particular value, as zooming and panning the map may be slow with the full set of over 7,000 dots on some browsers and platforms. The selection table also allows for setting a year to show the state of description as of that year instead of the present. (The endangerment data is not indexed by year, while the description data naturally is indexable according to the year of publication.)

A mouse over and click on a map glyph shows the language in question and the associated information on description and endangerment status, along with the sources for these statuses and links back to Glottolog.

The statistics table counts the number of languages at each description and endangerment level according to the selection.

Comparison of different regions and yearly states can be achieved simply by opening several browser windows and juxtaposing them (see Figure 5).

All tasks identified in Section 1 can be achieved with GlottoScope. The only drawback is that in language-dense regions in wide zoom, too many glyphs have to compete for the same space, rendering the visualization insufficient to appreciate the actual state.

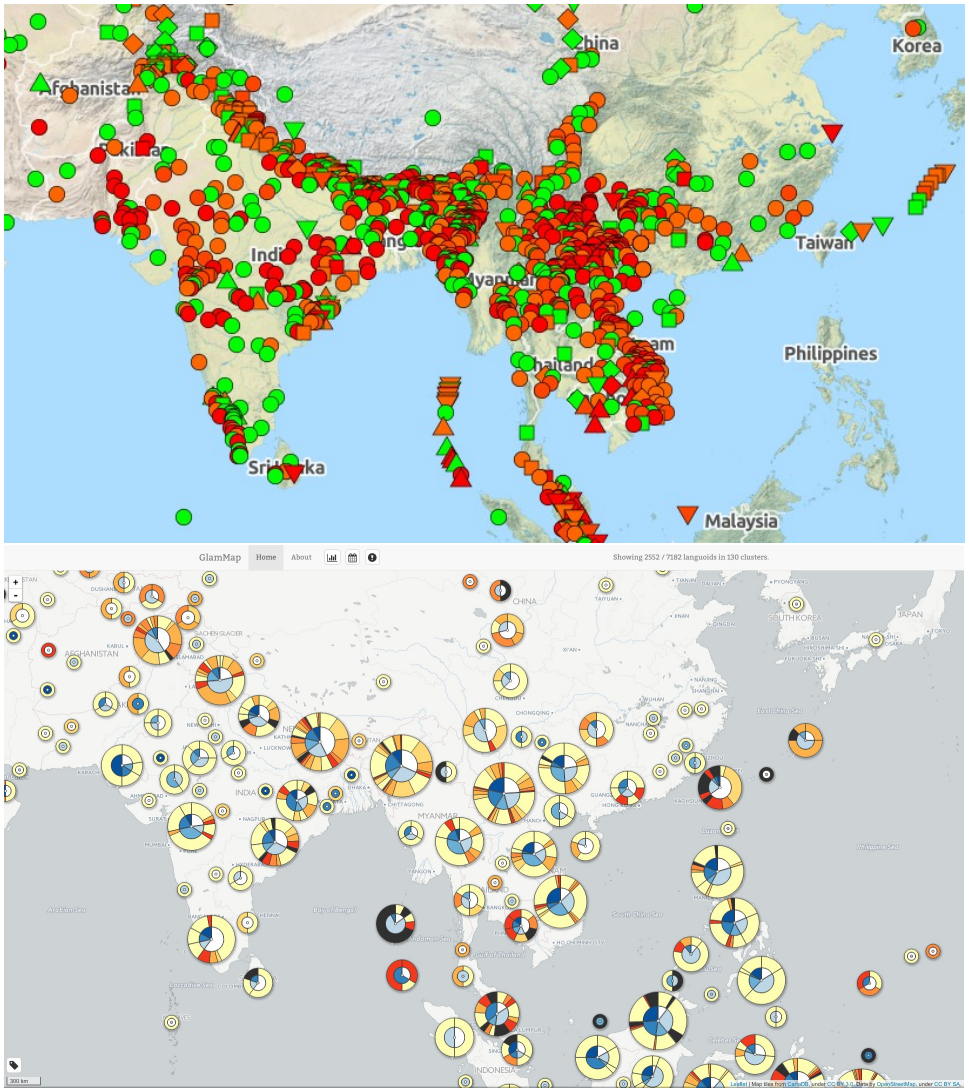
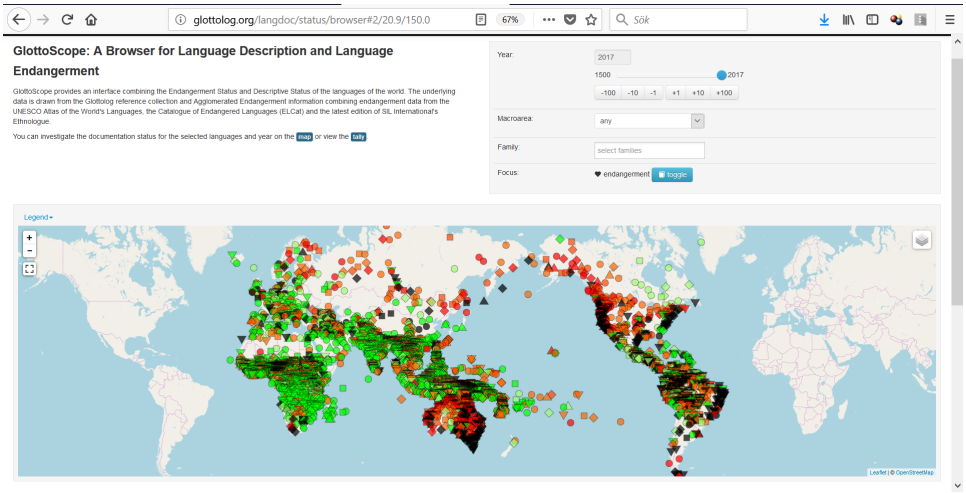


Figure 2. *Top:* A sample screenshot of GlottoScope; *bottom:* A sample screenshot of GlottoVis.

3.1.2 GlottoScope implementation and performance GlottoScope is implemented as part of the Glottolog web application, thus using the clld framework (Forkel 2014; Forkel & Bank 2016). The clld framework offers some support for interactive geographical maps, building on the leaflet Javascript library. This functionality is used by GlottoScope, which means the task of displaying maps is reduced to providing clld's mapping component with data encoded in the widely used GeoJSON format (Butler et al. 2016).

Interactivity of the map is provided by three different components: basic functionality like panning, zooming and switchable base layers is already part of the leaflet

library and third-party plugins. Loading info window content dynamically is provided by clld, and since GlottoScope is part of the Glottolog application, the data can be retrieved directly from the Glottolog database. The map legend, including the functionality to toggle map markers by endangerment or descriptive status, is implemented in bespoke code within the Glottolog application.



		Language is ...						
		safe	vulnerable	definitely endangered	severely endangered	critically endangered	extinct	
Most extensive description is a ...		●	●	●	●	●	●	**total**
long grammar	○	642	290	388	83	101	74	1578
grammar	□	401	163	191	62	60	59	936
grammar sketch	◇	807	323	434	82	115	175	1936
phonology/text	△	484	148	195	51	52	86	1016
wordlist or less	▽	924	251	570	120	112	310	2287
	total	3258	1175	1778	398	440	704	7753

Figure 3. A sample screenshot of GlottoScope.

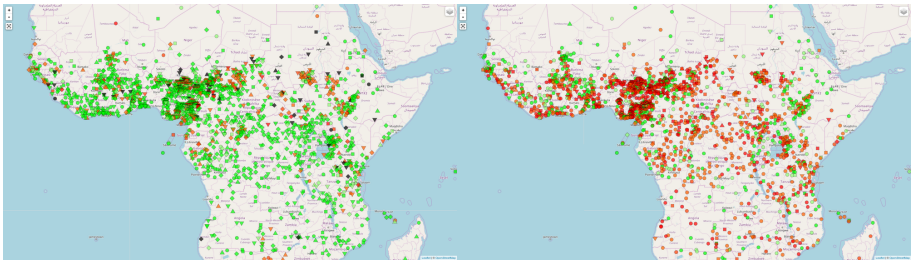
The main performance issue with this somewhat naive implementation stems from the fact that no clustering of language markers is done, i.e., the browser always has to keep the full set of language markers in memory.

The source code of the Glottolog web application, including GlottoScope, is maintained as Open Source project on GitHub. Released versions of the code are archived with and available from ZENODO (https://zenodo.org).

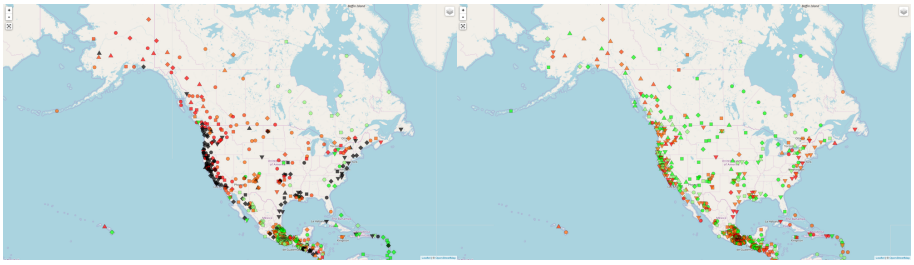
4. GlottoVis GlottoVis (http://glammap.net/glottovis/, Figure 6) provides an enhanced visualization of Agglomerated Endangerment and Descriptive Status. Figure 9 shows a screenshot, showing the selection table, map, and statistics table.

GlottoVis provides a zoomable geographic map overlaid with bivariate glyphs in the form of circles with an inner and outer section. The inner and outer sections

Figure 4. GlottoScope screenshots showing the effect of toggling the visualisation focus. Note how differently endangerment and descriptive status relate to each other in different parts of theworld: While in Central Africa endangerment seems to be less problematic, descriptive status certainly is; whereas North America shows the opposite.



(**a**) Central Africa on the left focusing on endangerment, on the right on the descriptive status



(**b**) North America on the left focusing on endangerment, on the right on the descriptive status

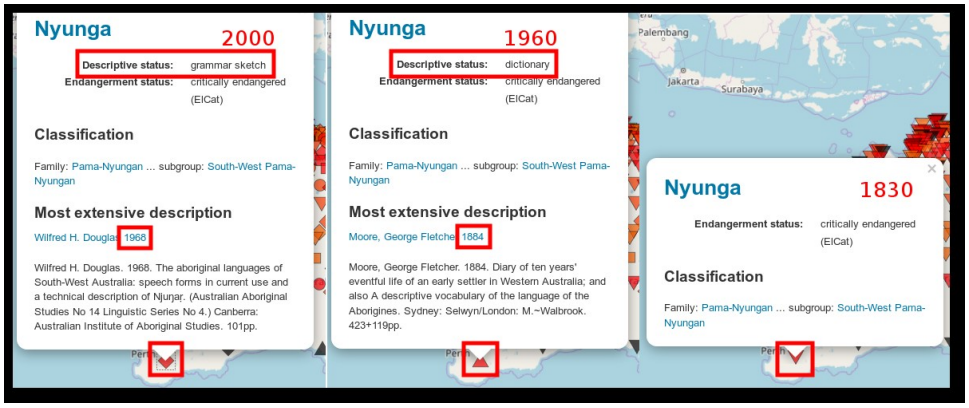


Figure 5. GlottoScope info window for the Nyunga language at three different points in time.

represent descriptive and endangerment status, respectively, but these roles can be reversed and/or combined into a single composite fill (see below). To avoid overlap of glyphs, clustering and (visual) data aggregation at each zoom level are applied. In other words, what would be too many circles at the same place are combined into larger circles reflecting the combined message of the individual circles. To achieve this, GlottoVis uses an agglomerative hierarchical clustering algorithm which guarantees

disjoint outcome glyphs on the map (see Castermans et al. 2017 for the relation to alternative techniques). A single clean glyph is easier to read than an arbitrary number of overlapping ones.

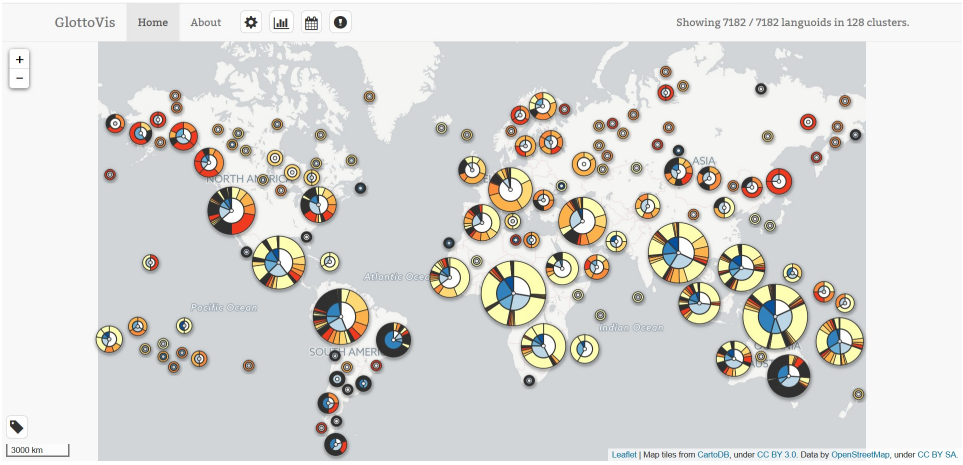


Figure 6. A sample screenshot of GlottoVis.

4.1 GlottoVis Design and Functionality The size of each glyph is increased proportionally to the number of languages it represents. Increasing its size can cause the newly constructed glyph to overlap other glyphs, so the merging process needs to be repeated until no more overlap remains. To allow users to see more details as they zoom in, the glyphs are not scaled at the same rate as the map when zooming in/out. If the user pans, the clustering does not change. Furthermore, if two glyphs are in the same cluster at a particular zoom level, then they should remain grouped at lower zoom levels.

GlottoVis uses an agglomerative approach to compute the hierarchical clustering. That is, we start with individual glyphs at the highest zoom level, and merge glyphs when they start to overlap, as we decrease the zoom level. A merged glyph is positioned halfway between the centers of the two glyphs being merged. At a particular zoom level, we can of course encounter multiple overlaps among glyphs. We could simply merge every connected set of glyphs into one cluster, but doing so may merge more glyphs than strictly necessary, since merging two glyphs may actually remove overlaps with other glyphs. Therefore, at every zoom level, we merge overlapping glyphs incrementally according to decreasing area of overlap until all glyphs are disjoint. These merged glyphs are then used to compute the clustering for the next (lower) zoom level.

4.1.1 Detailed algorithm description. The clustering hierarchy is computed only once, when initially loading the map. For n locations, there can be at most n-1 merges between two glyphs, and the next pair of glyphs to be merged can be easily found by testing all $O(n^2)$ pairs of glyphs. A naive approach would thus take $O(n^3)$ for n

locations, but this would be prohibitively slow. Instead, the hierarchical clustering is computed with the following $O(n^2)$ algorithm.

Let $P = p_1, \ldots, p_n$ be the set of point locations on the map. For every point $p_i \in P$ we compute its nearest neighbor $NN(p_i)$ in P. We then construct circles $C = c_1,\ldots,c_n$ where $c_i \in C$ is centered at p_i and its radius is the distance between p_i and $NN(p_i)$ (see Figure 7). As we zoom out, the glyphs grow. Now note that a particular glyph centered at p_i can never grow to the size of c_i, since at that point it would overlap with the glyph centered at $NN(p_i)$, and hence would have been merged earlier. Thus, to compute the next pair of glyphs that merge, it is sufficient to consider pairs (i,j) for which $c_i \cap c_j \neq \emptyset$. We can now use a geometric packing argument to show that there can be at most $O(n)$ of such pairs. In particular, every circle $c_i \in C$ can intersect with at most $O(1)$ circles in C that are at least as large as c_i.

Initially computing all nearest neighbors takes $O(n^2)$ time. Similarly, computing the pairs of circles that intersect also takes $O(n^2)$ time. Whenever two glyphs merge, the nearest neighbors and the pairs of intersecting circles must be recomputed, but only with respect to the newly created (merged) glyph. Hence, this information can easily be updated in $O(n)$ time per merge. Since there are at most n-1 merges, the total running time is $O(n^2)$.

4.1.2 Glyph design The default setup of GlottoVis shows description and endangerment status in different sections of the glyphs. We also explored four possibilities to combine the scales (see Figure 8). The first two scale combinations are formed by taking the lexicographical ordering in two different ways. This results in scales with 30 categories that are hard to read. The second two combinations are obtained by taking the index of two categories, one from each scale, and then either summing those or taking their product. For example, in the summing scale 1 + 3, 3 + 1, and 2 + 2 all map to the same category. Although the summing and the multiplication scale have fewer categories (10 and 17, respectively) than the lexicographic scales, both scales are still large. Furthermore, a different number of original categories is mapped to each combined category in the summation and multiplication scales, causing a disparity between various combined categories. Finally, the combined scales are not intuitive and it takes too much effort to translate from scale to language status. Hence, the default setting shows descriptive status and endangerment status separately inside each glyph.

The choice of glyph landed in sunburst charts (Stasko & Zhang 2000) which avoid a number of visual comparison drawbacks associated with, e.g., bar charts on a map (Castermans et al. 2017). A sunburst chart is a pie chart or donut chart that has multiple layers which should be read inside out. There is an ongoing debate about the usefulness of pie charts and donut charts (see, for example, https://eagereyes.org/blog/2015/ye-olde-pie-chart-debate) which naturally extends to sunburst charts as well. While we have not conducted a study of the usefulness of sunburst charts for visualizing descriptive and endangerment status of the languages of the world, Skau & Kosara (2016) conducted a user study on the effectiveness of pie charts and donut charts more generally. Skau & Kosara (2016) concluded that users

perform unexpectedly well in situations where they have to estimate a percentage by area only. The main intent of our glyphs is to let users estimate percentages relative to the languages represented by that glyph, so arguably sunburst charts are an appropriate choice of glyph for the present requirements. We leave a small hole in the center of our glyphs since otherwise many lines (separating wedges) might meet in a single point, giving a cluttered impression. Skau & Kosara (2016) saw no adverse effect from leaving out the center of the chart.

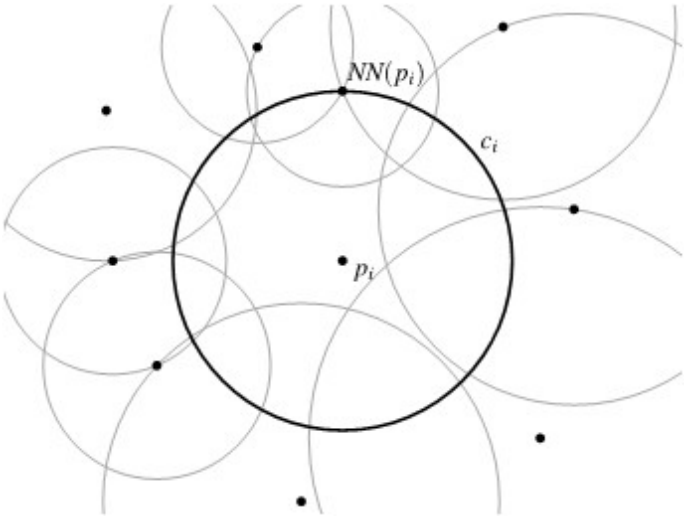


Figure 7. A location p_i, its nearest neighbor $NN(p_i)$, and the corresponding circle c_i. The other circles are shown in gray.

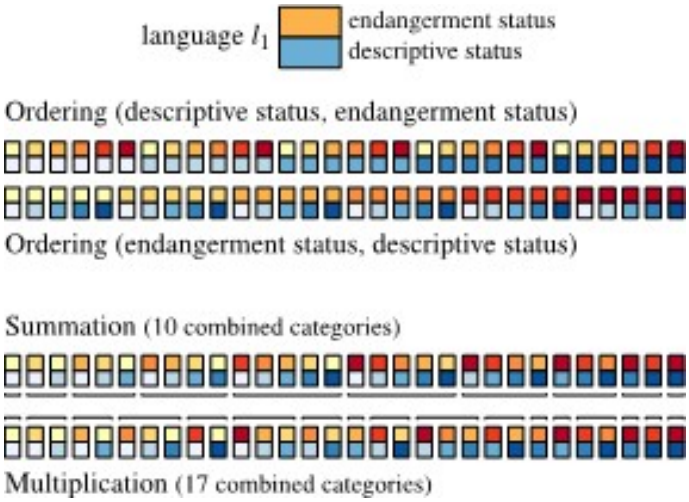


Figure 8. Four ways to combine the two status scales into a single one. The combined scale should go from overall light to dark.

To make the glyphs pop out from the map, GlottoVis uses a map layer in grayscale. The grayscale map contrasts nicely with the color palette for the endangerment status of languages, which is the outer ring of all glyphs in the default setup. GlottoVis are using the 6-class YlOrRd palette from ColorBrewer2.org (Brewer 2002; Harrower & Brewer 2003) for the endangerment status of languages, a scale ranging from yellow to red. Since the color red is associated with danger (Pravossoudovitch et al. 2014) it is intuitive to use darker red for more endangered languages, and black for extinct. The descriptive status of languages is displayed in the inner ring, for which we chose the 5-class Blues palette, which ranges from gray to blue. Because the lightest color of that palette is fairly gray, which has a low contrast with our gray map, we replaced that color with white. This indicates the highest level descriptive status, a grammar. Hence darker colors in both scales can be associated with a worse status: dark red or black means very endangered while dark blue means not, or barely, described.

Thus, scanning for lightness of colors has a meaning for scanning language status. Scanning visually for either descriptive status or endangerment status can then be done by focusing on a specific hue.

Drop shadows are utilized to indicate selection. Selected glyphs have a differently colored shadow, namely blue instead of almost black.

Figure 9 shows an overview of all components of GlottoVis except the timeline. The buttons that open components are located in the top of the interface, next to the menu. We refer to these buttons by their icon.

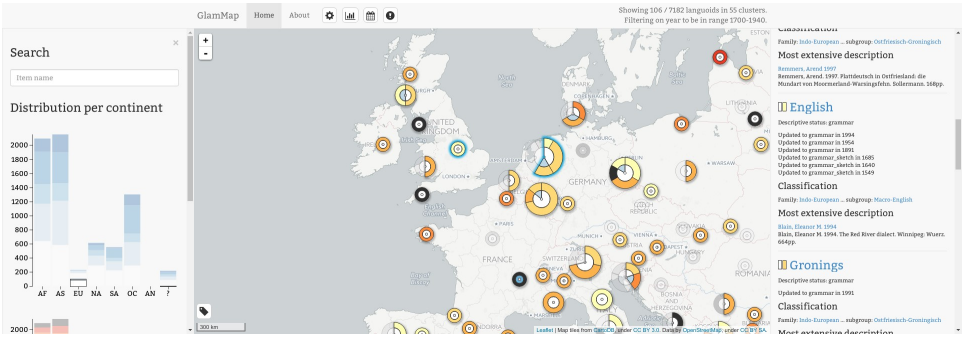


Figure 9. GlottoVis visualizes the endangerment and documentation status of the world's languages. This figure focuses on parts of Europe. A filter identifies languages whose documentation changed between 1700 and 1940. Two glyphs are selected and outlined in blue.

The main view of GlottoVis consists of a map with overlaid glyphs (see Figure 6). The map can be panned and zoomed with both keyboard and mouse. The upper right corner of the page displays the number of languages currently in view.

Mousing over a glyph opens a tooltip, which shows all languages represented by the glyph with two small squares in front that indicate their language status (see Figure 10). Mousing over these blocks shows the precise descriptive and endangerment status. The tooltips are meant to serve as quick reference, mostly in situations when not too many languages are aggregated in a glyph. For glyphs that represent

many languages, only five are shown and it is indicated how many languages are not displayed in the tooltip (see Figure 10).

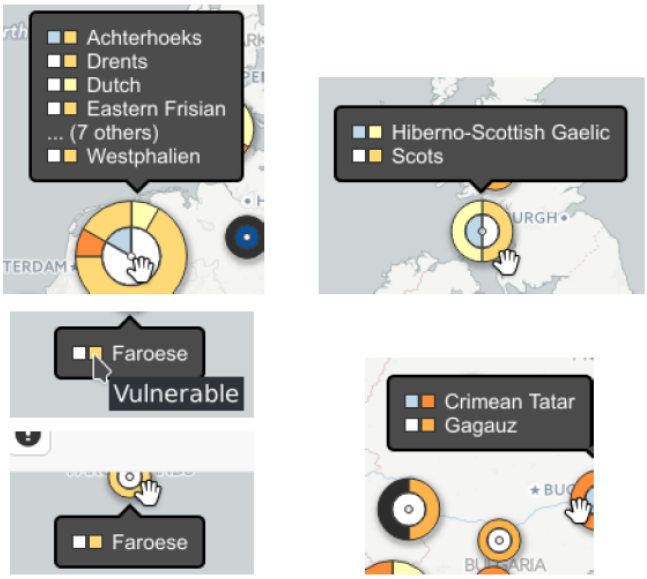


Figure 10. GlottoVis information shown when mousing over a glyph.

Selection allows the user to view more details of languages represented by a glyph, and also to view all languages represented by a big glyph. Clicking a glyph reveals a sidebar that will slide into view from the right (see Figure 11). Languages are shown in alphabetically sorted order. For every language, its name and language status are shown, but also its classification in a language family and the reference to the document that describes the language. All of these contain links to the Glottolog website, where more details can be found. There is also a brief overview of changes to the descriptive status of the language over time.

The legend can be toggled by a button in the lower left corner of the map. Clicking the legend will hide the text, which is still accessible by mousing over the colored squares. A small tooltip will then display the descriptive status or endangerment status that is indicated by the corresponding element.

The cogwheel button opens a dialog with settings. Users can change the color scheme and the configuration of the glyphs on the map: endangerment status on the inner ring and documentation status on the outer one, or only one ring, with the categories outlined above (Figure 8): summation and multiplication.

The bar chart button reveals a sidebar that slides into view from the left of the screen. It contains histograms detailing the statistical distribution of endangerment status and descriptive status of the languages currently in view (see Figure 13). To provide context, the distribution of all languages in the dataset is shown faded out in the background. Clicking a bar will apply a filter to the glyphs in view, so that only languages in the clicked continent are visible (see Figure 12). Note that it can happen

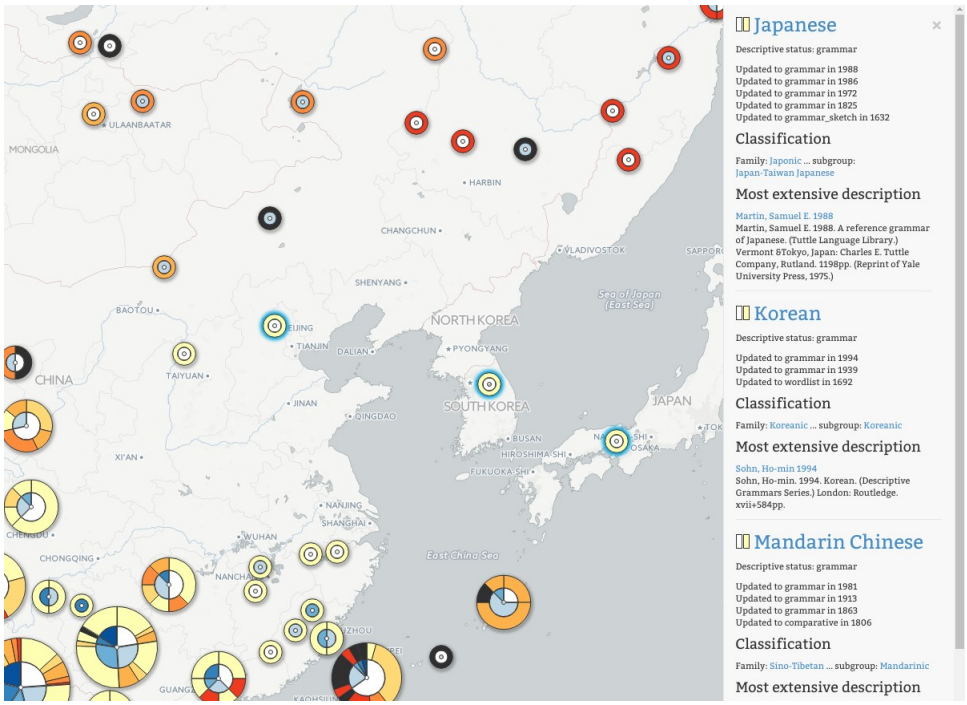


Figure 11. GlottoVis glyph selection; details of the selected languages are shown in a sidebar that slides into view from the right.

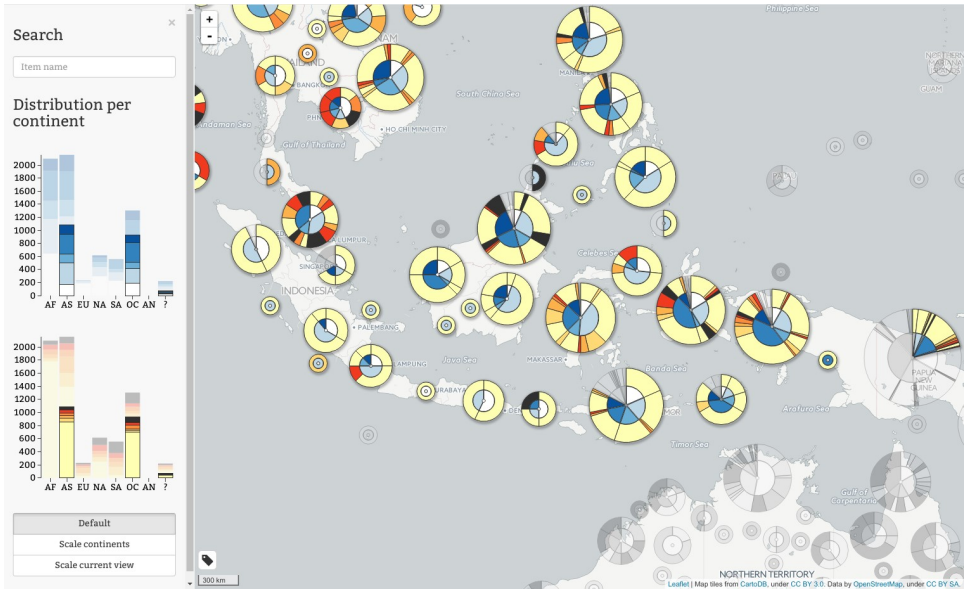


Figure 12. GlottoVis continent filters which can be activated by clicking a bar in the sidebar with histograms.

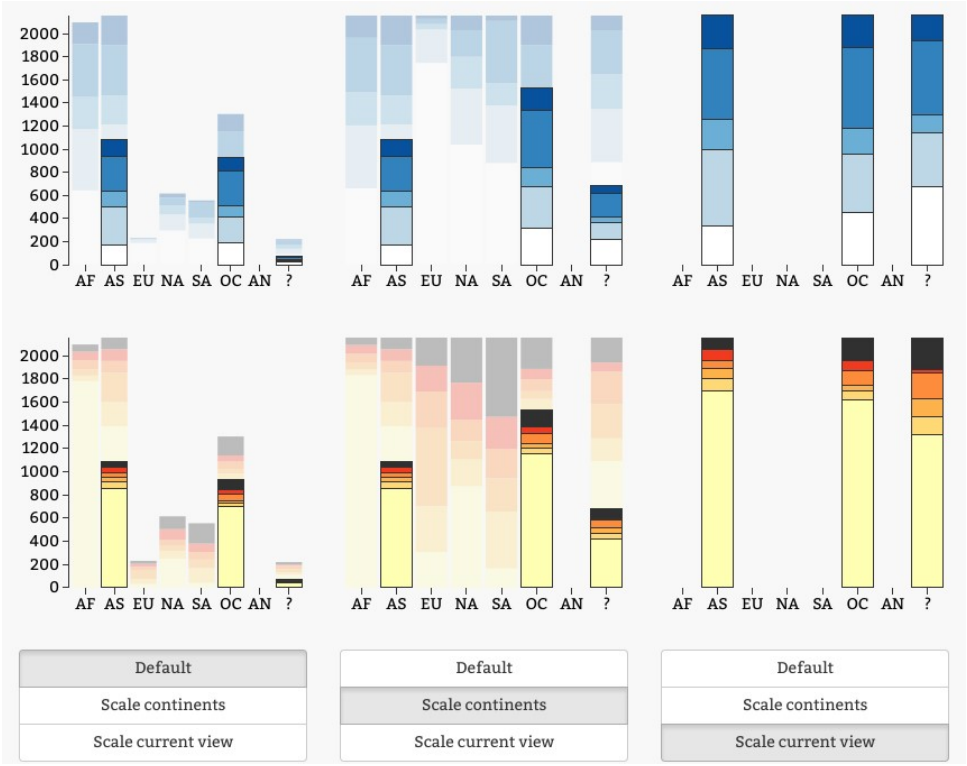


Figure 13. GlottoVis continent statistics can be displayed in histograms in a sidebar on the left (see Figure 12). The histograms can be scaled in various different ways.

that only part of a glyph is filtered out. Languages that are filtered out are grouped together within a glyph and their slices are shown in grayscale, in a semitransparent manner and also without a drop shadow. In this way they blend into the map but still give the user the appropriate context. Filtering can be stopped by clicking the same bar again. Clicking a different bar will change the filter.

The temporal aspect of the language description data is visualized in two places. Firstly, it is shown in the language detail sidebar (see Figure 11: the descriptive status of Japanese changed four times since its initial grammar sketch in 1632). Secondly, users can get an impression of the overall changes to the descriptive status of languages using the timeline. The calendar button reveals this timeline, which slides into view from the bottom of the screen (see Figure 14). The timeline is an area chart that displays the cumulative changes to the descriptive status of languages over time. This gives an impression of the documentation effort over time, for all languages combined.

The documentation effort has increased tremendously in relatively recent times. Hence, changes that occurred a long time ago can barely be seen, while the more recent part of the timeline cannot show details. There are various ways of addressing this issue. One possibility is to use a logarithmic scale on the vertical axis, but that is known to be deceptive for users. We decided on a different method, where we

explicitly emphasize the more recent part of the timeline in a two-step approach. Firstly, of the more than 18,000 changes, a very small number occur early on. We select the first 40 (an empirically determined number) changes, see in which year (1586) the last change occurs, and collect all other changes in 1586 as well. This gives us k early changes (as it happens, k=40 exactly in our case). Secondly, we allocate a fixed percentage (20%) of the width of the chart to these k changes. The latter makes the scale of the horizontal axis a piecewise linear scale with two parts. The first part, ranging from 1077 up to 1586, spans 20% of the chart width while the 1586–2016 time range spans the remaining 80%. Thus, the most recent 46% of the time range spans 80% of the width, so details can more easily be seen in the recent changes.

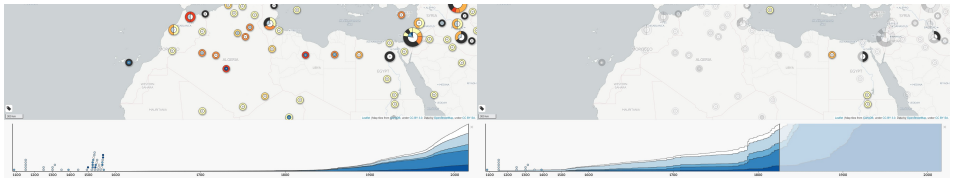


Figure 14. GlottoVis timeline visualization. *Left:* A timeline view displays the cumulative changes to descriptive status over time. A small number of initial changes is displayed separately as little discs, to make them stand out. *Right:* Filtering can be applied on the timeline. The area chart scales to the filtered part of the timeline, so that it is easier to read. Discs are hidden when the area chart overlaps them too much.

To address the problem that it is hard to see changes that occur very early on, we combine two techniques. Firstly, we explicitly show the k early changes as small discs at the corresponding horizontal position on the timeline. The color of the disc indicates the new documentation status of the represented language in that year. When discs would otherwise overlap, they are stacked vertically, sorted by documentation status. The area chart is drawn behind the discs. Users can mouse over discs to see the language and year in a tooltip. Secondly, users can select a time range (by dragging the lower and upper bounds). This causes the area chart to scale to this range (see Figure 14). Discs are hidden when the area chart overlaps more than half of their area.

Selecting the time range will not only change the timeline view, but also applies a filter to the map view. Languages for which no changes in descriptive status occurred in the selected time range are filtered out. For example, Figure 14 reveals that languages in North Africa became documented only after about 1825.

4.2 GlottoVis Implementation and Performance GlottoVis is implemented as a web application. On the server side, the clustering is performed, and requests are handled by a Ruby on Rails v4 server. Various results are cached, such as the clustering and processed history data. These are stored in JSON format and served directly to the client side when available.

The client side is implemented in JavaScript. The map view is built using Leaflet, and all charts and the glyphs on the map are constructed using D3.js v3. The tiles

that form the map are taken from CartoDB. We use Bootstrap v3 for a base layout, in combination with Font Awesome icons. Finally, we make use of several JavaScript libraries: jQuery v1.12, jQuery UI v1.11, Tooltipster v3.3, and JSS v1. Finally, we use color palettes from ColorBrewer.

In Leaflet a map can consist of multiple layers. The base layer contains the map tiles, each of which is a PNG image of 256 × 256 pixels. The glyphs are rendered in a separate layer as SVG elements. Because it can be computationally expensive for browsers to render vector graphics, we construct only glyphs that are in view, i.e., if their bounding box intersects the viewport. This improves the performance substantially.

The application needs to load a substantial amount of data (approximately 14 MB), so the initial load time depends on the speed of the users' internet connection. Download progress is indicated by a progress bar in the upper right corner of the screen. The rendering time is dependent on the browser (Google Chrome performed best at the time of writing). On a modern laptop and the internet connection of the authors, the application loads in about three seconds. Panning and zooming the map does not cause GlottoVis to download more data, so the only bottleneck there is how fast the browser can redraw the glyphs. The detailed information in the sidebar is loaded on demand. The historical data is loaded the first time the timeline is opened. Both of these types of requests have a low impact on the performance.

5. Conclusion We have presented GlottoScope and GlottoVis, two open and free web-based tools to visualize the data for language description and language endangerment. The particular challenge of this data set is the fact that two variables – language endangerment and descriptive status – need to be visible simultaneously and in close correlation. State-of-the-art databases provide the data for these visualizations but are replaceable as components.

GlottoScope is a straightforward interface that is easy to grasp quickly, but is rendered difficult for the eye to appreciate when there are too many glyphs on the map close to each other. GlottoVis is a visualization which dynamically combines what would be too many glyphs in the same place to one composite larger glyph, and allows for more design choices in the glyph make-up.

By providing these interfaces we hope to inform efforts and priorities towards language documentation and description.

6. Acknowledgements Comments and beta-testing of GlottoVis and GlottoScope were provided by a number of individuals including Hedvig Skirgård, Martin Haspelmath, Matti Miestamo, Mark Dingemanse, Lyle Campbell, Tapani Salminen, Frank Seifart, and audiences at several demo sessions.

The Netherlands Organisation for Scientific Research (NWO) is supporting B. Speckmann under project no. 639.023.208, K. Verbeek under project no. 639.021.541, and A. Betti, H. van den Berg, and T. Castermans under project no. 314.99.117.

The work of Hammarström was partly made possible thanks to the financial support of the Language and Cognition Department at the Max Planck Institute for

Psycholinguistics, Max-Planck Gesellschaft, and a European Research Council's Advanced Grant (269484 "INTERACT") awarded to Stephen C. Levinson.

References

Abley, Mark. 2003. *Spoken here: Travels among threatened languages*. London: Heinemann.

Adelaar, Willem. 1991. The endangered languages problem: South America. In Robins, R. H. & E. M. Uhlenbeck, (eds.), *Endangered languages*, 45–92. New York: Berg.

Adelaar, Willem. 2007a. Meso-America. In Moseley, Christopher (ed.), *Encyclopedia of the world's endangered languages*, 197–210. London: Routledge.

Adelaar, Willem. 2007b. Threatened languages in Hispanic South America. In Brenzinger, Matthias (ed.), *Language diversity endangered*, 9–28. Berlin: Mouton de Gruyter.

Becker-Donner, Etta. 1962. Guaporé-Gebiet. *Bulletin of the International Committee on Urgent Anthropological and Ethnological Research* 5. 146–150.

Been, Ken, Eli Daiches & Chee Yap. 2006. Dynamic map labeling. *IEEE Transactions on Visualization and Computer Graphics* 12(5). 773–780.

Bertin, Jacques. 1967. *Semiology of graphics: Diagrams, networks and maps*. Madison: The University of Wisconsin Press. (Translation from French: *Sémiologie graphique*).

Bird, Steven. 2010. A scalable method for preserving oral literature from small languages. In Chowdhury, Gobinda, Chris Koo & Jane Hunter (eds.), *Proceedings of the Role of Digital Libraries in a Time of Global Change and the 12th International Conference on Asia-Pacific Digital Libraries*, 5–14. Berlin: Springer.

Bird, Steven, Florian R. Hanke, Oliver Adams & Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 1–5. Stroudsburg, PA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W14/W14-2201.

Blench, Roger. 2007. Endangered languages in West Africa. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 140–162. Berlin: Mouton de Gruyter.

Bradley, David. 2007a. East and Southeast Asia. In Moseley, Christopher (ed.), *Encyclopedia of the world's endangered languages*, 349–424. London: Routledge.

Bradley, David. 2007b. Language endangerment in China and Mainland Southeast Asia. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 278–302. Berlin: Mouton de Gruyter.

Brenzinger, Matthias. 2007a. Language endangerment in Northern Africa. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 123–139. Berlin: Mouton de Gruyter.

Brenzinger, Matthias. 2007b. Language endangerment in Southern and Eastern Africa. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 179–204. Berlin: Mouton de Gruyter.

Brenzinger, Matthias, Arienne M. Dwyer, Tjeerd de Graaf, Colette Grinevald, Michael Krauss, Osahito Miyaoka, Nicholas Ostler, et al. 2003. Language vitality and endangerment. Document submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages, Paris, 10–12 March 2003.

Brewer, Cynthia A. 2002. ColorBrewer. http://colorbrewer2.org/.

Butler, H., M. Daly, A. Doyle, S. Gillies, S. Hagen & T. Schaub. 2016. *The GeoJSON Format*. Internet Engineering Task Force (IETF) Request for Comments: 7946. https://tools.ietf.org/html/rfc7946.

Caines, Andrew, Christian Bentz, Dimitrios Alikaniotis, Fridah Katushemererwe & Paula Buttery. 2016. The Glottolog data explorer: Mapping the world's languages. In Hautli-Janisz, Annette & Verena Lyding (eds.), *Proceedings of VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, 38–53. Portorož, Slovenia: European Language Resources Association (ELRA).

Campbell, Lyle. 2017. About the catalogue of the endangered languages. http://endangeredlanguages.com/about/. Accessed 07-27-2017.

Campbell, Lyle & Anna Belew. 2018. Introduction: Why catalogue endangered languages? In Campbell, Lyle & Anna Belew (eds.), *Cataloguing the world's endangered languages*, 1–14. London: Routledge.

Campbell, Lyle & Kenneth Rehg. 2018. Introduction. In Campbell, Lyle & Kenneth Rehg (eds.), *The Oxford handbook of endangered languages*, 1–18. Oxford: Oxford University Press.

Capell, Arthur. 1962. Linguistic research needed in Australia. *Bulletin of the International Committee on Urgent Anthropological and Ethnological Research* 5. 23–28.

Castermans, Thom, Harald Hammarström, Bettina Speckmann, Kevin Verbeek & Michel Westenberg. 2017. GlottoVis: Visualizing language endangerment and documentation. In Collins, Christopher, Michael Correll, Mennatallah El-Assady, Stefan Jänicke, Daniel Keim & David Wrisley (eds.), *VIS4DH'17: 2nd Workshop on Visualization for the Digital Humanities*, 1–5. Phoenix: IEEE.

Chelliah, Shobhana L. & Willem J. de Reuse. 2011. *Handbook of descriptive linguistic fieldwork*. Dordrecht: Springer.

Connell, Bruce. 2007. Endangered languages in Central Africa. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 163–178. Berlin: Mouton de Gruyter.

Crevels, Mily. 2007. South America. In Moseley, Christopher (ed.), *Encyclopedia of the world's endangered languages*, 103–196. London: Routledge.

Crystal, David. 2000. *Language death*. Cambridge: Cambridge University Press. Dalby, Andrew. 2003. *Language in danger: The loss of linguistic diversity and the threat to our future*. New York: Columbia University Press.

Dimmendaal, Gerrit J. & F. K. Erhard Voeltz. 2007. Africa. In Moseley, Christopher (ed.), *Encyclopedia of the world's endangered languages*, 579–634. London: Routledge.

Evans, Nicholas. 2007. Warramurrungunji undone: Australian languages in the 51st millennium. In Brenzinger, Matthias (eds.), *Language diversity endangered* (Trends in Linguistics 181), 342–373. Berlin: Mouton de Gruyter.

Evans, Nicholas. 2009. *Dying words: Endangered languages and what they have to tell us*. Oxford: John Wiley & Sons.

Evans, Nicholas & Stephen Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32(5):429–492.

Forkel, Robert. 2014. The cross-linguistic linked data project. In Chiarcos, Christian, John Philip McCrae, Petya Osenova & Cristina Vertan (eds.), *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, 60–66. Reykjavik: European Language Resources Association (ELRA).

Forkel, Robert & Sebastian Bank. 2016. CLLD: A toolkit for cross-linguistic databases. Python Library Archived at zenodo.org (10.5281/zenodo.55099). https://-doi.org/10.5281/zenodo.55099.

Golla, Victor. 2007. North America. In Moseley, Christopher (ed.), *Encyclopedia of the world's endangered languages*, 1–96. London: Routledge.

Grenoble, Lenore A. & Lindsay J. Whaley. 1998. *Endangered languages: Language loss and community response*. Cambridge: Cambridge University Press.

Grinevald, Colette. 2007. Endangered languages of Mexico and Central America. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 59–86. Berlin: Mouton de Gruyter.

Hammarström, Harald. 2010. The status of the least documented language families in the world. *Language Documentation & Conservation* 4. 177–212. http://hdl.handle.net/10125/4478.

Hammarström, Harald. 2011. Automatic annotation of bibliographical references for descriptive language materials. In Forner, Pamela, Julio Gonzalo, Jaana Kekäläinen, Mounia Lalmas & Maarten de Rijke (eds.), *Multilingual and Multimodal Information Access Evaluation. CLEF 2011. Lecture Notes in Computer Science*, Vol. 6941, 62–73. Berlin: Springer.

Hammarström, Harald. 2015. Ethnologue 16/17/18th editions: A comprehensive review. *Language* 91(3). 723–737.

Hammarström, Harald, Sebastian Bank, Robert Forkel & Martin Haspelmath. 2017. Glottolog 3.1. Jena: Max Planck Institute for the Science of Human History. http://-glottolog.org. Accessed 12-01-2017.

Hammarström, Harald & Mark Donohue. 2014. Some principles on the use of macroareas in typological comparison. In Hammarström, Harald & Lev Michael (ed.), *Quantitative approaches to areal linguistic typology*, 167–187. Leiden: Brill. (Language Dynamics & Change Special Issue.)

Hammarström, Harald & Sebastian Nordhoff. 2011. LangDoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language* 3(2). 31–43.

Harrison, David K. 2007. *When languages die*. Oxford Studies in Sociolinguistics. Oxford: Oxford University Press.

Harrower, Mark & Cynthia A. Brewer. 2003. ColorBrewer.Org: An online tool for selecting colour schemes for maps. *The Cartographic Journal* 40(1). 27–37.

Hauk, Bryn & Raina Heaton. 2018. Triage: Setting priorities for endangered language research. In Campbell, Lyle & Anna Belew (eds.), *Cataloguing the world's endangered languages*, 259–304. London: Routledge.

Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.

Himmelmann, Nikolaus. 2006. Language documentation: What is it and what is it good for? In Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation* (Trends in Linguistics 178), 1–30. Berlin: Mouton de Gruyter.

Himmelmann, Nikolaus. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation & Conservation* 6. 187–207.

Kazakevich, Olga & Aleksandr Kibrik. 2007. Language endangerment in the Cis. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 233–262. Berlin: Mouton de Gruyter.

Kibrik, Aleksandr E. 1991. The problem of endangered languages in the USSR. *Diogenes* 153. 67–83.

Krauss, Michael E. 1992. The world's languages in crisis. *Language* 68(1). 1–10.

Krauss, Michael E. 2007. Mass language extinction and documentation: The race against time. In Miyaoka, Osahito, O. Sakiyama & Michael Krauss (eds.), *Vanishing languages of the Pacific Rim*, 3–24. Oxford: Oxford University Press.

Lee, Nala H. & John R. Van Way. 2018. The language endangerment index. In Campbell, Lyle & Anna Belew (eds.), *Cataloguing the world's endangered languages*, 66–78. London: Routledge.

Lewis, Paul M. & Gary F. Simons. 2010. Assessing endangerment: Expanding Fishman's Gids. *Revue Roumaine de Linguistique* 55(2). 103–120.

Lewis, Paul M., Gary F. Simons & Charles D. Fennig. 2013. *Ethnologue: Languages of the world*. 17th ed. Dallas: SIL International. http://www.ethnologue.com.

Moore, Denny. 2007. Endangered languages of Lowland Tropical South America. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 29–58. Berlin: Mouton de Gruyter.

Moseley, Christopher. 2010. *Atlas of the world's languages in danger*. 3rd ed. Paris: UNESCO Publishing.

Nettle, Daniel & Suzanne Romaine. 2000. *Vanishing voices: The extinction of the world's languages*. Oxford: Oxford University Press.

Owens, Jonathan. 2007. Endangered languages of the Middle East. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 263–277. Berlin: Mouton de Gruyter.

Pravossoudovitch, Karyn, Francois Cury, Steve G. Young & Andrew J. Elliot. 2014. Is red the colour of danger? Testing an implicit red-danger association. *Ergonomics* 57(4). 503–510.

Rhodes, Richard A. & Lyle Campbell. 2018. The goals of language documentation. In Campbell, Lyle & Kenneth Rehg (eds.), *The Oxford handbook of endangered languages*, 107–122. Oxford: Oxford University Press.

Salminen, Tapani. 2007a. Endangered languages in Europe. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 205–232. Berlin: Mouton de Gruyter.

Salminen, Tapani. 2007b. Europe and North Asia. In Moseley, Christopher (ed.), *Encyclopedia of the world's endangered languages*, 211–282. London: Routledge.

Sands, Bonny. 2017. The challenge of documenting Africa's least known languages. In Kandybowicz, Jason & Harold Torrence (eds.), *Africa's endangered languages: Documentary and theoretical approaches*, 11–38. Oxford: Oxford University Press.

Scheepens, Roeland, Huub van de Wetering & Jarke Jack van Wijk,. 2014. Non-overlapping aggregated multivariate glyphs for moving objects. In *Proceedings of the 7th IEEE Pacific Visualization Symposium*, 17–24. Seoul, South Korea: IEEE.

Seifart, Frank, Harald Hammarström, Nicholas Evans & Stephen C. Levinson. In press. Language documentation 25 years on. *Language*.

Simons, Gary F. & Charles D. Fennig. 2017. *Ethnologue: Languages of the world*. 20th ed. Dallas: SIL International. http://www.ethnologue.com.

Skau, Drew & Robert Kosara. 2016. Arcs, angles, or areas: Individual data encodings in pie and donut charts. *Computer Graphics Forum* 35(3). 121–130.

Skinner, Leonard E. & Marlene B. Skinner. 2000. *Diccionario Chinanteco de San Felipe Usila, Oaxaca*. Vol. 43. Serie de Vocabularios Y Diccionarios Indígenas Mariano Silva Y Aceves. Coyoacán, México: Instituto Lingüístico de Verano.

Stasko, John & Eugene Zhang. 2000. Focus + context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proceedings of the 6th IEEE Symposium on Information Visualization*, 57–65. Los Alamitos: IEEE Scientific Press.

Stone, Doris. 1962. Urgent tasks of research concerning the cultures and languages of Central American Indian Tribes. *Bulletin of the International Committee on Urgent Anthropological and Ethnological Research* 5. 65–69.

Swadesh, Morris. 1960. Problems in language salvage for prehistory. *Bulletin of the International Committee on Urgent Anthropological and Ethnological Research* 3. 15–19.

Thomason, Sarah G. 2015. *Endangered languages: An introduction*. Cambridge: Cambridge University Press.

Tryon, Darrell. 2007. The languages of the Pacific Region: The Austronesian languages of Oceania. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 391–409. Berlin: Mouton de Gruyter.

Tufte, Edward R. & P.R. Graves-Morris. 1983. *The visual display of quantitative information*. Vol. 2. Chesire, Connecticut: Graphics Press.

van Driem, George. 2007a. Endangered languages of South Asia. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 303–341. Berlin: Mouton de Gruyter.
van Driem, George. 2007b. South Asia and the Middle East. In Moseley, Christopher (ed.), *Encyclopedia of the world's endangered languages*, 283–348. London: Routledge.
Vehlow, C., T. Reinhardt & D. Weiskopf. 2013. Visualizing fuzzy overlapping communities in networks. *IEEE Transactions on Visualization and Computer Graphics* 19(12). 2486–2495.
Ward, Matthew O. 2008. Multivariate data glyphs: Principles and practice. In Chen, C., W. Härdle & A. Unwin (eds.), *Handbook of data visualization* (Springer handbooks of computational statistics), 179–198. Berlin: Springer.
Woodbury, Anthony. 2011. Language documentation. In Austin, Peter & Julia Sallabank (eds.), *Handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.
Wurm, Stefan. 1956. Die dringendsten linguistischen aufgaben in Neuguinea. In *Ethnologica, seconde partie et rapport general* III, 289–292. Actes du Ive Congrès International Des Sciences Anthropologiques et Ethnologiques. Vienna: Adolf Holzhausens.
Wurm, Stephen A. 1991. Language death and disappearance: Causes and circumstances. *Diogenes* 39(153). 1–18.
Wurm, Stefan. 2007a. Australasia and the Pacific. In Moseley, Christopher (ed.), *Encyclopedia of the world's endangered languages*, 425–578. London: Routledge.
Wurm, Stefan. 2007b. Threatened languages in the Western Pacific Area from Taiwan to, and including, Papua New Guinea. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 374–390. Berlin: Mouton de Gruyter.
Yamamoto, Akira. 2007. Endangered languages in USA and Canada. In Brenzinger, Matthias (ed.), *Language diversity endangered* (Trends in Linguistics 181), 87–122. Berlin: Mouton de Gruyter.
Zaborski, Andrzej. 1970. Cushitic languages: An unexplored subcontinent. *Bulletin of the International Committee on Urgent Anthropological and Ethnological Research* 12. 119–128.

Harald Hammarström
harald.hammarstrom@lingfil.uu.se
https://orcid.org/0000-0003-0120-6396

Thom Castermans
t.h.a.castermans@tue.nluthor
https://orcid.org/0000-0002-9282-6760

Robert Forkel
forkel@shh.mpg.de
https://orcid.org/0000-0003-1081-086X

Kevin Verbeek
k.a.b.verbeek@tue.nl
https://orcid.org/0000-0003-3052-4844

Michel A. Westenberg
m.a.westenberg@tue.nl

Bettina Speckmann
b.speckmann@tue.nl
https://orcid.org/0000-0002-8514-7858