LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 16 (2022), pp. 290-318
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/74681
Revised Version Received: 12 Jun 2022

Linking endangerment databases and descriptive linguistics: An assessment of the use of terms relating to language endangerment in grammars

Roberto Zariquiey
Pontificia Universidad Católica del Perú

Mónica Arakaki
Pontificia Universidad Católica del Perú

Javier Vera
Pontificia Universidad Católica del Perú

Guido Torres-Orihuela
Universidad Nacional de San Agustín de Arequipa

Claret Cuba-Raime
Universidad Nacional de San Agustín de Arequipa

Carlos Barrientos
Pontificia Universidad Católica del Perú

Aracelli García
Pontificia Universidad Católica del Perú

Adriano Ingunza
Pontificia Universidad Católica del Perú

Harald Hammarström
Uppsala University

E-ISSN 1934-5275

The world harbours 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. The increased urgency has led to the development of several global endangerment databases and a more fine-grained understanding of the language endangerment progression as well as its possible reversal. In the present paper, we explore the terminological correlates of this development as found in the descriptive linguistic literature, using a corpus of over 10,000 digitized grammatical descriptions. Comparing this with existing endangerment databases, we find that simply counting terms related to endangerment does signal endangerment, but the degree of endangerment is more difficult to assess from grammatical descriptions. The label endangered seems to be an umbrella term that covers different situations ranging from moribund languages with less than ten speakers to minority languages with several thousand speakers. For many languages considered endangered in existing databases, explicit terms to this effect cannot be found in their descriptions. The discrepancy is due to incompleteness of the search-term set, gaps in the literature, and projected rather than observed information in the databases. Our explorations illustrate the potential for database curation assisted by computational searches both to maintain accuracy of the databases and to investigate assumed language endangerment. Future work includes a larger cloud of search terms, usage of term frequencies, and prescreening of descriptive literature for the existence of a relevant section. From the perspective of descriptive linguistics, this study calls for a more careful correlation between the language endangerment indexes, as developed in the global endangerment databases, and the treatment of the endangerment status of individual languages in descriptive grammars.

1. Introduction The diversity of 6,500 mutually unintelligible languages (Hammarström 2015: 733) found in the world 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; Swadesh 1960; Becker-Donner 1962; Capell 1962; Stone 1962; Zaborski 1970; Adelaar 1991; Kibrik 1991; Wurm 1991; Krauss 2007; Sands 2017; Campbell & Rehg 2018), and especially since Krauss's (1992) seminal article, many minority languages are becoming endangered and will be lost forever if not documented.

There is now a range of books describing the endangerment processes and their consequences in generalized case studies (Grenoble & Whaley 1998; Crystal 2000; Nettle & Romaine 2000; Abley 2003; Dalby 2003; Harrison 2007; Evans 2009; Thomason 2015), as well as three global databases – UNESCO's *Atlas of the World's Languages in Danger* (Moseley 2010), *Ethnologue: Languages of the World* (Eberhard et al. 2021), and the Catalogue of Endangered Languages (ELCat)[1] – which report the endangerment status of individual languages.[2] Although the databases are extremely valuable, they struggle to stay updated, partly lack individual sources, and do not aim to systematically carry any further information than the static vitality label. For a better understanding of endangerment dynamics – especially as it concerns broad empirical trends – richer and more temporally controlled data are desirable. At the same time, a thorough collection of digitized descriptive literature is available for research purposes (see Virk et al. 2020 and also §2.1). Although there are obvious limitations of "blind" searches for terms, there is the potential that (semi-) automatic searches over this collection could enrich and speed up the collection of language endangerment data to some degree. The present study explores the immediate prospects for this avenue of investigation, calling for a more integrated approach to language endangerment in both grammatical descriptions and endangerment databases, as a strategy to promote more productive interactions between descriptive work and endangerment databases.

2. Data

2.1 Digital collection of descriptive literature The full collection consists of over 37,000 digitized books and articles relating to descriptive linguistics. The most important subset is made up of some 12,000 grammatical descriptions (see Virk et al. 2020), but the collection also includes dictionaries, sociolinguistic studies, phonologies, comparative studies, text collections, overviews, word lists, and bibliographies (Hammarström & Nordhoff 2011). The collection comprises (1) out-of-copyright texts digitized by national libraries, archives, scientific societies, and other similar entities; (2) texts posted online with a licence to be used for research, usually by

[1] See http://www.endangeredlanguages.com (accessed 2022-03-30).

[2] The database cited in Hammarström et al. (2018) also contains individual endangerment data synthesized from these three databases (see §2.2).

university libraries and nonprofit organizations (notably SIL International); and (3) texts under publisher copyright where quotations of short extracts are legal. A listing of the collection can be accessed via the open-access bibliography Glottolog (Hammarström et al. 2021a).[3] For each reference pertaining to the present study, this catalogue features manually curated annotations of

1. the language it is written in (the meta-language, usually English, French, German, Spanish, Russian, or Mandarin Chinese);
2. the language(s) described in it (the vernacular, typically one of the thousands of minority languages throughout the world); and
3. the type of description (e.g., comparative study, description of a specific feature, phonological description, grammar, bibliography, sociolinguistic study, overview).

The collection has been digitized into machine-readable text through *ABBYY Finereader 14*, an OCR (optical character recognition) software, using the meta-language as the recognition language. The original digital documents are of varying quality, from barely legible typescript copies to high-quality scans and even born-digital documents. Contemporary OCR techniques rely heavily on dictionaries of major languages, and consequently, most tokens of the meta-languages are accurately reflected, while tokens of the vernacular(s) are hopelessly misrecognized. Since the search terms in the experiments to follow are in the meta-language, we have little reason to believe that OCR quality plays any significant role.

For the present study, we selected the sociolinguistic studies and grammatical descriptions (which as a rule are prefaced with a sociolinguistic section) as the bibliographical types where we systematically expected to find endangerment information. In this experimental study, we only considered documents in the (meta-)language English, where the prospects seemed to be explorable in the most straightforward manner. The results in question can easily be transferred to cover other meta-languages by translating the relevant terms. We also restricted the search to documents describing exactly one language so that any term occurrences can arguably be related to exactly that language. The final selection amounted to 7,088 documents spanning 3,214 languages.

2.2 Language endangerment data For evaluation, we compared our search extraction results with existing language endangerment data. The database mentioned in Hammarström et al. (2018) combines the scales from the following three databases: (1) the UNESCO scale for Atlas of the World's Languages in Danger (Moseley 2010), (2) the EGIDS (Expanded Graded Intergenerational Disruption Scale) for Ethnologue, and (3) the LEI (Language Endangerment Index) for ELCat. The result is an Agglomerated Endangerment Scale (AES) for every language using EGIDS-inspired labels, as per Table 1.

[3] https://doi.org/10.5281/zenodo.4761960 (Accessed on 2021-05-20.)

Table 1. Mappings between the endangerment categories in the source databases and the Agglomerated Endangerment Scale (AES) from Hammarström et al. (2018: 372)[4]

UNESCO	LEI-ELCat	EGIDS	AES
Safe	At risk	1 (National)	Not endangered
		2 (Regional)	
		3 (Trade)	
		4 (Educational)	
		5 (Written)	
		6a (Vigorous)	
Vulnerable	Vulnerable	6b (Threatened)	6b (Threatened)
Definitely endangered	Threatened	7 (Shifting)	7 (Shifting)
	Endangered		
Severely endangered	Severely endangered	8a (Moribund)	8a (Moribund)
Critically endangered	Critically endangered	8b (Nearly extinct)	8b (Nearly extinct)
Extinct	Dormant	9 (Dormant)	10 (Extinct)
	Awakening	9 (Reawakening)	
		9 (Second language only)	
		10 (Extinct)	

Note: AES = Agglomerated Endangerment Scale; EGIDS = Expanded Graded Intergenerational Disruption Scale; ELCat = Catalogue of Endangered Languages; LEI = Language Endangerment Index.

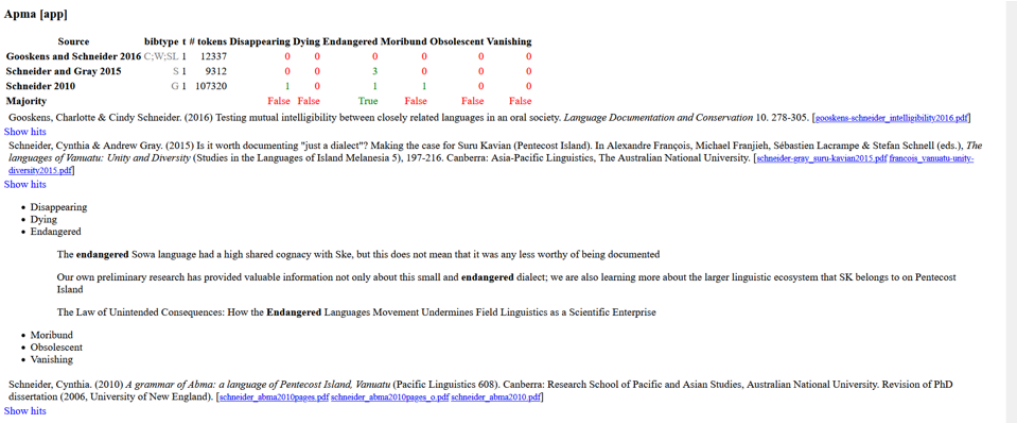
A problematic aspect is that the underlying databases often do not provide a source for the data indicated, and often when there is a source, it is an overview, which itself does not give individual sources (Hammarström et al. 2018: 366–369). Thus, the information often cannot be traced down to an underlying observation. This drawback is one of the motivations for the present study, which attempts to link data to individual sources and the observations therein.

3. Experiments Keyword searches (Hammarström, Her, & Tang 2021) were carried out with the Gramfinder tool (Hammarström 2021) by counting the number of occurrences of the terms *moribund, severely endangered, highly endangered, disap-*

[4] The mappings were elaborated by Dr. Frank Seifart based on the definitions in the respective scales.

pearing, *vanishing*, *dying*, *endangered*, and *obsolescent*. These terms are the most common terms relating to endangerment and the most obvious candidates for correspondences with the scales in the databases used for comparison (cf. Table 1). Morphological and capitalized variants of the terms were counted as well. For the multiple-word searches, such as *severely endangered* and *highly endangered*, matches were counted if the qualifier occurred in the same sentence as "endangered," not only when immediately adjacent. An initial screening revealed that the terms *disappearing*, *dying*, and *vanishing* were particularly prone to usage in a different context than the one sought after. For example, the term *dying* often occurred in language examples. Therefore, these three terms were counted *only if* preceded by "language is" or followed by "language," which effectively, and rather commonly, disambiguated the context to the desired one. Our search did not include the term *extinct* since we are interested in endangerment rather than cases of extinction. The existing databases are also more reliable with respect to extinction rather than degree of endangerment since the former can often be observed more securely and enduringly. Where relevant, we will denote the absence of hits of any of the terms by the label *not endangered*.

For each language, there may be several relevant sources. The Gramfinder tool output the findings for each source and, by default, assigned a result to the language as a whole by majority vote (with ties broken in favour of the positive) (Hammarström 2021). A snapshot of the search output is given in Figure 1.



Apma [app]

Source	bibtype	t	# tokens	Disappearing	Dying	Endangered	Moribund	Obsolescent	Vanishing
Gooskens and Schneider 2016	C;W;SL	1	12337	0	0	0	0	0	0
Schneider and Gray 2015	S	1	9312	0	0	3	0	0	0
Schneider 2010	G	1	107320	1	0	1	1	0	0
Majority				False	False	True	False	False	False

Gooskens, Charlotte & Cindy Schneider. (2016) Testing mutual intelligibility between closely related languages in an oral society. *Language Documentation and Conservation* 10. 278-305. [gooskens-schneider_intelligibility2016.pdf]

Show hits

Schneider, Cynthia & Andrew Gray. (2015) Is it worth documenting "just a dialect"? Making the case for Suru Kavian (Pentecost Island). In Alexandre François, Michael Franjieh, Sébastien Lacrampe & Stefan Schnell (eds.), *The languages of Vanuatu: Unity and Diversity* (Studies in the Languages of Island Melanesia 5), 197-216. Canberra: Asia-Pacific Linguistics, The Australian National University. [schneider-gray_suru-kavian2015.pdf francois_vanuatu-unity-diversity2015.pdf]

Show hits

- Disappearing
- Dying
- Endangered

The **endangered** Sowa language had a high shared cognacy with Ske, but this does not mean that it was any less worthy of being documented

Our own preliminary research has provided valuable information not only about this small and **endangered** dialect; we are also learning more about the larger linguistic ecosystem that SK belongs to on Pentecost Island

The Law of Unintended Consequences: How the **Endangered** Languages Movement Undermines Field Linguistics as a Scientific Enterprise

- Moribund
- Obsolescent
- Vanishing

Schneider, Cynthia. (2010) *A grammar of Abma: a language of Pentecost Island, Vanuatu* (Pacific Linguistics 608). Canberra: Research School of Pacific and Asian Studies, Australian National University. Revision of PhD dissertation (2006, University of New England). [schneider_abma2010pages.pdf schneider_abma2010pages_o.pdf schneider_abma2010.pdf]

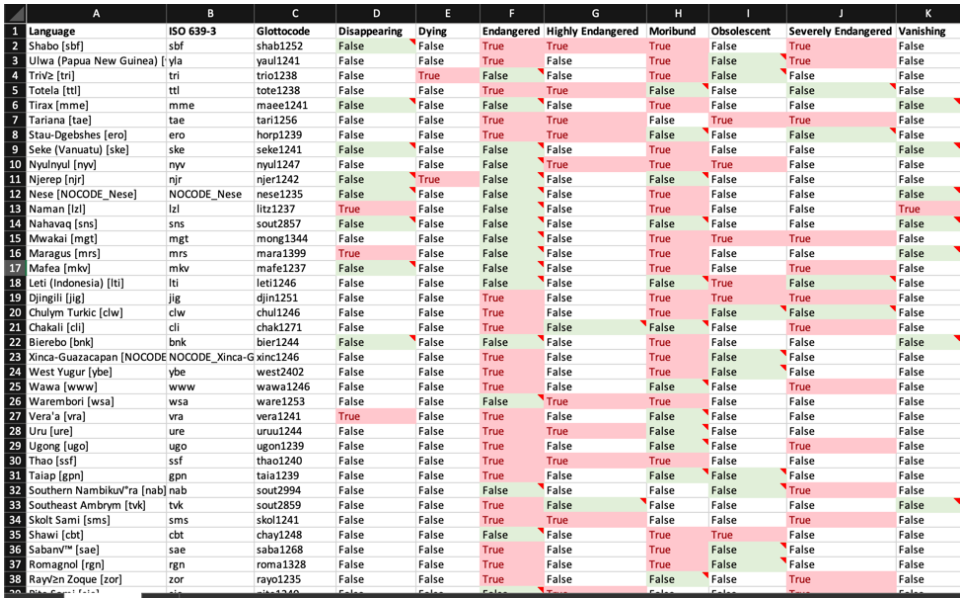
Show hits

Figure 1. Sample search/extraction output from the Gramfinder tool for the language Apma [app] in Vanuatu. For each language and corresponding grammatical descriptions, the number of hits is shown, alongside the threshold t (here set to 1 – one hit is sufficient – but can in general be automatically calculated; consult Hammarström, Her, & Tang 2021 for details). The sources are provided with links to full-text and displayable hit snippets.

Naturally, to expect the term occurrence to be perfectly indicative of the vitality status of the described language is naive and will likely result in false positives – for instance, if the term occurs in a different context, as in an occasional comment pertaining to another language than the main one in focus or in a negated sentence (Hammarström, Her, & Tang 2021). We thus contrast three usages of the search results. The naive usage (NU) uses the per-language label of the default Gramfinder output. The most recent source usage (RU) follows the most recent source for each language, conceding to the suspicion that this is more indicative of the language's actual vitality status. For the corrected usage (CU), NU hits are screened and corrected by a human. Normally, human curation of data from descriptive materials is a very time-consuming task, but because the hits were collected and organized by the computer, screening the appropriateness of positive hits can be done relatively quickly (Hammarström 2021). For the curation process, we checked each positive hit to determine whether the term was truly used to characterize the vitality status of the language under discussion. If this were the case, we coloured the cell in red (see Figure 2). Mistaken uses of an endangerment term (false positives) were coloured in green, and the coding value was changed from 'TRUE' to 'FALSE.' Each comment was double-checked by a different member of our team. We systematically added commentary indicating the context in which the term was used, thus allowing us to identify the most widespread types of induced mistakes, which are listed below:

- The term does not refer to a language.
- The term is part of the name of an institution, project, grant, etc.
- The term is used as part of a general statement about linguistic diversity.
- A single hit was counted as two (e.g., 'severely endangered' counted also as 'endangered').
- The term is part of the title of a bibliographic reference.
- The term refers to a language different from the one of the study.
- The term appears in the free translation of a linguistic example.

Figure 2 features a screenshot of the coding after the curation process, and Table 2 illustrates some of the errors identified. A total number of 422 false positives were identified out of the 708 languages for which at least one term was automatically found.



	A	B	C	D	E	F	G	H	I	J	K
1	Language	ISO 639-3	Glottocode	Disappearing	Dying	Endangered	Highly Endangered	Moribund	Obsolescent	Severely Endangered	Vanishing
2	Shabo [sbf]	sbf	shab1252	False	False	True	True	True	False	True	False
3	Ulwa (Papua New Guinea) [	yla	yaul1241	False	False	True	False	True	False	True	False
4	Triv2 [tri]	tri	trio1238	False	True	False	False	True	False	False	False
5	Totela [ttl]	ttl	tote1238	False	False	True	True	False	False	False	False
6	Tirax [mme]	mme	maee1241	False	False	False	False	True	False	False	False
7	Tariana [tae]	tae	tari1256	False	False	True	True	False	True	True	False
8	Stau-Dgebshes [ero]	ero	horp1239	False	False	True	True	False	False	False	False
9	Seke (Vanuatu) [ske]	ske	seke1241	False	False	False	False	True	False	False	False
10	Nyulnyul [nyv]	nyv	nyul1247	False	False	False	True	True	True	False	False
11	Njerep [njr]	njr	njer1242	False	True	False	False	False	False	False	False
12	Nese [NOCODE_Nese]	NOCODE_Nese	nese1235	False	False	False	False	True	False	False	False
13	Naman [lzl]	lzl	litz1237	True	False	False	False	True	False	False	True
14	Nahavaq [sns]	sns	sout2857	False	False	False	False	False	False	False	False
15	Mwakai [mgt]	mgt	mong1344	False	False	False	False	True	True	True	False
16	Maragus [mrs]	mrs	mara1399	True	False	False	False	True	False	False	False
17	Mafea [mkv]	mkv	mafe1237	False	False	False	False	True	False	True	False
18	Leti (Indonesia) [lti]	lti	leti1246	False	False	False	False	False	True	False	False
19	Djingili [jig]	jig	djin1251	False	False	True	False	True	True	True	False
20	Chulym Turkic [clw]	clw	chul1246	False	False	True	False	True	False	False	False
21	Chakali [cli]	cli	chak1271	False	False	True	False	False	False	True	False
22	Bierebo [bnk]	bnk	bier1244	False	False	False	False	True	False	False	False
23	Xinca-Guazacapan [NOCODE	NOCODE_Xinca-G	xinc1246	False	False	True	False	True	False	False	False
24	West Yugur [ybe]	ybe	west2402	False	False	True	False	True	False	False	False
25	Wawa [www]	www	wawa1246	False	False	True	False	False	False	True	False
26	Warembori [wsa]	wsa	ware1253	False	False	False	True	True	False	False	False
27	Vera'a [vra]	vra	vera1241	True	False	True	False	False	False	False	False
28	Uru [ure]	ure	uruu1244	False	False	True	True	False	False	False	False
29	Ugong [ugo]	ugo	ugon1239	False	False	True	False	False	False	True	False
30	Thao [ssf]	ssf	thao1240	False	False	True	True	True	False	False	False
31	Taiap [gpn]	gpn	taia1239	False	False	True	False	False	False	False	False
32	Southern Nambikuv"ra [nab]	nab	sout2994	False	False	False	False	False	False	True	False
33	Southeast Ambrym [tvk]	tvk	sout2859	False	False	True	False	False	False	False	False
34	Skolt Sami [sms]	sms	skol1241	False	False	True	True	False	False	True	False
35	Shawi [cbt]	cbt	chay1248	False	False	False	False	True	True	False	False
36	Sabanv™ [sae]	sae	saba1268	False	False	True	False	True	False	False	False
37	Romagnol [rgn]	rgn	roma1328	False	False	True	False	True	False	False	False
38	Rayv2n Zoque [zor]	zor	rayo1235	False	False	True	False	False	False	True	False

Figure 2. Screenshot of the curated database: The red colour indicates true positives, and the green colour is used for false positives (value changed from 'TRUE' to 'FALSE'). Each false positive is accompanied by a comment.

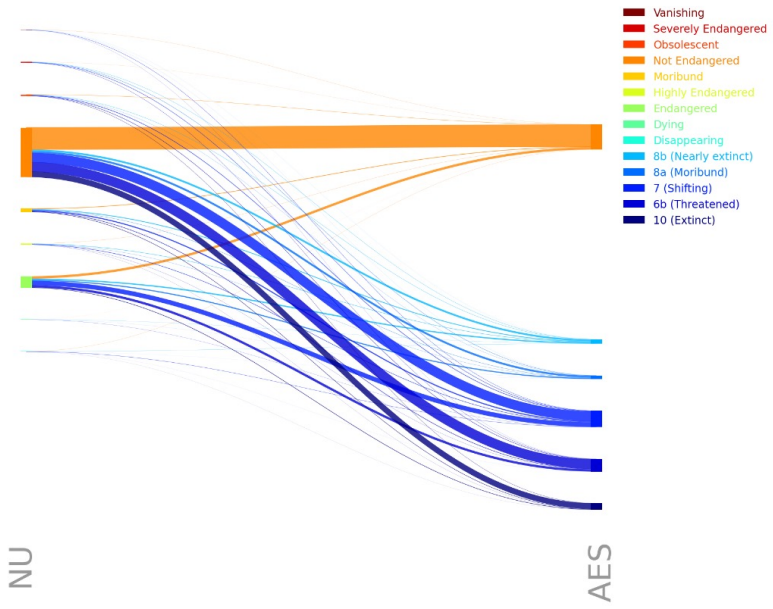
Table 2. Examples of false positives corrected through human data curation

Term	Language	ISO 639-3	Comments
Disappearing	Shabo	sbf	The references list includes an entry with the term in the title.
Disappearing	Western Sisaala	ssl	Moran's (2006: ii) grammar sketch includes the following dedication: "This work is dedicated to all who share in the pursuit of describing and preserving the world's disappearing languages."
Disappearing	Thavung	thm	The term appears in the text but does not refer to the language: "Malmkjaer (1991: 454), the integration of anthropology and linguistics in the tagmemic approach has provided invaluable documentation of many rapidly disappearing languages in remote regions."
Dying	Mape	mlh	The term appears in the text but refers to a different language: "This linguistic scenario has been observed elsewhere, for example, in the Gahuku speaking area of the Eastern Highlands Province, where the Gahuku language is dying rapidly as the younger generation are turning to Tok Pisin and English for their communicative needs (Tama 1994)."
Endangered	Mansim	–	The references list includes an entry with the term in the title.
Endangered	Manda-Matumba	mgs	The term appears in the text but refers to a different language: Vidunda.
Endangered	Kendayan	knx	The term appears as part of the name of the funding agency: "The Hans Rausing Endangered Language Documentation Project based at SOAS."

Term	Language	ISO 639-3	Comments
Highly endangered	Garig-Ilgar	ilg	The term appears in the text but refers to a different language: "Garig is also extremely close, in its phonology, grammar and lexicon, to Iwaidja, a highly endangered language, which today is spoken by perhaps 150 people."
Highly endangered	Farefare	gur	The term appears in the text but does not refer to the language: "traditional science, moral education, governance) enshrined in the language are highly endangered."
Moribund	Fuliiru	flr	The term appears in the text but does not refer to the language: "For example, Krauss (1992, Language 68(l):4-10) suggests that 50% of the languages currently spoken are 'moribund.'"
Moribund	Dime	dim	The term appears in the text but does not refer to the language: "without good quality documentation while the language is vital, […] later generations would have no hope of reviving a language once it is moribund or dead."
Obsolescent	Hup	jup	The references list includes an entry with the term in the title.
Obsolescent	Eyak	eya	The term is used to refer to a morphological marker.
Vanishing	Maskelynes	klv	The references list includes an entry with the term in the title.
Vanishing	Katso	kaf	The references list includes an entry with the term in the title.

Term	Language	ISO 639-3	Comments
Vanishing	Avava	tmb	The term appears in the text but refers to a different language: "One of the four, Naman: a vanishing language of Malakula (Vanuatu), had been submitted to Pacific Linguistics a couple of weeks earlier."

3.1 Terms and degrees of endangerment In order to evaluate the prospects for the automated extraction of the endangerment data, we first needed to better understand how terms in the literature actually relate to degree of endangerment – for example, whether *dying* reflects a higher degree of endangerment than *vanishing* or the generic *endangered*? By comparing the use of the relevant terms with the AES, we could calculate the degree of endangerment associated with each term on average. For each term across the three usages, we could check the AES for the languages with hits for the term. Further, if we measured the AES numerically from 0 (*not endangered*) to 5 (*extinct*), we obtained an average degree of endangerment for each term. Figure 3 shows alluvial diagrams of the terms versus AES associations, and Table 3 contains the corresponding statistics.



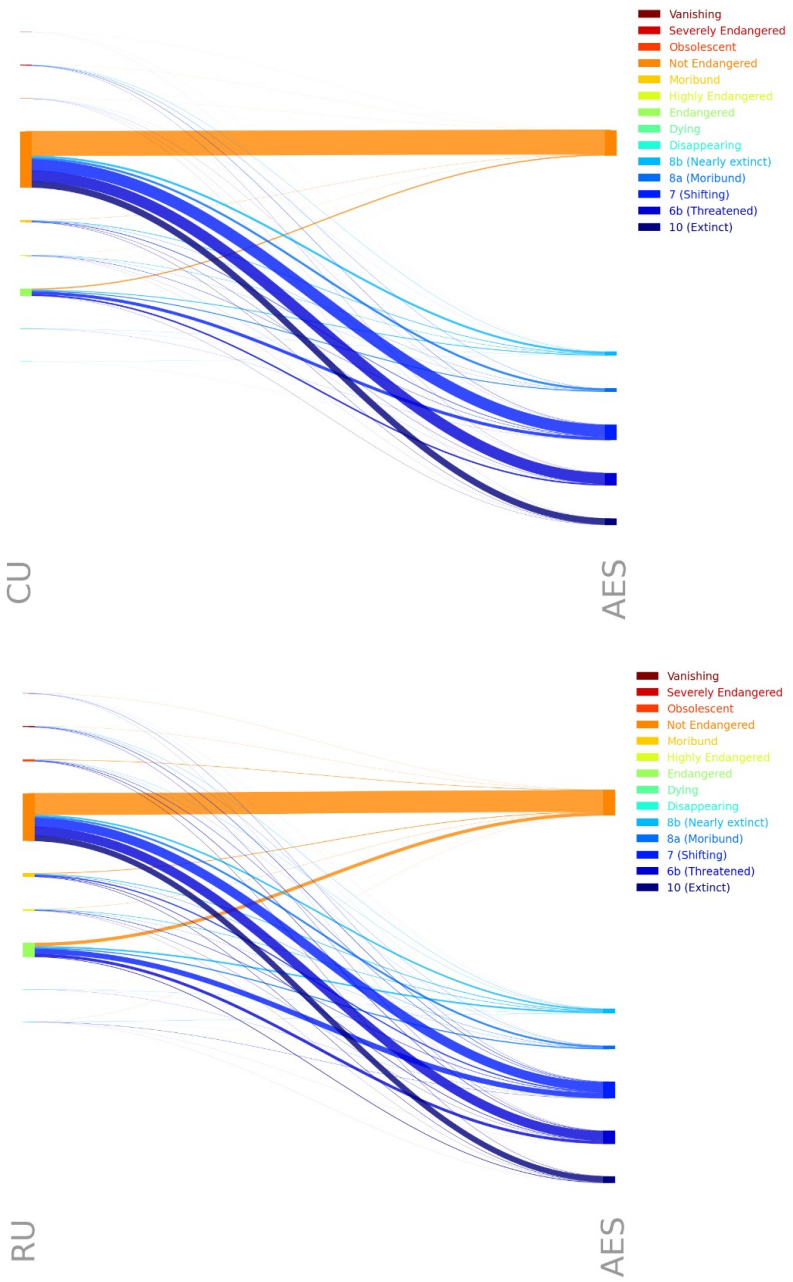


Figure 3. Alluvial diagrams showing the correspondence of search-term hits and AES labels across the three search-result usages: naive usage (*top*), corrected usage (*centre*), and most recent usage (*bottom*)

Table 3. Statistics on the correspondence of search-term hits and AES labels across the three search-result usages

Search-Result Usage	Term	Best AES Match			Degree of Endangerment
		AES Label	AES Score	No. lgs	
NU	*Not endangered*	Not endangered	0.44	(1102/2508)	1.37
	Vanishing	7 (Shifting)	0.48	(12/25)	1.68
	Endangered	7 (Shifting)	0.38	(222/582)	1.84
	Obsolescent	7 (Shifting)	0.29	(24/83)	2.20
	Severely endangered	7 (Shifting)	0.33	(18/54)	2.31
	Moribund	7 (Shifting)	0.30	(51/170)	2.38
	Disappearing	7 (Shifting)	0.44	(12/27)	2.30
	Highly endangered	7 (Shifting)	0.34	(21/61)	2.57
	Dying	7 (Shifting)	0.43	(6/14)	2.71
CU	*Not endangered*	Not endangered	0.43	(1176/2756)	1.39
	Vanishing	Not endangered	0.33	(1/3)	2.00
	Endangered	7 (Shifting)	0.40	(143/359)	1.90
	Obsolescent	7 (Shifting)	0.29	(5/17)	2.53
	Severely endangered	7 (Shifting)	0.33	(15/46)	2.46
	Moribund	8b (Nearly extinct)	0.29	(30/104)	2.75
	Disappearing	8b (Nearly extinct)	0.80	(4/5)	3.60
	Highly endangered	7 (Shifting)	0.35	(17/49)	2.76
	Dying	8b (Nearly extinct)	0.38	(5/13)	2.77
RU	*Not endangered*	Not endangered	0.44	(1053/2373)	1.37
	Vanishing	7 (Shifting)	0.46	(11/24)	1.67
	Endangered	7 (Shifting)	0.36	(251/702)	1.79

Search-Result Usage	Term	Best AES Match			Degree of Endangerment
		AES Label	AES Score	No. lgs	
	Obsolescent	6b (Threatened)	0.26	(24/94)	1.97
	Severely endangered	7 (Shifting)	0.38	(24/64)	2.19
	Moribund	7 (Shifting)	0.31	(59/188)	2.28
	Disappearing	7 (Shifting)	0.43	(12/28)	2.39
	Highly endangered	7 (Shifting)	0.31	(26/85)	2.55
	Dying	8b (Nearly extinct)	0.40	(6/15)	2.80

Note: AES = Agglomerated Endangerment Scale; CU = corrected usage; lgs = languages; NU = naive usage; RU = most recent usage.

The ranking of terms, save for anomalies relating to terms with few occurrences in CU, is relatively consistent. The rankings exhibited in NU and RU differ only in the placement of disappearing, where the stronger interpretation of this term appears to be preferable, judging from the numerical difference and the CU. An important finding revealed by comparing the different terminologies suggested by the database labels (see Table 1) is that *highly endangered* is used in the literature to a degree that surpasses *severely endangered* and *moribund*. We also found that several terms in the descriptive literature match the category labeled *shifting*. This result reinforces the impression that endangerment terminology is often informally used in grammatical descriptions.

3.2 Predicting individual endangerment With the degree-of-severity rankings for endangerment terms established in §3.1, when there are multiple-term hits, we can assess the status of a language described in a source as per the most severe term. We thus obtain a language-level endangerment status for all 3,214 languages covered for the three kinds of search-result usages, shown in Table 4.

Table 4. Language-level endangerment status for all 3,214 languages covered for the three kinds of search-result usages

Term	Naive (NU)	Corrected (CU)	Most Recent (RU)
Dying	14	13	15
Highly endangered	61	49	83
Disappearing	24	4	25
Moribund	137	87	152
Severely endangered	28	30	35
Obsolescent	58	11	66
Endangered	380	263	462
Vanishing	4	1	3
Not endangered	2,508	2,756	2,373
Total	3,214	3,214	3,214

Let us first consider the search results for NU versus their human correction, CU. The alluvial diagram in Figure 4 shows the language-level correspondence. Overall, the naive automatic assessment corresponds to its human correction by 91% (2,919 out of 3,214 languages), but this number is heavily dependent on the category of *not endangered* lacking hits. Of the languages with hits, only 411 out of 706 (58%) did not need some human adjustment. Hence, this appears to be the limit for individual accuracy of naively extracted positive hits. While far from the 91% accuracy considering all categories, 58% accuracy on eight levels is much better than random. As expected, and as can be gauged from Table 4, the majority of errors are "spurious" occurrences of the keywords (see Table 2 for some examples). Hence, the naive extraction tends to overestimate the amount of endangerment compared to a human reading.

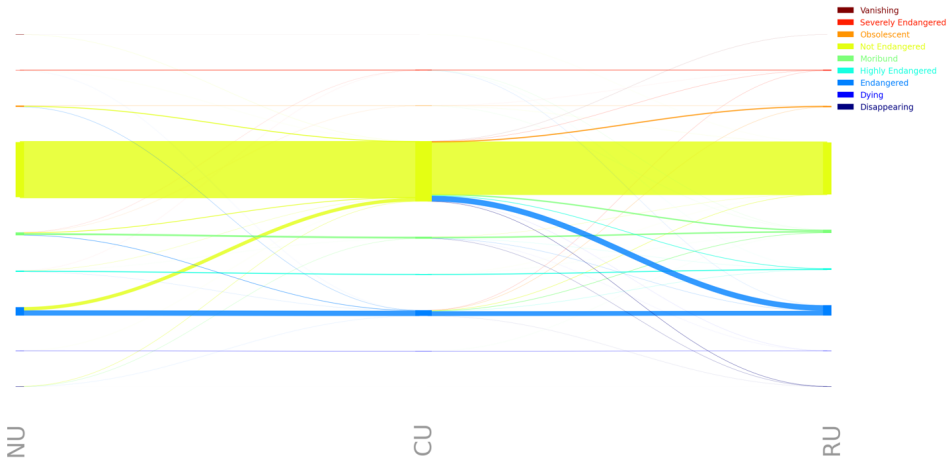
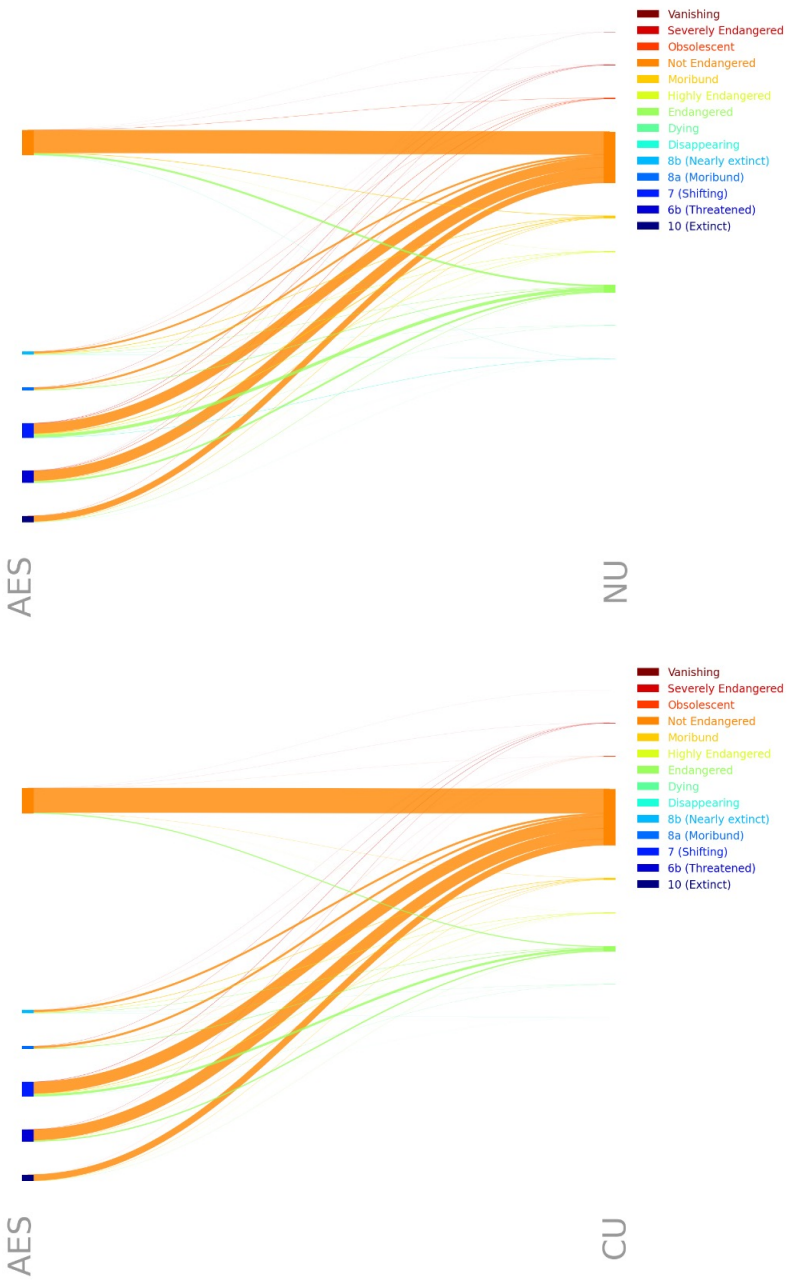


Figure 4. Language-level correspondences between naive usage (NU), corrected usage (CU), and most recent usage (RU)

The RU resembles the NU with an overall 92% correspondence (2,955 out of 3,214 languages; 74% on positive hits, or 626 out of 841 languages). However, RU differs in that it finds overall **more** endangerment and is less in agreement with CU, showing 84% overall correspondence (2,711 out of 3,214 languages; 43% on positive hits, 363 out of 841 languages). In other words, giving primacy to more recent descriptions finds more endangerment, as perhaps suspected, but does not resemble human reading more than that of the average for all available grammars (NU).

Now let us compare the results of automated searching to the individual AES data points. While the endangerment degrees of the search terms were established in §3.1, one further step remains to relate the per-language assessments (derived from the search terms) to AES assessments. The alluvial diagrams of Figure 5 show the per-language correspondences (which are not identical to those in Figure 3, where several search terms can be counted for each language), and Table 5 has the corresponding statistics.


Vanishing
Severely Endangered
Obsolescent
Not Endangered
Moribund
Highly Endangered
Endangered
Dying
Disappearing
8b (Nearly extinct)
8a (Moribund)
7 (Shifting)
6b (Threatened)
10 (Extinct)
AES
NU
Vanishing
Severely Endangered
Obsolescent
Not Endangered
Moribund
Highly Endangered
Endangered
Dying
Disappearing
8b (Nearly extinct)
8a (Moribund)
7 (Shifting)
6b (Threatened)
10 (Extinct)
AES
CU

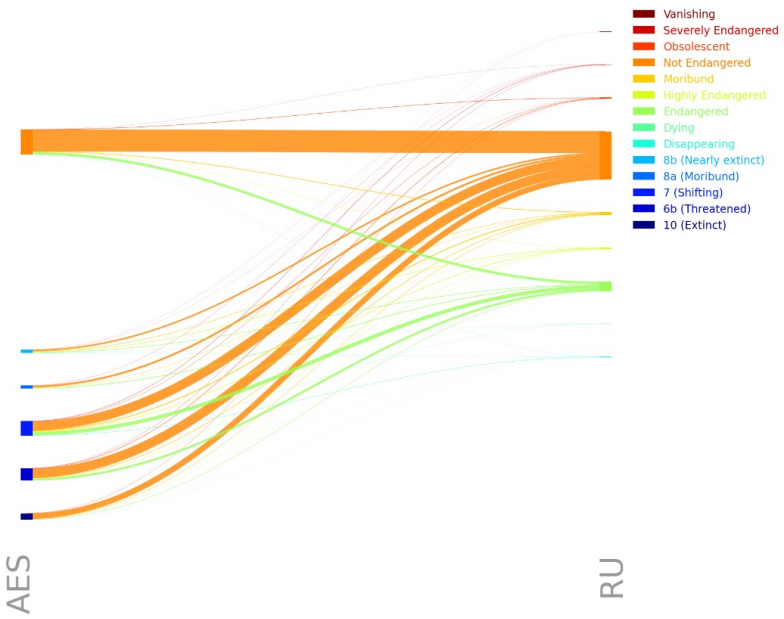


Figure 5. Alluvial diagrams showing the correspondence between the AES labels and language-level labels in the three search-result usages: naive usage (top), corrected usage (centre), and most recent usage (bottom)

Table 5. Statistics on the correspondence of automated language-level assessments and AES labels across the three search-result usages

	Naive Usage (NU)		
Term	Best AES Match		
Not endangered	Not endangered	0.44	(1102/2508)
Vanishing	6b (Threatened)	0.50	(2/4)
Endangered	7 (Shifting)	0.38	(146/380)
Obsolescent	7 (Shifting)	0.28	(16/58)
Severely endangered	7 (Shifting)	0.39	(11/28)
Moribund	7 (Shifting)	0.27	(37/137)
Disappearing	7 (Shifting)	0.50	(12/24)
Highly endangered	7 (Shifting)	0.34	(21/61)
Dying	7 (Shifting)	0.43	(6/14)
	Corrected Usage (CU)		
Term	Best AES Match		
Not endangered	Not endangered	0.43	(1176/2756)
Vanishing	Not endangered	1.00	(1/1)
Endangered	7 (Shifting)	0.40	(106/263)
Obsolescent	7 (Shifting)	0.27	(3/11)
Severely endangered	7 (Shifting)	0.40	(12/30)
Moribund	7 (Shifting)	0.28	(24/87)
Disappearing	8b (Nearly extinct)	0.75	(3/4)
Highly endangered	7 (Shifting)	0.35	(17/49)
Dying	8b (Nearly extinct)	0.38	(5/13)
	Most Recent Usage (RU)		
Term	Best AES Match		
Not endangered	Not endangered	0.44	(1053/2373)
Vanishing	6b (Threatened)	0.67	(2/3)
Endangered	7 (Shifting)	0.36	(168/462)
Obsolescent	Not endangered	0.30	(20/66)
Severely endangered	7 (Shifting)	0.37	(13/35)

Moribund	7 (Shifting)	0.29	(44/152)
Disappearing	7 (Shifting)	0.48	(12/25)
Highly endangered	7 (Shifting)	0.30	(25/83)
Dying	8b (Nearly extinct)	0.40	(6/15)

Note: AES = Agglomerated Endangerment Scale.

While better than random, it is difficult to predict a specific AES label with high precision. The overall precision for positive hits is around 35% (251 out of 706 languages, 36% [NU]; 171 out of 458 languages,37% [CU]; 290 out of 841 languages, 34% [RU]). Most individual term-based language assessments are only associated with a specific AES label 25–50% of the time (or occur too few times to be significant). The same pattern is found with little variation across the terms as well as across the different usages – including the human-corrected readings! We can thus conclude that AES endangerment assessments cannot be directly read off key terms in the literature, either by a human or a machine. This, again, reinforces the idea that endangerment labels are informally used in descriptive grammars.

The single best-matching AES label corresponding to most endangerment terms is *shifting*, even though the terms are known (from §3.1) to have different endangerment weight. If we are satisfied with predicting whether a language is endangered or not according to the AES (i.e., disregarding the degree), this can be done with around 80% precision (565 out of 706 languages, 80% [NU]; 391 out of 458 languages, 85% [CU]; 651 out of 841 languages, 77% [RU]). We note that the highest precision in the human-corrected case testifies to the validity of this labour.

So far, we have only discussed precision in automated extraction, and we now turn to the equally important question of recall. An observant reading of Table 5 reveals a recall level of around 0.44; in other words, of the large collection of languages where the search found no endangerment keywords, the AES predicted otherwise for slightly more than half. At first blush, this seems like an unexpectedly low recall, but the matter is more complicated. There are at least four categories of languages not categorized as *not endangered* by the AES:

Languages that are already extinct. As explained in §3, this set is not targeted in our study but would have been easy to exclude from consideration. There are 325 languages in our set that are extinct according to the AES and whose literature (consequently) does not contain endangerment keywords.

Languages that are endangered but whose endangerment is not discussed in the particular items of literature included in our collection. An important subset includes publications that discuss the endangerment of entire arrays of languages. These are not included in our study because it would not be straightforward to attribute the occurrence of a term in the document to a specific language or set of languages (even in cases where we know the list of languages are treated in the publication as a whole). However, such publications are found frequently as sources in the databases

underlying the AES.

Languages that are listed as endangered by the AES based on presumption rather than observation. Informal inspection suggests that many languages are presumed endangered based on their locations and general considerations rather than actual observations of broken transmission or lack of regular use. In such cases, we should actually not expect there to be a descriptive publication with endangerment observations.

Languages that are endangered but the descriptive publications explain this without using any of the key terms.

It is only the recall of the last category that is cause for concern and can be improved with an expanded, cleverer keyword search. The remaining categories have to be addressed by other means. Since the databases underlying the AES do not always contain a traceable source, an exact understanding of the size of each of the above categories is difficult to achieve. For the purpose of this study, we sampled ten languages at random for closer inspection. These are portrayed in Table 6. Although this is a small sample, it is clear that we should not take the AES information as "right" (if anything, the opposite) and the "low" recall of the automated searches as useless. Endangerment databases are often based on imprecise information, which sometimes may actually contradict fieldworkers' reports. Our experiment illustrates quite well how automated searches can in fact help endangerment databases in their struggle to check concrete information, keep information updated, and keep it consistent. This, of course, will work better if endangerment terminology is more systematically used in the linguistic literature.

A recent suggestion in the domain of Natural Language Processing is the introduction of *data statements* – a characterization of a dataset that provides context – in order to reduce the risk for unwarranted generalizations that may be harmful (Hovy & Spruit 2016; Bender & Friedman 2018). Possibly, such a scheme could be adopted for sociolinguistic descriptions to improve systematicity and clarity.

The sample evaluation also shows that there is room for real improvements in the automated-searching process. The range of terms considered can be expanded to include common phrasal expressions for endangerment such as "very few speakers" or "nearly extinct." It also seems worthwhile to attempt to automatically determine whether the source carries a substantial sociolinguistic section or not, as it seems a fair share of descriptions do not have a detailed sociolinguistic section that subsumes language transmission.

Table 6. Closer inspection of ten languages selected at random (languages considered endangered by the AES but not by automatic extraction)

Language	ISO 639-3	AES label	Inquiry
Aguaruna	agr	6b (Threatened)	AES derives from ELCat which cites Overall (2014). But Overall (2014) contains no information to sanction the status given in ELCat. Gramfinder searched Overall (2007), which indeed contains no vitality/endangerment information.
Sanumá	xsu	7 (Shifting)	AES derives from ELCat which cites Crevels (2012) who claims that Sanumá is "endangered" in Venezuela (Crevels 2012:189) and "potentially endangered" in Brazil (Crevels 2012:221). Gramfinder searched Borgman (1990) who gives no indications of actual or potential endangerment of the Sanumá in Brazil. In fact, Borgman (1990:17) describes the Sanumá as largely monolingual, as does the recent very detailed survey of Ferreira et al. (2019:34-38).
Mongghul	mjg	7 (Shifting)	AES derives from ELCat which cites Faehndrich (2007) which carries the observation. Gramfinder also searched Georg (2003:286-287) which only hints at endangerment (by saying that the number of speakers is considerably smaller than the ethnic group) and Üjiyedijn Chuluu (1994) which is silent on endangerment. The automatic search draws the correct conclusion from the most recent source but draws the wrong conclusion on majority vote.

Language	ISO 639-3	AES label	Inquiry
Lamkang	lmk	7 (Shifting)	AES derives from ELCat which cites Haokip (2011:96) who in turn cite Thounaojam and Chelliah (2007). Thounaojam and Chelliah (2007:2), however, only point to the possibility of endangerment: "Whether the number of speakers is closer to five or ten thousand, social, political and economic factors threaten Lamkang's longevity. Integration and competition for government jobs with the socioeconomically dominant Meitei, who populate the state capital Imphal in the central Manipur valley, point to a possible erosion of the linguistic situation of Lamkang."
Baré	bae	8b (Nearly extinct)	AES derives from ELCat which has the correct observation. Gramfinder searched Aikhenvald (1995:3-4) who says "there are very few speakers of this language left" rather than any of the specific endangerment terms.
Mandeali	mjl	7 (Shifting)	AES derives from UNESCO but UNESCO in turn carries no source. Gramfinder searched Ranganatha (1981) which has sociolinguistic information but says nothing to the effect that Mandeali would be endangered (Ranganatha 1981:15).
Hozo	hoz	7 (Shifting)	AES derives from ELCat which cites Lewis (2009) which in turn carries no source. Gramfinder searched Kassa (2014) which focused on morphosyntax and therefore carries no sociolinguistic information. Bahiru (2015:5) does consider Hoozo endangered due to the potential Oromo pressure, even though Hoozo is dominant in all generations.

Language	ISO 639-3	AES label	Inquiry
Gayo	gay	6b (Threatened)	AES derives from Eberhard et al. (2021) which carries no source. Gramfinder searched Eades (2005:6-8) who declares Gayo to be dominant in all ages but reports possible register loss.
Wara	wbf	7 (Shifting)	AES derives from ELCat which cites Lewis (2009) which in turn carries no source. Gramfinder searched Ouattara (2015:8-9) who says that other languages are used for interethnic communication but Wara is used in everyday Wara life in all age groups.
Old Japanese	ojp	10 (Extinct)	AES naturally marks Old Japanese as extinct and Gramfinder searches a number of publications, none of which exhibit endangerment terms.

Note: AES = Agglomerated Endangerment Scale; ELCat = Catalogue of Endangered Languages.

3.3 Predicting overall endangerment As we have seen, automated searches both overestimate endangerment (via "spurious" hits) and underestimate endangerment (via lack of sources and search-term poverty). It is useful to know to which extent these effects cancel out and what the net degree of over-/underestimation is, compared to the AES. If we again measure the AES numerically from 0 (*not endangered*) to 5 (*extinct*), the average AES score for all 3,214 languages considered in this study is 1.49. If we use the language-level labels of the automated searches and the degree ranking of terms established in §3.1 and map them to the same 0–5 scale, the average for all languages is lower: 0.49 (NU), 0.34 (CU), and 0.59 (RU). About half of the difference is due to already extinct languages. If we remove these languages from consideration, the AES average drops to 1.09 (while the computational estimates remain much the same, as expected). The full set of values, also broken down by macro-area, is shown in Table 7. It is also possible to use the term *degree values* from §3.1, which have been calculated to fit the AES. The global averages would match closely because of this "training" and would not be interesting to compare in the present study but could be used to get estimates of overall endangerment in new datasets.

In the sense considered above then, computational estimates on the whole underestimate endangerment compared to the AES. We expect some of this difference to be due to cautious older literature, as indicated by the higher estimate from the RU. From the results in §2, we also expect some of the difference to be due to projected, as opposed to observed, endangerment in the AES. Projected endangerment

may be as real and relevant as observed endangerment, but perhaps a terminological distinction is needed between projected (as practised in many overviews and databases) and observable endangerment (with a literature observation).
We recapitulate that the comparison here only concerns the set of languages for which there are grammars, sketches, or sociolinguistic studies. We are not in a position to assess AES data concerning languages for which no such literature is available.

Table 7. Average endangerment on a scale of 0 (*not endangered*) to 5 (*extinct*) according to the AES and the computational approaches. The numbers in parentheses show counts where extinct languages have been excluded.

Macro-area	No. lgs	AES	NU	CU	RU
Africa	777 (764)	0.57 (0.50)	0.38 (0.38)	0.23 (0.23)	0.43 (0.43)
North America	395 (291)	2.82 (2.04)	0.53 (0.61)	0.34 (0.41)	0.78 (0.90)
Papua	767 (755)	0.95 (0.88)	0.49 (0.49)	0.33 (0.33)	0.52 (0.52)
South America	191 (178)	2.01 (1.79)	0.95 (0.95)	0.61 (0.60)	0.98 (0.98)
Eurasia	900 (801)	1.56 (1.14)	0.47 (0.49)	0.34 (0.37)	0.60 (0.64)
Australia	184 (100)	3.83 (2.84)	0.56 (0.66)	0.40 (0.54)	0.63 (0.79)
Total	3,214 (2,889)	1.49 (1.09)	0.49 (0.51)	0.33 (0.34)	0.59 (0.60)

Note: AES = Agglomerated Endangerment Scale; CU = corrected usage; lgs = languages; NU = naive usage; RU = most recent usage.

4. Conclusion We investigated the usage of terms relating to endangerment in the descriptive literature. Our findings confirm that terms relating to endangerment found in a document describing a specific language, do indicate endangerment but also reveal that there are "spurious" usages and that the degree of endangerment is more difficult to assess, at least when compared to existing databases. These results show that endangerment terminology is informally used in descriptive literature, despite the fact that endangerment databases are becoming more popular and widespread. Almost half the languages considered endangered in existing databases do not exhibit such explicit terms in their descriptions. This difference is a combination of poverty of search phrases, gaps in explicit literature, and projections of endangerment in current databases. This is certainly a problematic aspect of such databases. Our explorations illustrate the potential for database curation assisted by computational searches to verify positive hits, stay up to date, as well as investigate languages without hits but expected otherwise. While projected endangerment may be as real and relevant as observed endangerment, a terminological distinction between pro-

jected (as practised in many overviews and databases) and observable endangerment (with a literature observation) would be a significant improvement to transparency.

From our findings, several issues require future work. As described above, the procedure for investigating full-text corpora for language endangerment started out with the most obvious search terms. This can be improved with some straightforward strategies, such as prescreening descriptive literature for a relevant section relating to endangerment. Another possibility, or perhaps necessity, is to consider a larger or a more complex cloud of search terms. In this endeavour, common techniques from information retrieval, such as *term frequency–inverse document frequency* (TF-IDF) (Manning et al. 2008) could easily be adopted to rank the results obtained. Yet another tool is to investigate *collocations* involving endangerment terms analogously. It also remains to be investigated whether the frequency and constellation of endangerment terms (rather than their mere existence) in a document can signal the degree of endangerment more robustly. Because of the general heterogeneity of the documents in the collection, this may be less feasible than what appears at first blush.

References

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

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

Aikhenvald, Alexandra. 1995. *Bare* (Languages of the World/Materials 100). Munich: Lincom.

Bahiru, Getachew Kassa. 2015. *A grammar of Hoozo*. Addis Ababa, Ethiopia: Addis Ababa University. (Doctoral dissertation.)

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

Bender, Emily M. & Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6. 587–604. doi:10.1162/tacl_a_00041

Borgman, Donald M. 1990. Sanuma. In Derbyshire, Desmond C. & Geoffrey K. Pullum (eds.), *Handbook of Amazonian languages*, vol. 2, 15–248. Berlin: Mouton de Gruyter.

Campbell, Lyle & Kenneth Rehg. 2018. Introduction: Endangered languages. 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 Ethnological Research* 5. 23–28.
Chuluu, Üjiyedijn. 1994. *Introduction, grammar, and sample sentences for Monguor*. Philadelphia: University of Pennsylvania Press.
Crevels, Mily. 2012. Language endangerment in South America: The clock is ticking. In Campbell, Lyle & Verónica Grondona (eds.), *The Indigenous languages of South America: A comprehensive guide* (The World of Linguistics 2), 167–233. Berlin: Mouton.
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.
Eades, Domenyk. 2005. *A grammar of Gayo: A language of Aceh, Sumatra* (Pacific Linguistics 567). Canberra: Research School of Pacific and Asian Studies, Australian National University.
Eberhard, David M., Gary F. Simons, & Charles D. Fennig. 2021. *Ethnologue: Languages of the world*, 24th edn. Dallas: SIL International.
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.
Faehndrich, Burgel R. M. 2007. *Sketch grammar of the Karlong variety of Mongghul, and dialectal survey of Mongghul*. Honolulu: University of Hawai'i at Mānoa. (Doctoral dissertation.)
Ferreira, Helder Perri, Ana Maria Antunes Machado, & Estevão Benfica Senra. 2019. *Línguas Yanomami no Brasil*. São Paulo: Instituto Socioambiental (ISA) and Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN).
Georg, Stefan. 2003. Mongghul. In Janhunen, Juha (ed.), *The Mongolic languages* (Routledge Language Family Series), 286–306. London: Routledge.
Grenoble, Lenore A. & Lindsay J. Whaley. 1998. *Endangered languages: Language loss and community response*. Cambridge: Cambridge University Press.
Hammarström, Harald. 2015. Ethnologue 16/17/18th editions: A comprehensive review. *Language* 91(3). 723–737.
Hammarström, Harald. 2021. *Gramfinder: Human and machine reading of grammatical descriptions of the languages of the world*. Trier, Germany: DBLP.
Hammarström, Harald, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A. Westenberg, & Bettina Speckmann. 2018. Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation* 12. 359–392. (http://hdl.handle.net/10125/24792)
Hammarström, Harald, Robert Forkel, Martin Haspelmath, & Sebastian Bank. 2021a. Glottolog 4.4. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://doi.org/10.5281/zenodo.4761960) (Accessed on 2021-05-20.)

Hammarström, Harald, Robert Forkel, Martin Haspelmath, & Sebastian Bank. 2021b. Glottolog 4.5. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://doi.org/10.5281/zenodo.5772642) (Accessed on 2021-12-10.)

Hammarström, Harald, One-Soon Her, & Marc Tang. 2021. Term-spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. In Ljunglöf, Peter, Simon Dobnik, & Richard Johansson (eds.), *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020)*, Gothenburg, 25–27 November 2020, 27–34. Linköping: Linköping University Electronic Press.

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

Haokip, Pauthang. 2011. The languages of Manipur: A case study of the Kuki-Chin languages. *Linguistics of the Tibeto-Burman Area* 34(1). 85–118.

Harrison, K. David. 2007. *When languages die: The extinction of the world's languages and the erosion of human knowledge* (Oxford Studies in Sociolinguistics). Oxford: Oxford University Press.

Hovy, Dirk & Shannon L. Spruit. 2016. The social impact of natural language processing. In Erk, Katrin & Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, 7–12 August, 591–598. Berlin: Association for Computational Linguistics. doi:10.18653/v1/P16-2096

Kassa, Getachew. 2014. *A brief grammar of the Hoozo language*. Asosa, Ethiopia: Mao-Komo Language Development Project.

Kibrik, Aleksandr E. 1991. The problem of endangered languages in the USSR. *Diogenes* 39(153). 67–83. doi:10.1177/039219219103915305

Krauss, Michael. 1992. The world's languages in crisis. *Language* 68(1). 1–10. doi:10.1353/lan.1992.0075

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

Lee, Changsoo. 2018. How are 'immigrant workers' represented in Korean news reporting? A text mining approach to critical discourse analysis. *Digital Scholarship in the Humanities* 34(1). 82–99. doi:10.1093/llc/fqy017

Lewis, Paul M. (ed.). 2009. *Ethnologue: Languages of the world*, 16th edn. Dallas: SIL International.

Manning, Christopher D., Prabhakar Raghavan, & Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Moran, Steven Paul. 2006. *A grammatical sketch of Isaalo (Western Sisaala)*. Ypsilanti: Eastern Michigan University. (Master's thesis.)

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

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

Ouattara, Virpi. 2015. *A phonological and tonal analysis of Samue using Optimality Theory*. Turku, Finland: University of Turku. (Doctoral dissertation.)

Overall, Simon E. 2007. *A grammar of Aguaruna*. Melbourne: La Trobe University. (Doctoral dissertation.)
Overall, Simon E. 2014. Clause chaining, switch reference and nominalisations in Aguaruna (Jivaroan). In van Gijn, Rik, Jeremy Hammond, Dejan Matiā, Saskia van Putten, & Ana Vilacy Galucio (eds.), *Information structure and reference tracking in complex sentences*, 309–340. Amsterdam: John Benjamins Publishing Company.
Ranganatha, M. R. 1981. *Mandeali*. India: Office of the Registrar General, Language Division.
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.
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 Ethnological Research* 5. 65–69.
Swadesh, Morris. 1960. Problems in language salvage for prehistory. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research* 3. 15–19.
Thomason, Sarah G. 2015. *Endangered languages: An introduction* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.
Thounaojam, Harimohon & Shobhana L. Chelliah. 2007. The Lamkang language: Grammatical sketch, texts and lexicon. *Linguistics of the Tibeto-Burman Area* 30(1). 1–189.
Virk, Shafqat Mumtaz, Harald Hammarström, Markus Forsberg, & Søren Wichmann. 2020. The DReaM Corpus: A multilingual annotated corpus of grammars for the world's languages. In Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, & Stelios Piperidis (eds.), *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, 11–16 May, 871–877. Marseille, France: European Language Resources Association.
Wurm, Stefan. 1956. Die dringendsten linguistischen Aufgaben in Neuguinea. In *Ethnologica, seconde partie et rapport général* (Actes du IVe Congrès International des Sciences Anthropologiques et Ethnologiques, Vienne, 1–8 September, 1952) vol. 3, 289–292. Wien: Adolf Holzhausens.
Wurm, Stephen A. 1991. Language death and disappearance: Causes and circumstances. *Diogenes* 39(153). 1–18. doi:10.1177/039219219103915302
Zaborski, Andrzej. 1970. Cushitic languages: An unexplored subcontinent. *Bulletin of the International Committee on Urgent Anthropological Ethnological Research* 12. 119–128.

Roberto Zariquiey
rzariquiey@pucp.edu.pe