Language endangerment, community size and typological rarity

Jan Wohlgemuth

1 Introduction

Publications on endangered languages frequently point out that endangered languages possess features or characteristics that are cross-linguistically rare or even unique. As Nettle and Romaine (2000: 11) put it:

> In fact, from the evidence we have to date, it would appear that the most grammatically complex and unusual languages are [...] often spoken by small tribes whose traditional way of life is under threat.

It is a truism that, if these languages become extinct, their rare features vanish with them, thus diminishing the diversity of human languages. While this loss in itself is already lamentable enough, it also has serious impact on the field of linguistics: If these languages are not documented, our impression of the range of possible human languages and possible variability of grammatical-typological parameters becomes irreparably skewed and narrow. This has been discussed e. g. by Dixon (1997: 116 *passim*) Hale (1998), Nettle and Romaine (2000: 11–12), Crystal (2000: 55, 64).

While it may at first seem surprising that the existence or absence of particular, cross-linguistically rare grammatical features in a language should somehow correlate with the degree of endangerment of that language, there seems to be at least slight evidence pointing into this direction. With a random distribution of rare features across all of the world's languages, one should expect these *rara* to be found in endangered languages at basically the same frequency as in non-endangered languages. It seems, however, that cross-linguistically "exotic" features are indeed to some extent more likely to be found in the former ones.

To my knowledge, this interrelationship has not been examined quantitatively yet and still warrants a plausible explanation. In this paper, I therefore approach the question as to whether endangered languages indeed are "rarer" or, looking at the issue from the opposite perspective, whether languages with unusual characteristics are in fact generally endangered or *more* endangered

than "average" languages. Lacking a more fine-tuned, comparative assessment of the world's languages with regard to their degree of endangerment, I will take their speaker community size as the decisive criterion.

2 Terminology and data basis

2.1 Features and characteristics

Since the present study mainly draws upon observations which themselves are based on data from *The World Atlas of Language Structures* (hereafter: *WALS*; Haspelmath et al. (eds.) 2005), it seems expedient to briefly introduce the terminology used therein at least as far as it is employed in this paper.

The 142 typological parameters analyzed in WALS are called *features*. One such feature is e. g. "Position of Tense-Aspect Affixes " (WALS chapter and Map 69; Dryer 2005).

For each feature, between 120 and 1,370 languages are given along with the information as to whether and / or how this feature exists in each observed language. This information is called *(feature) value*, and for above example, such values are e. g. "tense-aspect prefixes" or "no tense-aspect inflection".

I call this combined information on feature plus feature value for a single language a *characteristic*. The evaluation of rarity is based on the overall frequency of such characteristics in the entire WALS sample, as will be explained in Section 2.3.

2.2 Rarity

In accordance with Frerick's (2006: 10–15) criticism of Plank's (2000) only vaguely defined terminology and his inconsistent use thereof, I will apply the terms *rarum / rare* and *unicale / unique* to refer to grammatical characteristics found only in very few languages (*rara*) or one language (*unicalia*) respectively.

To be more precise, "found only in very few languages" shall, for the purposes of this paper, mean that the feature value in question is accounted for in less than five percent of the languages represented in WALS. To the extent that WALS can be considered an adequate, representative depiction of the world's linguistic diversity, one may consequently assume that the feature is also rare beyond the WALS sample.

Since it is not relevant for the study at hand, I will not systematically differentiate *rara* further between *rara*, *rarissima*, and *unicalia*. The terminology concerning rare linguistic features is discussed in further detail on pages 1–2 of Cysouw and Wohlgemuth (this volume), anyway.

2.3 Degree of rarity

I will use the *rarity index level* values calculated by Cysouw (2004, 2005, forthc.) on the basis of WALS as a measure for the cross-linguistic degree of "rareness", i. e. the absolute number of rare features found in a given language and their relative rarity in a cross-linguistic perspective. Cysouw's index has the advantage that it is unbiased and built upon a huge amount of typological information, as it is based on data from the extensive sample of languages used in WALS. This is much more objective than the mere impressionistic assumption that a quirky feature one finds in any particular language ought to be rather rare:

> The basic idea behind the rarity index is to compute the chance of occurrence for all the characteristics of a particular language, and then take the mean over all these chances. In essence, the lower this mean, the more rare characteristics this language has. (Cysouw forthc.)

A high *rarity index value* therefore basically means that the language has either a few extremely rare features or relatively many features that are at least moderately rare on a global scale. To normalize for distortion effects caused by the different number of characteristics coded for a language in WALS, Cysouw calculated a rarity index level by comparing the rarity index values with those of 1,000 fictitious languages per feature. For details on the simulation and the generation of the fictitious languages and feature values see Cysouw (forthc.)

A high *index level value* (given in percent) means that the (high) rarity index value is robust. This is to avoid the term *significant*, which would imply the result of a statistical analysis. See Cysouw (forthc.) for a discussion on why it is nevertheless very similar to a significance test result.

Cysouw (forthc.) computed separate rarity indexes for single languages, yielding an index of absolute rarity, and areal groups of languages, yielding an index of relative rarity. Unless explicitly stated otherwise, I will use the rarity index calculations for individual languages and thus discuss absolute *rara* in this paper only.

2.4 Endangerment and community size

There are numerous ways of classifying endangered languages as such and evaluating the degree(s) of their endangerment (e. g. Krauss 1992: 101–102; Wurm 1998: 192; Crystal 2000: 20–21; Grenoble and Whaley 1998: 24–25). Most of these classifications incorporate a multitude of factors which can have various grades of impact on the endangerment of a given language. For the purposes of this paper these classifications turn out to have one major drawback: While such a multitude reflects reality more accurately, a large number of factors makes it difficult to account for all of them in cross-linguistic comparison and in calculations like the ones done here.

Although I am fully aware of the pitfalls of determining the degree of endangerment simply through looking at the number of speakers, I chose that criterion as a proxy. As indicated above, it would have been impossible for me to retrieve and assess the necessary information on most, let alone all, of the proposed endangering factors for all of the 2,560 languages listed in WALS.

For these practical considerations, I decided to use primarily the classification as "nearly extinct" in Gordon (2005)[1] as the relevant criterion. This classification is based on a community size characterized as "only a few elderly speakers are still living" (Gordon 2005), which essentially means that *all* of these languages have less than 100, more often than not only a few dozen, fluent native speakers, occasionally only a handful or just one last speaker.

At any rate, community size itself has also been suggested as a relevant factor promoting the emergence of typologically rare features e. g. by Nettle (1999b: 138 and *passim*) using the example of object-initial word order:

> [. . .] one could predict that the rare, non-optimal orders would be more likely to be found in small communities than in large ones, since these would be more vulnerable to drift away from optimal states. (Nettle 1999b: 139).

This point is taken up again in Section 4.

As a matter of fact, not all endangered languages are necessarily actually "small" with respect to their community size — even languages with hundreds of thousands of speakers can be in a critical situation (cf. Crystal 2000: 13). Nevertheless, one can in good conscience assume that having only a very small community size normally means that these languages are endangered. This, then, brings us back to the question whether *rara* are more likely to be found in endangered languages.

2.5 Data basis

To check for the correlations between rarity and endangerment, I chose the 561 languages classified as "nearly extinct" from Gordon (2005). Of these languages, 152 are also featured in WALS and thus have a rarity index value calculated by Cysouw. These 152 languages constitute my sample of small, endangered languages. I refer to this sample as *the small languages*.

To have a control set of data, I chose from the top 550 languages with the most speakers (all languages with more than 2,000,000 speakers; numbers according to Gordon 2005) the first 152 which have a rarity index value calculated by Cysouw. I call this set *the big languages*.

The rarity index itself is based on 2,489 of the 2,560 languages from WALS, only excluding sign languages and a few other languages for the lack of (sufficient) comparable data (cf. Cysouw forthc.). In the following sections, I will nonetheless refer to the languages of this sample as *the / all WALS languages*.

For the sake of space, I will not list the names and rarity index values of all the languages in these three samples. However, an overview of the two 152-language-samples is given in the Appendix.

3 Rarity distributions across small and big languages

3.1 Statistical analysis

In order to determine whether there are differences in the distribution of "rare" languages between these samples, one first has to calculate the distribution within the three samples. Table 1 on the following page shows the results of these calculations.

The histogram in Figure 1 on the next page shows the distribution of the WALS languages across the rarity index level values, indicating that the WALS languages show all degrees of rarity. It is essentially a design feature of the rarity index that its median should be at exactly 50.0 and that all languages are distributed rather evenly across the entire range of the rarity scale. One can, however, observe that the distribution is slightly shifted towards the lower end of the rarity index scale, and the first quartile (Q_1), cutting off the lowest, i. e. first, 25 % of the data sample, is at 17.40 instead of the hypothetical 25.0 where it should be in an absolutely even distribution. Similarly, the

Table 1. Rarity index level distribution of the three samples

Sample:	WALS languages	small languages	big languages
number of languages	2,489	152	152
minimum	0.00	0.70	0.50
first quartile	17.40	41.22	19.60
median	46.00	67.35	53.45
third quartile	75.00	88.70	78.23
maximum	100	100	99.5
mean	47.29	61.16	49.75

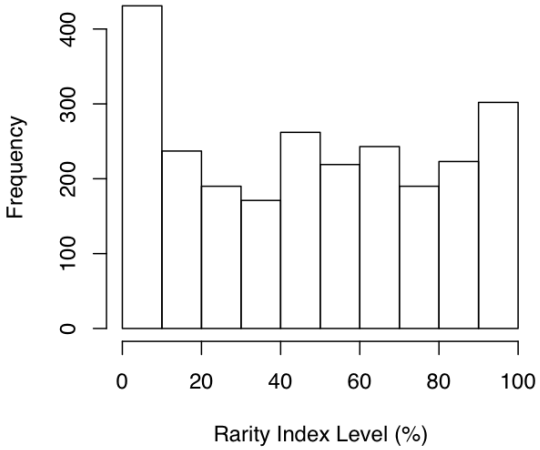

Figure 1. Distribution of the WALS languages across the rarity index level values

median (Q_2) with a value of 46 is only fairly close to the hypothetical 50, and only the third quartile (Q_3) is exactly at 75.0 where it should be by design.[2]

Compare this "overall distribution" to Figure 2 on the facing page, showing the distribution of the small languages over the rarity scale. It can clearly be seen that a relatively high number of the small languages show a higher degree of rarity with 35 (i. e. 23 %) of the languages having index level values in the top segment between 90 and 100. Accordingly, the median for this sample is rather high at 67.35.

The big languages of the control sample (cf. Figure 3 on the next page) are distributed as follows: Q_1 is at 19.60, which is rather close to the value

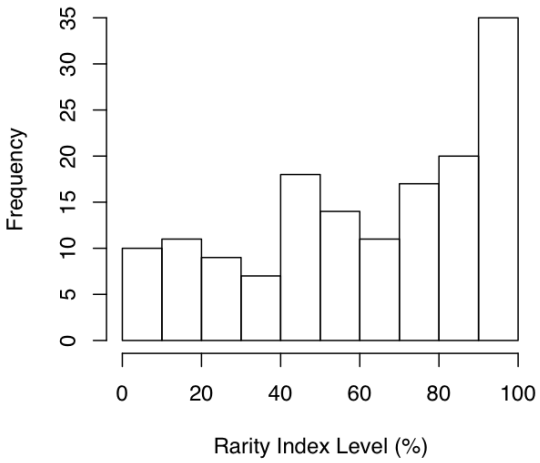

Figure 2. Distribution of the *small languages* across the rarity index level values

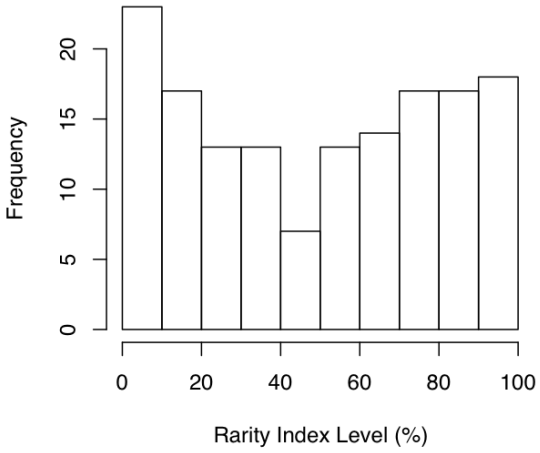

Figure 3. Distribution of the *big languages* across the rarity index level values

found in the WALS sample, the median is at 53.45, which is also fairly near to the design value of 50, and Q_3 at 78.23 is similarly near the WALS sample value. As can also be seen from the graph, the languages of this sample are thus distributed rather towards *both* ends of the rarity scale than to one end.

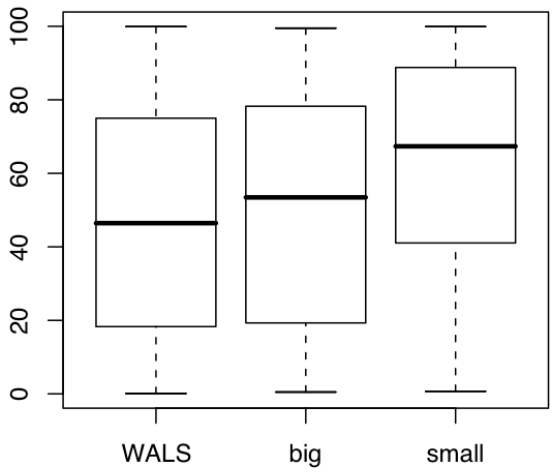

Figure 4. Comparison of the three samples' rarity index level distributions

In Figure 4, three box plots show the rarity index level distributions of all three samples in direct comparison. The "whiskers" and dotted lines show the total range of values, here 0 to 100 by design, whereas the lower and upper limits of the boxes indicate Q_1 and Q_3 respectively; the median (Q_2) is indicated by the thick horizontal line through the boxes. Comparing the three box plots for the three samples, one can identify two fairly obvious differences between the small languages and the two other samples: The small languages' box is notably shifted towards the upper end of the rarity scale and the distance between Q_1 and Q_3 is shorter compared to the more even distribution of all WALS languages and the big languages. Both of these groups appear to be very similar and are close to the normal distribution of values intended by Cysouw.

This divergence of the small languages, which can already be seen with the naked eye, is confirmed to be a significant one by means of a t-test which yields a value of $p = 2.293 \times 10^{-7}$ for the WALS sample vs. the small languages, cf. Table 2 on the next page.

This result proves that the observable difference is truly a significant one. One cannot avoid the conclusion that the small languages of our sample actually *do* have more cross-linguistically rare features or – in other words –

Table 2. T-Test results compared

sample pairs	p-value	significance
WALS > small	0.000,000,229,3	very high
WALS > big	0.442,2	none
small > big	0.001,434	moderate

that there *is* a significantly higher likelihood to find small (endangered) languages in the upper end of the rarity scale.

3.2 A heterogeneous picture

Yet, the whole picture is not as simple as the last paragraph of the preceding subsection could make believe. Despite the significant shift towards the rarer end of the scale, one does find small languages across the entire range of rarity index values, and the languages with comparably rare features display considerable variation of speaker community sizes. To illustrate this, Table 3 on the following page shows the top and bottom 15 languages of the rarity index scale, which is based on the results of Cysouw's (2005, forthc.) calculations of the mean rarity index and index level values for the WALS languages. The data in the table is augmented by the speaker numbers from Gordon (2005). The languages are sorted first by descending index level values and second by ascending mean rarity index values.

Looking at these results and interpreting them, one has to bear a few caveats in mind. First, the sample of small languages and the control sample just alike are both rather small and thus much less representative than the WALS sample or the original collection of small languages from Gordon (2005). The 152 languages each account for only 6.1 % of the WALS languages and 27.1 % of the endangered languages listed in Gordon (2005). This discrepancy is due to the fact that the members of the small languages sample were selected only by one criterion, namely whether the languages are in WALS and hence have a rarity index value available.

This point leads to the second problem: The complete data set is likely distorted because the WALS sample itself already includes some small languages only *because* of their odd characteristics, which then are coded in WALS, while other, more "ordinary" features of such languages often do not appear in WALS. This may also be connected with the following point inas-

Table 3. Top and bottom 15 languages (mean rarity index level) and their size

Rank	Language (Genus)	Features in WALS	Mean Rarity Index	Index Level (%)	Speakers
1	Wari' (Chapacura-Wanhan)	115	2.36	100	5
2	Dinka (Nilotic)	45	3.45	100	320,000
3	Jamul Tiipay (Yuman)	44	3.76	100	220
4	Nuer (Nilotic)	28	3.42	100	804,000
5	Karó (Arára) (Tupi-Guarani)	24	6.16	100	150
6	Winnebago (Siouan)	7	11.37	100	230
7	Chalcatongo Mixtec (Mixtecan)	113	2.05	99.9	15,000
8	Kutenai (Kutenai)	113	2.02	99.9	12
9	Kombai (Awju-Dumut)	38	3.27	99.9	4,000
10	Dahalo (Southern Cushitic)	17	5.86	99.9	< 400
11	Maxakali (Maxakali)	15	6.95	99.9	728
12	Warrwa (Nyulnyulan)	20	3.74	99.8	2
13	Bunuba (Bunuban)	16	4.21	99.8	< 100
14	Eyak (Eyak)[3]	16	4.05	99.8	(*1*)
15	Yawuru (Nyulnyulan)	15	4.51	99.8	30
⋮	⋮	⋮	⋮	⋮	⋮
2,474	Kalam (Madang)	19	0.50	0.1	15,000
2,475	Guhu-Samane (Binanderean)	12	0.42	0.1	12,761
2,476	Shira Yughur (Mongolic)	5	0.31	0.1	3,000
2,477	Mawng (Iwaidjan)	106	0.70	0.0	200
2,478	Bagirmi (Bongo-Bagirmi)	106	0.69	0.0	44,761
2,479	Khasi (Khasian)	102	0.68	0.0	865,000
2,480	Brahui (Northern Dravidian)	93	0.67	0.0	2,000,000
2,481	Daga (Dagan)	91	0.64	0.0	6,000
2,482	West Makian (North Halmaheran)	48	0.57	0.0	12,000
2,483	Kaliai-Kove (Oceanic)	42	0.52	0.0	6,750
2,484	Selepet (Finisterre-Huon)	36	0.49	0.0	7,000
2,485	Ndut (Northern Atlantic)	34	0.55	0.0	35,000
2,486	Cornish (Celtic)[4]	32	0.52	0.0	(*500*)
2,487	Tulu (Southern Dravidian)	29	0.51	0.0	1,949,000
2,488	Sougb (East Bird's Head)	18	0.44	0.0	12,000
2,489	Bisa (Eastern Mande)	15	0.46	0.0	371,000

much as there is not always information available about the "average" features of small languages.

Furthermore, the WALS data underlying this study could also be skewed because the scholarly papers on small languages some of the WALS data is based on is biased. The authors of such papers tend to emphasize cross-linguistic peculiarities for various reasons. One of them is to point out the need to do more extensive research on that language, another one is the aim to underscore differences with neighboring languages in order to establish it as a separate language, or simply to make the language more "attractive" or interesting.

These limitations notwithstanding, the difference in the mean rarity value and rarity index level distributions is significant and calls for an explanation, as do some particular facts: there is not only a substantial amount of unendangered languages with rare characteristics but also a number of endangered languages without rare characteristics.

3.2.1 Unendangered languages with rare characteristics

As could already be seen in Table 3, not all of the "rarest" individual languages are endangered. The first ("most exotic" or "rarest") one – Wari' – definitely is endangered, and so are many others of the highest ranking languages in Cysouw's rarity index level list. On the other hand, languages like Dinka and Nuer, ranking similarly high in the rarity index level list, each have hundreds of thousands of speakers and are not acutely endangered.

Table 4. Cluster of three very large languages in the top 100 by rarity index level

Rank	Language (Genus)	Features in WALS	Mean Rarity Index	Index Level (%)	Speakers
60	Mandarin (Chinese)	130	1.55	98.3	940,856,000
66	German (Germanic)	129	1.40	98.0	92,113,000
69	Cantonese (Chinese)	76	1.58	98.0	59,570,000

Furthermore, one also finds some of the largest languages of the world within the top 100 languages of the rarity index level list, cf. e. g. the ones given in Table 4. The fact that one finds several such large languages ranking

high in the rarity index level list prohibits any generalization along the lines that rara would only or predominantly be found in small languages.

3.2.2 Endangered/small languages without rare characteristics

A similar picture is found at the other – lower – end of the scale as it is shown in the lower half of Table 3. Among the languages with the lowest rarity index level values there are also some severely endangered languages like Mawng or Cornish right next to reasonably safe languages as Tulu or Brahui. With respect to the generalizations on endangerment and rarity this means that *not all* endangered languages possess rare characteristics.

As became obvious from the data given in Section 3.2, a substantial number of the small languages is found in the upper quarter of the rarity scale. Nevertheless, the languages of this sample are distributed over the whole scale. This basically means that at least *some* of the small languages actually appear to be very "un-unusual" in cross-linguistic comparison.

This finding, of course, must not be misconstrued as a statement that such small "average" languages were less worthy of description or that their documentation was of minor relevance or had a lower priority. Quite to the contrary, documentation and description is, of course, the prerequisite to *any* analysis that then reveals the typological makeup of a language and thereby allows the detection of rare characteristics. One cannot know beforehand whether a small language contains *rara* or not. But – as the the data presented here show – one has some reason to expect it does.

3.2.3 Interim summary

In summary, the analysis given in this section shows that there is a significantly higher chance that a given small – and hence usually endangered – language has cross-linguistically rare or unique features.

There is, however, no incontrovertible evidence for a direct correlation of language endangerment and rarity or a solid implication in either direction, as the distributions shown above also give ample counter-evidence. The statement in the previous paragraph is therefore not an unconditional correlation in either direction but rather an implication based on an increased likelihood.

The explanation of the findings presented here therefore boils down to this basic problem: of what nature is the relationship between the degree of

a language's endangerment and the presence of rare or unique grammatical features or characteristics in it? In other words: Are the significant differences a sign of a (weak) correlation in one or the other direction between these two factors, or are they a case of covariation and both dependent on a third, different, factor, namely the size of the speaker community?

4 Looking for an explanation

4.1 Community size

Trudgill (2004: 318) referring to Nettle (1999b: 147) points out that small speaker community size favors the development of unusual phonological systems. Taking up this point and applying it to all aspects of a language, one can then argue that small communities also might be more apt to develop and / or maintain unusual grammatical characteristics in general.

Nettle (1999a, 1999b, 1999c) has demonstrated by means of computer simulations that in languages with very small speaker communities of under 400 speakers "structures against which there is a bias in acquisition can evolve and persist for more of the time than in large ones" (Nettle 1999a: 129). This is the case because small community size makes a language more susceptible to language change, even if that change involves the innovation and diffusion typologically "unexpected" or "marked" characteristics:

> "If a group consists of just a few hundred people, the idiosyncrasies of one influential individual can spread through it very easily. This is not the case if the group consists of thousands or tens of thousands of people. In general, the smaller the community, the greater the probability that a given variant that has no functional advantage at all but is neutral or slightly disadvantageous, can replace the existing item and become the norm." (Nettle 1999b: 139)

This explains why typologically unique or rare innovations generally seem to appear more frequently in small languages. The question whether a characteristic's rarity always means that it is "marked" or has "no functional advantage at all" must remain open here. Judging from the *rara* discussed in the present volume and in Cysouw and Wohlgemuth (eds., 2010), though, I would object to the generalization that all of them were necessarily "neutral or slightly disadvantageous" in their nature.

Evidence for Nettle's explanation cited above can be found in Kulick (1992: 2 *passim*), who mentions several case studies from Papua New Guinea

of deliberate manipulation of a languages' structure in order to distinguish it from neighboring languages by means of idiosyncratic characteristics. Such deliberate changes could more easily diffuse to become a common standard within a smaller community.

Another factor to be taken into consideration is that "large" languages, even while having considerable internal variation, often tend to have one "normalized" variant which is also learned as a foreign language by many (adult) speakers of (small) minority languages and thus more likely subject to simplification than small languages that are not learned by outsiders.

These factors would already go a long way to explain a co-dependency of both, language endangerment and typological rarity, on a third factor, namely community size — which is exactly the factor used in the calculations here.

4.2 Enclave situations

Bickel (2006) adds to this that the trend to "normalization" under contact with other, normally larger, languages of a less rare typological profile can only be avoided in so-called enclave situations (cf. Bickel and Nichols 2003: 30) where they may remain more or less unaffected by majority language influence and effects of globalization. Such speech communities can be rather small but need not necessarily be below the "critical mass" threshold of being severely endangered.

Enclave situations can also explain the fact that some non-endangered languages, regardless of size, contain rare characteristics which did not spread into neighboring small languages if these are in a type of location that Bickel and Nichols (2003: 30) call "preservation enclaves". These are situations where the relative isolation of their speaker community allows these languages to maintain their ("usual") typological profile because they are not under immediate pressure from the (bigger, *rara*-containing) language. The larger languages' *rara* thus also stay rare because they do not diffuse into other languages which would render them more frequent.

This view is supported by the dialectological study of Andersen (1988), who supposes that

> "there is a connection between the limited social-spatial function of a dialect, its relative closedness, and its ability to sustain exorbitant phonetic developments" (Andersen 1988: 70).

4.3 Endangerment and *rara*

Community size is thus probably not the sole decisive factor in language death, as can be seen e. g. from languages which have only comparatively few speakers but are nonetheless rather stable while other languages are endangered despite their comparatively large speaker community. Similarly, having a high rarity index value does not necessarily imply either endangerment or a relatively small number of speakers, as can be seen from some of the larger languages in Tables 3 and 4, where one also finds languages with rare characteristics.

One thus has to differentiate the generalization mentioned in the beginning and keep the notions of endangerment and rarity separate: Neither do all or most endangered languages possess typologically unusual features, nor are all languages with rare features endangered.

Furthermore, *rara* themselves can be endangered – independent from the endangerment or safety of the language they occur in – by various other extra-linguistic factors, among them globalization and global standardization. These factors can endanger rare features or characteristics cross-linguistically and in a particular language without endangering the whole language (cf. Wohlgemuth and Köpl 2005).

An example involving *rara* discussed in the present volume is the introduction and spread of decimal (base-10) numeral systems may already have caused the demise of unique and rare numeral systems in at least some regions of the world. It is quite evident that some of the rare(r) numeral systems were replaced as a consequence of strong cultural pressure (cf. Comrie 2005; and the remarks by Hammarström (this volume): 28, 32).

This kind of scenario, too, can explain why even in regions with a high degree of genealogical diversity and lots of small languages not as many *rara* are found as one might have expected.

Being small *and* having *rara* (which to a certain extent actually seems to be favored by small speaker communities) can mean that there is a higher probability that the language in question is endangered. Claiming, however, that endangered languages *per se* are "rarer" than average appears like inappropriately turning the causality on its head. From being endangered, languages do not come to have rare characteristics they did not have before being endangered. If the endangerment situation has a direct impact at all, these languages rather tend to *lose* their "exotic" features during phases of attrition,i. e. when they are being assimilated by a larger majority language.

Normally, *rara* already exist in the language before it becomes endangered. If rare features actually arise in conjunction with the language becoming endangered, it is rather the small(er) size of the speaker community that can favor the spread of innovative *rara*.

4.4 A multiplicity of factors

The exact mechanisms that are involved in the possible impact of extra-linguistic factors on the emergence, maintenance, frequency, or diffusion of linguistic (i. e. grammatical) parameters still need to be explored. Such factors besides community size are, among others, location (or degree of geographical isolation) of the speaker community (cf. the enclave situations mentioned above), cultural factors and community structure (cf. e. g. the study of Güldemann et al. (in prep.) on the (historical) linguistics of hunter-gatherer languages) or different language contact scenarios and situations (cf. e. g. Kelkar-Stephan (2010) for an account of how a *rarum* emerged owing to particular circumstances of language contact).

Furthermore, any generalizations concerning the interaction of linguistic and extra-linguistic factors need to be put on an empirical basis to be use- and meaningful. This has already been discussed by Nettle (1999b: 138–141), but is nonetheless still true a decade later, as I pointed out in a different context:

> Yet, so far there is no such linguistic discipline as sociolinguistic typology [. . .] This means that there is no solid basis for the cross-linguistic evaluation and comparative classification of sociolinguistic settings and contact scenarios and the different parameters defining their nature. These, however, are indispensable prerequisites to test for correlations of these extra-linguistic parameters with linguistic facts and factors of the languages involved. (Wohlgemuth 2009: 298–299).

This also applies to the study and explanation of the cross-linguistic distribution of rare and unusual typological characteristics and the extra-linguistic factors having an impact on them.

5 Conclusions

The question as to whether there is a direct correlation between the degree of endangerment and the rarity or uniqueness of a language could not clearly be answered. There is no incontrovertible evidence for a direct and unconditional

correlation of these two parameters. What can be observed, though, is rather the covariation of both factors depending on another factor — the size of the speaker community.

Yet, other extra-linguistic factors must be taken into account more systematically to explain the endangerment of a language and the emergence and / or retention of rare linguistic features.

In summary it can be said that there are significant differences between the rarity index distributions of small languages versus the huge sample of WALS languages. However, lacking comparative data on extra-linguistic factors in a similar fashion as the typological data of WALS, one cannot establish direct correlations other than the rather vague implication that rare characteristics can be found "with clearly more than chance frequency" in languages which have a small speaker community and thus very likely are endangered.

Acknowledgments

This paper is based on two presentations I gave in Leipzig in February and March 2006 and on a poster presented during the third Oxford-Kōbe seminar on the linguistics of endangered languages in April 2006. I am grateful to the audience of these presentations, two anonymous referees, and to Eric Holman and Bernard Comrie for their feedback that had an impact on the evolution of this paper's focus and direction. I am particularly thankful to Michael Cysouw for kindly letting me use his data and rarity index calculations and for also refreshing my memory on statistical methods and terms, and to Hans-Jörg Bibiko for helping me with the map in Figure 5. Of course, these people must not be blamed for any errors, shortcomings or misinterpretations in this paper, which are entirely my own.

Appendix

Listed below are the languages in the two samples of the "smallest" and "biggest" WALS languages having a rarity index value as discussed in Section 2.5. Language names basically follow the form used in WALS or Ethnologue, however I did not append preposed adjectival parts but rather left them in front of the (proper) name, e. g. *Central Pomo*, not *Pomo, Central*.

It is partly an artifact of sampling (e. g. the unavailability of data on African languages[5] or the mean community size over the threshold of 300

speakers) that there are very few languages from Eurasia and Africa represented. The overall geographical distribution of the small languages sample can be seen from the map in Figure 5 on the facing page.

A final caveat: Like Eyak (cf. note 3 on page 274), some of the "small" languages may in fact already be extinct even though they still had (few) speakers listed in Gordon (2005).

The sample of 152 "small languages"

Ainu, Achumawi, Ahtena, Alawa, Angosturas Tunebo, A-Pucikwar, Arabana, Atsugewi, Atzingo Matlatzinca, Baadi, Bädi Kanum, Badimaya, Baré, Baure, Berbice Creole Dutch, Biri, Boruca, Cahuilla, Catawba, Central Pomo, Central Sierra Miwok, Chinook, Cholon, Clallam, Coast Miwok, Coeur d'Alene, Coos, Cupeño, Darling, Dhargari, Djingili, Dyaabugay, Dyirbal, Eyak, Gagadu, Ganggalida, Gunya, Hupa, Itonama, Itzá, Jabutí, Kalapuya, Kalispel-Pend d'Oreille, Kamilaroi, Kamu, Karadjeri, Kashaya, Kato, Kawaiisu, Kerek, Kiliwa, Klamath-Modoc, Kokata, Koyukon, Kumbainggar, Kuwama, Lake Miwok, Lamu-Lamu, Laragia, Lardil, Limilngan, Lushootseed, Madngele, Mandan, Mapoyo, Mara, Maranunggu, Margany, Martuyhunira, Menomini, Miriwung, Mogholi, Mono, Movima, Mullukmulluk, Munsee, Muruwari, Ngadjunmaya, Ngalakan, Ngawun, Ngura, Nisenan, Northeast Maidu, Northern Haida, Northern Sierra Miwok, Northwest Maidu, Nyawaygi, Nyulnyul, Omagua, Ona, Osage, Pakanha, Panamint, Paulohi, Pawnee, Pipil, Pitta Pitta, Plains Miwok, Principense, Puelche, Quileute, Rama, Resígaro, Serrano, Shasta, Sirenik Yupik, Southeastern Pomo, Southern Haida, Southern Puget Sound Salish, Southern Sierra Miwok, Squamish, Tanaina, Tariano, Taushiro, Tehuelche, Thao, Thaypan, Tübatulabal, Tuscarora, Tyaraity, Udihe, Unami, Upper Chehalis, Ura, Uradhi, Uru, Vod, Wadjiginy, Wambaya, Wangaaybuwan-Ngiyambaa, Wappo, Waray, Warluwara, Warrgamay, Warungu, Wasco-Wishram, Washo, Western Abnaki, Western Yiddish, Wichita, Wintu, Wirangu, Yámana, Yidiny, Yinggarda, Yir Yoront, Yokuts, Yuchi, Yugh, Yuki, Yurok, Záparo

The sample of 152 "big languages"

Afrikaans, Akan, Albanian, Alemannic, Algerian Spoken Arabic, Amharic, Armenian, Assamese, Awadhi, Balochi, Belarusan, Bengali, Bhojpuri, Bokmaal Norwegian, Bosnian, Bugis, Bulgarian, Bundeli, Burmese, Catalan-Valencian-Balear, Cebuano, Central Khmer, Chhattisgarhi, Chittagonian, Croatian, Czech, Danish, Dec-

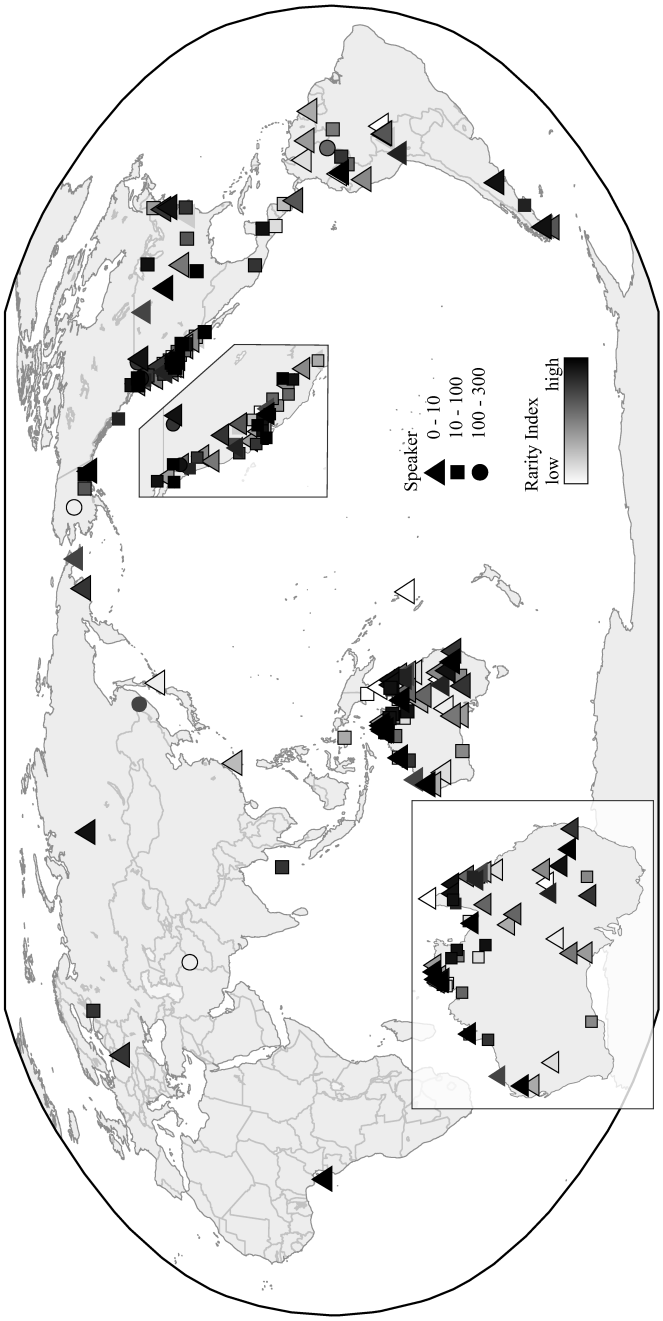

Figure 5. The "small languages" sample and its global distribution

can, Dutch, Eastern Farsi, Eastern Oromo, Eastern Panjabi, Egyptian Spoken Arabic, English, Finnish, French, Fulfulde, Gan Chinese, Georgian, German , Gikuyu, Greek, Gujarati, Haitian Creole French, Hakka Chinese, Haryanvi, Hausa, Hebrew, Hijazi Spoken Arabic, Hiligaynon, Hindi, Hungarian, Igbo, Ilocano, Indonesian, Italian, Japanese, Javanese, Jinyu Chinese, Jula, Kanauji, Kannada, Kashmiri, Kazakh, Kituba, Korean, Krio, Kurdi, Kurmanji, Libyan Spoken Arabic, Lingala, Lombard, Luba-Kasai, Luri (Lri), Madura, Magahi, Maithili, Malagasy, Malay, Malayalam, Marwari, Mesopotamian Spoken Arabic, Min Bei Chinese, Min Nan Chinese, Minangkabau, Mòoré, Moroccan Spoken Arabic, Najdi Spoken Arabic, Napoletano-Calabrese, Nepali, North Azerbaijani, North Levantine Spoken Arabic, North Mesopotamian Spoken Arabic, Northeastern Thai, Northern Thai, Northern Zhuang, Nyanja, Oriya, Paraguayan Guaraní, Polish, Portuguese, Romanian, Rundi, Russian, Rwanda, Sa'idi Spoken Arabic, Sanaani Spoken Arabic, Santali, Serbian, Shona, Sicilian, Sindhi, Sinhala, Slovak, Somali, South Azerbaijani, South Levantine Spoken Arabic, Southern Sotho, Southern Thai, Spanish, Sudanese Spoken Arabic, Sukuma, Sunda, Swahili, Swedish, Sylhetti, Tagalog, Ta'izzi-Adeni Arabic, Tajiki, Tamil, Tatar, Telugu, Thai, Tigrigna, Tunisian Spoken Arabic, Turkish, Turkmen, Ukrainian, Umbundu, Urdu, Vietnamese, West-Central Oromo, Western Egyptian Bedawi Spoken Arabic, Western Farsi, Western Panjabi, Wu Chinese, Xhosa, Xiang Chinese, Yoruba, Yue Chinese, Zulu

Notes

1. For an updated list, see http://www.ethnologue.org/nearly_extinct.asp
2. The exact reason(s) why this shift towards the lower end occurs is a mathematical problem of the index calculation which has yet to be solved (Cysouw, p. c.), but since the actual values rather than the hypothetical ones will be the basis for comparison in this paper, this deviation can be disregarded here.
3. As a matter of fact, Eyak became extinct in January 2008; the calculations of this paper were, however, done in 2006 and are based on data from 2005.
4. Cornish became extinct in 1777, but is being revived (cf. Gordon 2005), there is a chance that the "new" version of the language is more "normal" in terms of the rarity index value, as there has been a long break of transmission and an unknown amount of information has probably been lost.
5. The sample of small languages contains very few languages from Africa. Apparently, only one of the endangered languages mentioned in Gordon (2005) is also featured in WALS *and* has a rarity index value: Principense. All other African languages from WALS are simply "too big" to show up in this sample.

References

Andersen, Henning
1988 Center and periphery: adoption, diffusion, and spread. In *Historical Dialectology. Regional and Social*. Jacek Fisia (ed.), 39–83. Berlin / New York: Mouton de Gruyter. (Trends in Linguistics; 37).

Bickel, Balthasar and Johanna Nichols
2003 *Typological enclaves*. Paper presented at the 5th Biannual Conference of the Association for Linguistic Typology (ALT V), Cagliari, September 18, 2003. [Slides: http://www.uni-leipzig.de/∼autotyp/download/enclaves@ALT5-2003 BB-JN.pdf]

Bickel, Balthasar
2006 *What favors the development of rara? A Himalayan case study*. Paper presented at the conference Rara & Rarissima — Collecting and interpreting unusual characteristics of human languages, Leipzig, 29 March–1 April 2006. [Slides: http://www.uni-leipzig.de/∼bickel/research/presentations/himalayan _rara2006.ppt.pdf]

Comrie, Bernard
2005 Endangered numeral systems. In *Bedrohte Vielfalt. Aspekte des Sprachentods. Aspects of language death*. Jan Wohlgemuth and Tyko Dirksmeyer (eds.), 203–230. Berlin: Weißensee.

Crystal, David
2000 *Language Death* Cambridge etc.: Cambridge University Press.

Cysouw, Michael
2004 *On the distribution of rare characteristics*. Manuscript. Leipzig. http://email.eva.mpg.de/∼cysouw/pdf/cysouwRARA.pdf

Cysouw, Michael
2005 What it means to be rare: the case of person marking. In *Linguistic Diversity and Language Theories*. Zygmunt Frajzynger, Adam Hodges and David S. Rood (eds.), 235–258. Amsterdam: Benjamins.

Cysouw, Michael
forthc. *Quantitative explorations of the world-wide distribution of rare characteristics, or: the exceptionality of north-western European languages.* To appear in *Exception in Language.* Horst Simon and Heike Wiese (eds.). Berlin / New York: Mouton de Gruyter.

Wohlgemuth, Jan and Michael Cysouw
2010 *Rara & Rarissima: Documenting the fringes of linguistic diversity*. (Empirical Approaches to Language Typology; 46). Berlin / New York: Mouton de Gruyter.

Dixon, R. M. W.
1997 *The rise and fall of languages.* Cambridge etc.: Cambridge University Press.

Dryer, Matthew S.
2005 Position of Tense-Aspect Affixes. In *The World Atlas of Language Structures*. Martin Haspelmath et al. (eds.), 282–285. Oxford etc.: Oxford University Press.

Frerick, Daniela
2006 *Raritäten in den Sprachen der Welt*. M.A. thesis, Westfälische Wilhelms-Universität Münster.

Gordon, Raymond G. Jr.
2005 *Ethnologue: Languages of the world* 15th edition. Dallas: Summer Institute of Linguistics. http://www.ethnologue.com

Grenoble, Lenore A. and Lindsay J. Whaley
1998 Toward a typology of language endangerment. In *Endangered Languages. Current issues and future prospects*. Lenore A. Grenoble and Lindsay J. Whaley (eds.), 22–54. Cambridge etc.: Cambridge University Press.

Güldemann, Tom, Patrick McConvell and Richard Rhodes (eds.)
in prep. *Hunter-gatherers and linguistic history: a global perspective*. submitted to Cambridge University Press.

Hale, Ken
1998 On endangered languages and the importance of linguistic diversity. In *Endangered Languages. Current issues and future prospects*, Lenore A. Grenoble and Lindsay J. Whaley (eds.), 192–216. Cambridge etc.: Cambridge University Press.

Hammarström, Harald
this volume Rarities in numeral systems. In *Rethinking Universals: How rarities affect linguistic theory*. Jan Wohlgemuth and Michael Cysouw (eds.), 11–59. (Empirical Approaches to Language Typology; 45). Berlin / New York: Mouton de Gruyter.

Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.)
2005 *The World Atlas of Language Structures*. Oxford / New York: Oxford University Press.

Kelkar-Stephan, Leena
2010 Future tense to express habitual past or present, and past tense to express immediate future! In *Rara & Rarissima: Documenting the fringes of linguistic diversity*, Jan Wohlgemuth and Michael Cysouw (eds.), 211–233. (Empirical Approaches to Language Typology; 46). Berlin / New York: Mouton de Gruyter.

Krauss, Michael
1992 The world's languages in crisis. *Language* 68 (2): 4–10.

Kulick, Don
1992 *Language shift and cultural reproduction. Socialization, self, and syncretism in a Papua New Guinean village*. Cambridge: Cambridge University Press.

Nettle, Daniel
1999a Is the rate of linguistic change constant? *Lingua* 108: 119–136.

Nettle, Daniel
1999b *Linguistic diversity* Oxford etc.: Oxford University Press.

Nettle, Daniel
1999c Using Social Impact Theory to simulate language change. *Lingua* 108: 95–111.

Nettle, Daniel and Suzanne Romaine
2000 *Vanishing Voices. The Extinction of the World's Languages*. Oxford etc.: Oxford University Press.

Plank, Frans
2000 *Das grammatische Raritätenkabinett. A leisurely collection to entertain and instruct*. Manuscript. Universität Konstanz.
http://ling.uni-konstanz.de/pages/proj/Sprachbau/rara.html

Trudgill, Peter
2004 Linguistic and social typology: The Austronesian migrations and phoneme inventories. In *Linguistic Typology* 8 (3): 305–320.

Wohlgemuth, Jan
2009 *A Typology of Verbal Borrowings*. Berlin / New York: Mouton de Gruyter. (Trends in Linguistics, Studies and Monographs; 211).

Wohlgemuth, Jan and Sebastian Köpl
2005 Endangered Subsystems. In *Bedrohte Vielfalt. Aspekte des Sprachentods. Aspects of language death*. Jan Wohlgemuth and Tyko Dirksmeyer (eds.), 177–186. Berlin: Weißensee.

Wurm, Stephen A.
1998 Methods of Language Maintenance and Revival, with Selected Cases of Language Endangerment in the World. In *Studies in Endangered Languages*. Kazuto Matsumura (ed.), 191–212. (ICHEL Linguistic Studies; 1). Tokyo: Hituzi Syobo.