Some more details about the definition of rarity

Michael Cysouw

Replying to the many stimulating comments raised by Dahl, I am first rather astound by his assertion that I did not define the term ‘rare’. In fact, the whole of Section 3 defines the precise mathematical operalization of my notion of rarity. And indeed, my notion of rarity is a relative one (and I would even go as far as to argue that a notion of ‘absolute rarity’ is meaningless, cf. Cysouw 2003). Still stronger, also the evaluation of the (relatively defined) Rarity Indices is relative. I explicitly do not presuppose any absolute norm separating ‘low’ from ‘high’ Rarity Indices, because I would not know of any data that could help us set such a norm. Thus, the only observations I make in the paper are about the most extreme (relative) rarities as compared to all other (relative) rarities. The list of rare traits of Northwestern European languages in Section 7 is thus a list of ‘relative relative rarity’. Whether these traits are really all noteworthy is of course open to interpretation. Looking at the values of the Mean Group Rarity Index for the traits themselves (as reported on in the first column of Table 4), I would suggest that the first four are really much more significant rarities in northwestern Europe than the other in the list. Still, I find it hight stimulating to know what other European characteristics should be considered rare when the notion of rarity is interpreted a bit more lenient. Just to take up the least extreme case of relative pronouns (as referred to by Dahl), this is indeed found in 7.2 % of the world’s languages, which one might (or might not) find rare. However, looking at the worldwide distribution of relative pronouns, shown here in Figure 1 (Comrie & Kuteva 2005 = WALS 122), it is clear that it actually is a clear example of a regionally bound rarity.

Next, Dahl discusses two possible problems with my notion of rarity. First, from the context of the theme of the present collection of papers he warns that the intuitive notions of rarity and exceptionality do not necessarily coincide. In principle, I completely agree with this comment, as I write in the introduction to the paper “exceptionality is a more encompassing term than rarity.” However, I think that the difference proposed by Dahl does not differentiate the two. For something to be called an exception, Dahl argues, there has to be some presupposed generalization relative to which it can be an exception. Now, when a trait

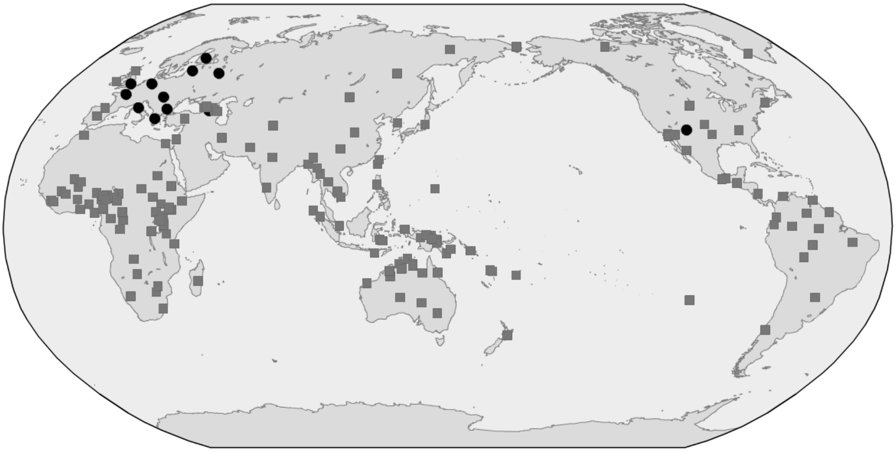

Figure 1. Usage of relative pronouns (dots) compared with other relativization strategies (squares) for the relativization of subject (adapted from Comrie & Kuteva 2005).

'X' is rare, but the opposite trait 'not-X' is not be definable (or only negatively definable by saying it is not X), then it is difficult to argue relative to what X is an exception. Here I disagree. The only generalization that is necessary is the presence of one trait (or a group of traits) that is common, and then everything else can be declared both exceptional and rare relative to the common case(s). One example discussed by Dahl concerns the typology of comparative constructions (Stassen 2005 = WALS 121). There are four types distinguished, one of which is more common than the others: Locational (47%), Exceed (20%), Conjoined (20%), and Particle (13%). Now, relative to the Locational strategy, all other are (more or less) rare and (more or less) exceptional. The radical situation would be an extremely fine-grained typology of the world's languages in which all types are rare (implying of course that there are very many different types). In this situation, I do not think anybody would want to claim that all types are exceptions, because indeed there is nothing to be an exception against. However, in my operalization of rarity this situations would also not result in the presence of any rare types. In the Rarity Index, as proposed in (3) in the paper, the proportion of occurrence is taken relative to the number of types that are distinguished. The result is that in the hypothetical situation with very many roughly equally frequent small types, the Rarity Index will consider all types to be not rare. So, as far as there are problems with the definability of the 'non rare' counterpart, I think the interpretations of rarity and exceptionality coincide.

Secondly, Dahl argues that a trait might be a composition of various independent characteristics, only the combination of which is rare. In such situations rarity should be assessed relative to the expected intersection of the traits in isolation. I completely agree with this, but the problem is caused by the unstructured coding of the values of WALS. Unfortunately, WALS does not include explicit information on the finer-grained structure of the traits distinguished. For the present paper, I decided not to perform any recoding of the WALS data, as this would be a project in it's own right (see Footnote 4 of the paper and the reference therein). But suppose one would perform such recoding, as suggested by Dahl, then the computation of the rarity index for composed traits would indeed change. As an example, let's consider the WALS map on uvular consonants (Maddiesson 2005 = WALS 6) that was brought up by Dahl. There are four different types distinguished in this map that can easily be decomposed as an intersection of two binary parameters, as shown in Table 1. There is a strong correlation between these parameters (Fisher's Exact $p < 10^{-7}$). This implies that the twelve cases of 'uvular continuants without uvular stops' are actually much less than would be expected by chance alone (expected frequency is $480 \times 60/566 = 50.9$).

Table 1. Typological distribution of uvular consonants

		Uvular Stops		
		No	Yes	Total
Uvular Continuants	No	468	38	506
	Yes	12	48	60
	Total	480	86	566

The Rarity Index, as shown in (3) in the original paper, is actually of the form "expected proportion divided by observed proportion" (E/O). The observed proportion (O) is the frequency of a trait f_i divided through the total number of languages f_{tot} (i.e. 12/566 in the current example). The expected proportion (E) that I used in the paper was simply the expectation under assumption of independence, viz. $1/n$, where n is the number of values distinguished (i.e. 1/4 in the current example). The Rarity Index for this trait is thus $E/O = f_{tot}/(n \times f_i) = 566/(4 \times 12) = 11.8$. However, when the feature is decomposed as shown in Table 1, then the expected proportion changes: the expected proportion is the product of the independent proportions of the decomposed traits. In the example the expected proportion is the proportion of ‘no uvular stops’ times the proportion of ‘yes uvular consonants’ (i.e. $480/566 \times 60/566 = 0.09$, which is noteably smaller than 1/4 as assumed in the paper). In this way, composed traits that have a lower expectation than 1/n get a lower ‘Composed’ Rarity Index. For the present example this index would be $480/566 \times 60/566 \times 566/12 = 4.24$, which is clearly smaller than the 11.8 from the index as used in the paper. In general, when a feature f is decomposed into a set of co-occurring features f_1, f_2, f_3, É f_t then the expected proportion for f_i is the product of all independent proportions, see (1), and the Rarity Index (RI) changes accordingly, as shown in (2). However, this all of course highly depends on any proposed decomposition of WALS features. In the current example the decomposition is rather unproblematic, but for many other features in WALS this is not as easy.

(1)
$$E(f_i) = \prod_{s=1}^{t} \frac{f_{s_i}}{f_{tot}}$$

(2)
$$RI(f_i)I\frac{E}{O} = E \times \frac{1}{O} = \left(\prod_{s=1}^{t} \frac{f_{s_i}}{f_{tot}}\right) \times \frac{f_{tot}}{f_i}$$

Finally, building on the discussion in DahlÕs reply, I would like to suggest that the relation between complexity and rarity is of implicational nature, in the sense that complexity probably implies rarity, but clearly not vice versa. As for the relation between areal diversity and rarity, I am not convinced that there should be any relation. Of course, in highly diverse areas more rarities will be found, but so would common traits. The real question should be whether the proportion of rare traits to common traits correlates with diversity. As far as I am concerned, the verdict on this matter is still open.

References

Comrie, Bernard, and Tania Kuteva
2005 Relativization on subjects. In *The World Atlas of Language Structures*, Martin Haspelmath, Matthew Dryer, David Gil and Bernard Comrie (eds.), 494–497. Oxford: Oxford University Press.

Cysouw, Michael
2003 Against implicational universals. *Linguistic Typology* 7: 89–10.

Maddiesson, Ian
2005 Uvular consonants. In *The World Atlas of Language Structures*, Martin Haspelmath, Matthew Dryer, David Gil and Bernard Comrie (eds.), 30–33. Oxford: Oxford University Press.

Stassen, Leon
2005 Comparative constructions. In *The World Atlas of Language Structures*, Martin Haspelmath, Matthew Dryer, David Gil and Bernard Comrie (eds.), 490–493. Oxford: Oxford University Press.