Quantitative explorations of the worldwide distribution of rare characteristics, or: the exceptionality of northwestern European languages

Michael Cysouw

Abstract. In this article, the distribution of rare features among the world's languages is investigated, based on the data from the *World Atlas of Language Structures* (Haspelmath et al. 2005). A Rarity Index for a language is defined, resulting in a listing of the world's languages by mean rarity. Further, a Group Rarity Index is defined to be able to measure average rarity of genealogical or areal groups. One of the most exceptional geographical areas turns out to be northwestern Europe. A closer investigation of the rare characteristics that make this area exceptional concludes this article.

1. Introduction[1]

From a cross-linguistic perspective, the notion of exceptionality is intricately intertwined with assumptions about normality. A language showing an 'exceptional' characteristic is much too often just a language with a different trait as commonly found in the few 'normal' European national standard languages widely investigated in current linguistics. Unfortunately, from a worldwide perspective it is these European national standard languages that often turn out to be atypical – as will be shown later on in this article. Instead of assuming knowledge about what is normal or exceptional for a human language, I will investigate exceptionality empirically by taking account of the worldwide linguistic diversity.

One way to empirically approach the notion of exceptionality is to replace it with the notion of rarity. Strictly speaking, exceptionality is a more encompassing term than rarity. However, rarity is much easier to operationalise when dealing with large amounts of data. In this article, a trait will be considered ex-

1. I thank Bernard Comrie, the editors of the present volume, and one anonymous reviewer for their comments and input on the basis of an earlier version of this paper.

ceptional when it is rare with regard to the known worldwide diversity. Such an approach can only be taken given a large amount of data about the world's linguistic diversity. Such a database has recently become available in the form of the *World Atlas of Language Structures* (WALS, Haspelmath et al. 2005), and I will gratefully draw on this enormous dataset for the present investigation of rarity among the world's languages.

This paper is organised as follows. First, in Section 2, I will introduce the *World Atlas of Language Structures* from which the typological data are drawn that form the basis for my calculations of rarity. In the following Section 3, the quantitative approach to compute rarity from typological data is explained. Section 4 then looks at the overall rarity for individual languages, claiming the South American language Wari' to be one of the languages with the highest index level of rare characteristics. In Section 5, the calculation of rarity is extended to encompass groups of languages, and this calculation is applied to genealogical families. The Kartvelian and Northwest Caucasian language families turn out to be the families with the highest index level of rare characteristics. In Section 6, the calculation of group rarity is used to investigate areal centres of high rarity. Various geographical areas with a high level of rarity are identified. Most fascinatingly, northwestern Europe ends up on top as the linguistically rarest geographical area in the world. Section 7 investigates the exceptionality of northwestern Europe more closely, identifying twelve features that make this area so unusual from a worldwide perspective. These characteristics are all linguistically independent from each other, indicating that the exceptionally high level of rarity is probably a historical coincidence, possible enlarged by some structural bias of European scholarly tradition in linguistics.

2. Using the *World Atlas of Language Structures*

The *World Atlas of Language Structures* (WALS, Haspelmath et al. 2005) is a large database of structural (phonological, grammatical, and lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of more than 40 authors, many of them the leading authorities on the subject.[2] It is published as a printed book in traditional atlas format, but also

2. The WALS is an exceptionally large collaborative project, involving many different authors. As suggested by the editors of the WALS, I will not refer to all the authors of the WALS when referring to the WALS as a whole. However, when the content of one particular chapter is discussed, the author of this chapter will be referred to in the usual way, like reference is made to chapters in edited books. Yet, because I have been using the complete data as supplied by the WALS for the cal-

accompanied by a fully searchable electronic version of the database. The atlas consists of 142 maps with accompanying texts on diverse features of human language (such as vowel inventory size, noun-genitive order, passive constructions, and 'hand'/'arm' polysemy), each of which is the responsibility of a single author (or team of authors). Each map shows between 120 and 1,370 languages. Altogether more than 2,600 languages are shown on the maps, and more than 55,000 dots give information on structural characteristics of these languages.[3]

In informal discussion, some doubts have been uttered as to the reliability of the data in the WALS. The reason for these doubts is that most data points have been coded by typologists on the basis of extant descriptive material, and not by specialist of the languages in question. As a test case, Wälchli (2005) checked the 119 coding points for Latvian and found these WALS-data to be reasonably well representative of the language. Latvian is a 'hard' case for reliability, because the editors urged all authors to include this language in their map (Latvian is one of the so-called 'basic 100-language sample'). Further, Latvian is a well-known and well-described language, but the problem for typologists is that there is no central reference work to check for any information on this language. This led to a few errors in the WALS, because authors sometimes based

culations of the rarity indices, I take this opportunity to thank the editors and all the authors for making this kind of research possible (in alphabetical order): Andreas Ammann, Matthew Baerman, Dik Bakker, Balthasar Bickel, Cecil H. Brown, Dunstan Brown, Bernard Comrie, Greville G. Corbett, Sonia Cristofaro, Michael Cysouw, Östen Dahl, Michael Daniel, Ferdinand de Haan, Holger Diessel, Nina Dobrushina, Matthew S. Dryer, Orin D. Gensler, David Gil, Rob Goedemans, Valentin Goussev, Martin Haspelmath, Johannes Helmbrecht, Oliver A. Iggesen, Paul Kay, Ekkehard König, Maria Koptjevskaja-Tamm, Tania Kuteva, Ludo Lejeune, Ian Maddieson, Luisa Maffi, Elena Maslova, Matti Miestamo, Edith Moravcsik, Vladimir P. Nedjalkov, Johanna Nichols, Umarani Pappuswamy, David Peterson, Maria Polinsky, Carl Rubino, Peter Siemund, Anna Siewierska, Jae Jung Song, Leon Stassen, Thomas Stolz, Cornelia Stroh, Stephan Töpper, Aina Urdze, Johan van der Auwera, Harry van der Hulst, Viveka Velupillai, Ljuba N. Veselinova and Ulrike Zeshan. Further, I would like to thank Hans-Jörg Bibiko for supplying the WALS *Interactive Reference Tool*, with which the maps in this paper are made.

3. Note that with about 142 features and 2,600 languages, there should be as many as 369,000 datapoints. With the actually available 55,000 datapoints 'only' about 15 % of the data matrix is filled. For many statistical approaches this low coverage is a problem, and only carefully selected parts of the data, resulting in a higher coverage, can be used. In the approach presented in this paper, I will attempt to use the complete data, notwithstanding the many missing values. However, special statistical corrections, as described in Section 3, are needed to work around the problem of missing values.

their judgements on sources that were not the best for their particular question. Wälchli (2005) notes five such errors (= 4.2%), in which it is understandable from the sources used that a linguist might be led to the wrong conclusions. Further, Wälchli found two errors in the WALS that appear to be practical mistakes (= 1.7%). From all information supplied by the authors (e.g. from the examples included), it is clear that the author knew the right coding. However, by some unidentifiable problem in the long chain of work-phases, starting with the collection of the data up to the final publication of the WALS, somewhere an error arose. In a large-scale enterprise like the WALS, it is impossible to avoid such practical errors completely. Their low number for Latvian even argues for the high reliability standard of the WALS.[4]

3. Computing a rarity index

The principal idea of the present investigation is to use this enormous WALS-database for 'holistic' typology. In the WALS, there are features coded from all areas of linguistic structure, so it is possible to look for correlations between widely different aspects of linguistic structure. For the present analysis, I will not look at the content of the features, but only consider their relative ubiquity. Are there languages, families or areas that have more rare characteristics than others? To investigate this question, I devised a rarity index – a calculation to estimate the relative ubiquity of characteristics of a language, as measured by the data in the WALS. The basic idea behind the rarity index is to compute the chance of occurrence for all characteristics of a particular language, and then take the mean over all these chances of occurrence. In essence, this results in an average rarity for a language. However, there are various confounding factors mediating between chance and rarity, which make it necessary to introduce a few extra steps in the evaluation of the chances of occurrence.

Before I explain these confounding factors and the resolution used, let me first introduce some WALS-terminology. The data in the WALS is organised into FEATURES and VALUES. A feature is a parameter of linguistic variation, shown as a double-paged map in the printed atlas (e.g. the first map depicts the size of the consonant inventory, Maddieson 2005a). Within each feature,

4. The data as brought together in the WALS is beyond doubt the largest and best organised survey of structural linguistic characteristics of the world's languages. However, there are various problems with the coding structure of the data that make it difficult to use the data for large-scale quantitative investigations without recoding them (cf. Cysouw et al. 2005). In this paper, I disregarded these problems and took the data as supplied in the atlas without doing any recoding.

each language has a VALUE. A value is the characterisation of the language for the feature in question (e.g. in the first map on consonant inventories, English – with 24 consonants – has the value 'average', defined as the range between 19 and 25 consonants). As a first approach to a rarity index, the rarity of a value might be formalised by simply taking the chance occurrence of that value. For example, the value 'average' of the feature 'consonant inventories' occurs in 181 languages out of a total of 561 languages coded for this feature. There is thus a chance occurrence of 181/561 = 0.322 for this value. However, this chance cannot simply be interpreted as an indication of the rarity of the value.

The first problem is that different maps distinguish different numbers of values, and the chance occurrences thereby have different impact on the evaluation of rarity. For example, in the map on consonant inventories there are five different values distinguished (small, moderately small, average, moderately large, large), but in the next map on vowel quality inventories (Maddieson 2005b) there are only three different values distinguished (small, average, large). Now, consider the value 'large' of the feature 'vowel quality inventory'. This value has a chance occurrence of 183/563 = 0.325, almost exactly the same as for 'average' consonant inventory discussed previously. However, with only three values distinguished for vowel quality inventories, such a chance of around one-third should count as just average rarity. In contrast, with the five values as distinguished for consonant inventories, a chance of one-third is actually higher than expected from an equal distribution (in which the chance would be one-fifth), and should thus be counted as relatively low rarity (or 'common'). Conversely, in a hypothetical feature with only two values distinguished, a chance expectation of around one-third would count as relatively high rarity (or 'unusual').

The simplest solution to this problem is to multiply the chance occurrence of each value with the number of values distinguished, as shown in the definition of the Rarity Index in (1). The feature 'consonant inventories' distinguished five different values, so the rarity index for the value 'average' is $5 \cdot 0.322 = 1.61$, which is higher (and thus less rare) than the index for the value 'large' of the feature 'vowel quality inventory $3 \cdot 0.325 = 0.975$. Note that a rarity index of around 1.0 means that the chance occurrence of a particular value approaches the chances for equally distributed features. For a feature with x values, an equal distribution would mean a chance of occurrence for each value of $1/x$. If the empirically established chance occurrence of a particular value approaches $1/x$, the rarity index for this value approaches $x \cdot (1/x) = 1$. For practical reasons, I used the inverse of this index, as shown in (2). The higher this index, the higher the rarity of the value in the WALS data. Using this inverse has the nice effect that the mean of all indices over all languages coded for a particular feature is

also exactly one, as shown in (3). The equation in (3) can easily be verified by writing out the terms in the summation.

(1) $$R_{f_i} = n \cdot \frac{f_1}{f_{tot}}$$

n = number of values of a particular feature
f_i = frequency of value i
f_{tot} = total number of languages coded for this feature

(2) $$R_{f_i} = \frac{f_{tot}}{n \cdot f_i}$$

(3) $$\frac{\sum_{i=1}^{n} (R_{f_i} \cdot f_i)}{f_{tot}} = 1$$

The formula in (2) thus defines the rarity-index of a value. The next step is now to compute a rarity index for a language on this basis. The basic idea for computing a rarity index of a language is to take the mean of all rarity indices for all the characteristics of this language, throughout all the maps in the WALS. However, a second confounding factor is the number of maps in which a particular language occurs. The data of the WALS is not complete, meaning that not every language is coded in every map. Many languages are only coded in very few maps. For this reason, simply taking the mean rarity over all values is not a good measure to evaluate which language has the most unusual characteristics. If a particular language is only coded for few features in the WALS, there will be strong random effects. Languages with few code-points in the WALS will show more extreme values of mean rarity, both to the high and the low side. This effect can be observed in Figure 1, in which the mean rarity for all 2,600 languages in the WALS is plotted against the number of features coded (each point in the figure represents one language).[5] The fewer features are coded for a language, the more extreme mean rarities occur.

To normalize this effect, I evaluated the distribution of mean rarity by a randomization technique. The randomization proceeded as follows. For each

5. For clarity of depiction, the logarithm of mean rarity is shown in this figure. Using the logarithm has the visual effect of separating the out the values some more, thereby showing more clearly the distribution of the points in the figure. Another effect is that the mean rarity now centers around zero, because log (1) = 0.

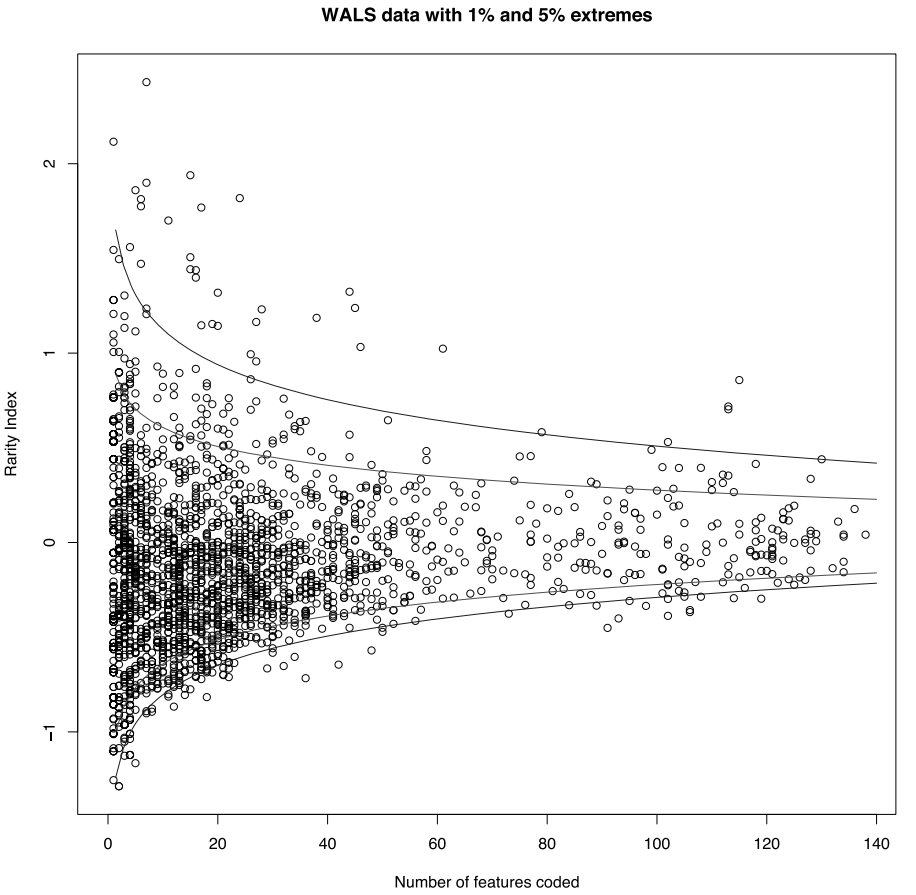

Figure 1. Plot of mean rarity indices against the number of features coded, with lines indicating 1 % (outer lines) and 5 % (inner lines) extremes as measured by a randomization procedure.

number of features coded (ranging between 1 and 139),[6] a thousand fictitious languages were created. For each invented language, a set of features was selected completely at random. Within each feature, a value was selected semi-randomly. The value selection was guided by the actual chance occurrences of each value in the WALS. In this way, each set of a thousand fictitious lan-

6. The WALS has 142 maps, but for the present investigation the two maps on sign languages and the map on writing systems have been disregarded, leading to a maximum of 139 features available.

guages has the same distribution of values as the real WALS. For example, the number of languages with average consonant inventory will be around 32.2% in each set of thousand languages. One such set of a thousand languages was made with each language being coded for one feature only. Then one set was made with each language being coded for two features, etcetera, finishing with a set of thousand languages in which each language was coded for 139 features. The mean rarity for all these invented languages was computed, thus giving a thousand mean rarity values for each number of features.

Using all these fictitious languages, the mean rarity of a real language can be evaluated. For example, Dutch is coded for 67 features and has a mean rarity of 1.66. The question now is how extreme this value is. The mean rarity is higher than 1.00, so there appears to be a relatively high level of rarity in this language. But is this really much higher than 1.00, or is a value of 1.66 still within the expected variation? To evaluate this, the set of thousand fictitious languages coded for 67 features were used. Among this set of thousand made-up languages, there turned out to be 96 (= 9.6%) with a mean rarity higher than 1.66. Thus 904 (= 90.4 %) fictitious languages had a smaller mean rarity. From this it can be concluded that the mean rarity of Dutch is really rather high (even among the highest 10 %). Note that this value is not a real significance value, as given by statistical analyses, although it is a somewhat similar concept. This value indicates the relative unusualness of a particular language within the WALS dataset. Using such evaluations, lines representing the 1 % and 5 % extremes can be drawn in Figure 1. These lines show the boundary between the extremes in the fictitious languages, indicating which of the real languages (represented by the dots) belong to these extremes.

4. Rarity indices for individual languages

Using this evaluation of mean rarity by randomization, the languages with the most extreme mean rarity are shown in Table 1. In this table, a mean rarity 'index level' is indicated by a percentage in the last column. For example, 100 % means that this particular mean rarity is higher than all thousand fictitious languages for the number of features coded. The first six languages all fall in the level of this most extreme mean rarity. As can be seen in the penultimate column, the actual values of mean rarity differ widely. Winnebago has a very high mean rarity (11.37), which is even high considering that this language is only coded for 7 features (judging from the index level of 100 %). In contrast, Wari' is also included among the most extreme index levels with a mean rarity of 'only' 2.36 (remember that the mean over all the data in the WALS is 1.00).

Table 1. Top 15 of languages according to mean rarity index level. Within each level, they are ordered to the number of features coded, though this is for presentational purposes only.

Language	Genus	Features Coded	Mean Rarity	Index Level
Wari'	Chapacura-Wanhan	115	2.36	100
Dinka	Nilotic	45	3.45	100
Jamul Tiipay	Yuman	44	3.76	100
Nuer	Nilotic	28	3.42	100
Karó (Arára)	Tupi-Guarani	24	6.16	100
Winnebago	Siouan	7	11.37	100
Chalcatongo Mixtec	Mixtecan	113	2.05	99.9
Kutenai	Kutenai	113	2.02	99.9
Kombai	Awju-Dumut	38	3.27	99.9
Dahalo	Southern Cushitic	17	5.86	99.9
Maxakali	Maxakali	15	6.95	99.9
Warrwa	Nyulnyulan	20	3.74	99.8
Bunuba	Bunuban	16	4.21	99.8
Eyak	Eyak	16	4.05	99.8
Yawuru	Nyulnyulan	15	4.51	99.8

However, this value is achieved with as much as 115 features being coded, and for such many features, a mean rarity of 2.36 is apparently still highly significant.

Although such a listing of the world's languages as to the level of rarity satisfies a currently widespread felt need for rankings, its merits are doubtful. It would be interesting if particular genealogical or areal groups showed up high in this listing. However on first inspection this is not the case. There are two Nilotic and two Nyulnyulan languages among the top 15, which is indicative, though not convincing. Areally, among the top 15 as presented in Table 1, only the languages from Eurasia are absent. The majority of the top 15 is from the Americas (eight languages), three are from Africa and four from Australia/New Guinea. However, this is partly an effect of the random cut-off point of the top 15, chosen here for reasons of space. In Figure 2, a world map is presented, showing the geographical distribution of the top 5 % languages (i.e. all languages with an index level of 95 % and higher). There appears to be a relatively high density of languages in Africa (around the equator) and northern Australia/New Guinea, but these are also regions with a high number of languages represented in the WALS data. I would argue that from this distribution alone, there does not ap-

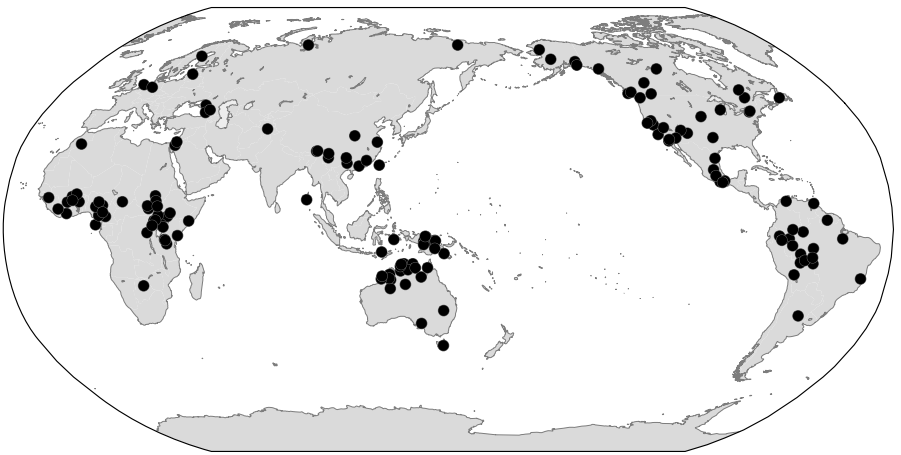

Figure 2. World map showing the top 5 % on the rarity index level of the languages in the WALS.

pear to be a reason to declare any group of languages to stand out as showing a particular high level of unusualness.

5. Rarity indices for groups of languages

To further investigate the distribution of rarity among the world's languages, I computed rarity for groups of languages. based the index levels for each language (as discussed in the previous section). Such values for Group Rarity (GR) are useful to evaluate the relative rarity of a genealogical or an areal group of languages. As a measure of Group Rarity, I have used a weighted mean of the rarity *index levels* of the individual languages. Basically, to compute this weighted mean, I took the mean of all index levels of the individual languages (not the mean rarity itself), and weighted the languages according to the logarithm of the number of features coded, as shown in the formula in (4). Because of this logarithm, the languages with more features coded have slightly less influence on the resulting value. Also, languages that are only coded for one feature do not have any influence, because $\log(1) = 0$.

$$GR = \frac{\sum_{i=1}^{n} \log(L_i) \cdot (\%R)_i}{\sum_{i=1}^{n} \log(L_i)} \tag{4}$$

n = number of languages in a group
L_i = number of features coded for language i
$\%R_i$ = rarity index level for language i

Using the measure of group rarity on genealogical groups results in an interesting set of linguistic families showing a high level of rarity. The top 10 linguistic families as to group rarity are shown in Table 2. Only families with more than three languages included in the WALS are shown, because I want to show effect on the level of the family. In families with only few members coded in the WALS (or few members existing in the world), high rarity of individual languages will raise the level of the whole family unproportionally.

Table 2. Top 10 of weighted rarity for linguistic families (only families shown with more than 3 languages included in the WALS data).

Family	No. of Languages	Group Rarity
Northwest Caucasian	7	87.8
Kartvelian	4	83.7
Caddoan	5	82.2
Wakashan	7	80.2
Iroquoian	8	76.3
Khoisan	11	74.5
Arauan	6	71.8
Salishan	24	71.2
Na Dene	23	70.2
Algic	31	69.9

Two families from the Caucasus (Northwest Caucasian and Kartvelian) take the first two positions on the ranking of families (the third indigenous family from the Caucasus, Nakh-Dagestanian, has only slightly higher than average rarity). Further, families from Northern America are strongly represented: Caddoan, Wakashan, Iroquoian, Salishan, Na Dene and Algic all made it into the top 10. Hokan, Eskimo-Aleut, Kiowa-Tanoan and Penutian just did not make it all the way up, though they still show an extremely high level of group rarity. From a genealogical perspective, the Caucasus and Northern America clearly stand out as having families showing a high level of group rarity.

6. Areal distribution of rarity

To evaluate whether there are geographical areas with a high preponderance of rare features, I investigated groups of languages that are geographically con-

tiguous. For each language in the database, I took the thirty nearest languages (using a simple Euclidean distance, not taking account of natural barriers) and computed the rarity for all such areal groups. The rarity index for each group is plotted on a map on the location of the centre of the group. Such an approach necessarily will show some areal consistency, because two neighbouring languages will share many of their neighbours. However, it is interesting to see where the centres of areally consistent groups are. These centres are indicative of the location of geographical areas with a high level of rarity. The higher the rarity index for a group around a particular language, the darker the dot on the map as shown in Figure 3.

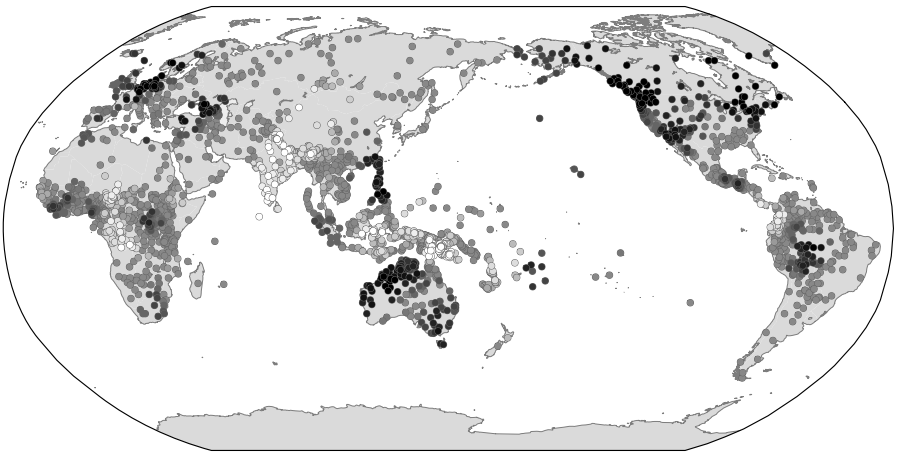

Figure 3. World map showing areal centres of rarity.

In this map, there are fifteen centres of high rarity, as summarised in Table 3. For all these areas, a centre is indicated. These centre languages are the first languages that show up in the ranking of group rarity for the areal groups. This central language is not necessarily of any importance itself. For example, Frisian only turns out to be the centre of the Northwest European cluster because it is roughly in the middle of the area including English, French and German, among others. The fact that there are fifteen centres (and not more or less) depends on the decision to compute group rarity for areal groups of thirty languages around each centre. More centres of rarity appear when, for example, groups of only ten languages are taken. However, these centres mostly split up groups found in the map shown here. When groups larger than thirty languages are used in the computations, then the clear distinctions between the various centres start

Table 3. Areas of high rarity, grouped by Macroareas.

Macroarea	Location of area with high rarity	Centre language
Eurasia	North-western Europe	Frisian
	Caucasus	Adyghe
Oceania	Philippines	Bikol
	Sumatra	Minangkabau
	Pacific	East Futuna
	Northern Australia	Walmatjarri
	Southeast Australia	Ngiyambaa
America	Northwest America	Lummi
	Northeast America	West Greenlandic
	Western North America	Havasupai
	Central America	Zapotec
	Amazonia	Pirahä
Africa	West Africa	Guro
	Central Africa	Mende
	Southern Africa	Zulu

to diminish. For the current purpose of investigating worldwide areal patterns in the WALS data, a group size of about thirty appears to be most suitable.

It is interesting to speculate why these centres appear in this worldwide survey of rarity. Several of these areal groups are considered to be typological areas (or 'Sprachbünde'). However, some areas with high rarity have no accompanying claim for areality, and many traditionally claimed linguistic areas do not show up as areas with high rarity. Although it is tempting to hypothesize that strong influence between languages might lead to the spreading of otherwise rare phenomena, the overlap between rare areas and known areal groupings is at present only approximate. However, the quantitative notion of rarity as used in this paper might be particularly useful to investigate linguistic areas as the strongest evidence for areality stems from traits that are common in a particular area, but rare elsewhere.

7. Rare characteristics of northwestern Europe

Probably the most surprising area to appear in the list of geographical areas with a high level of rarity is northwestern Europe. This area is centred on Frisian. Many of the thirty languages around Frisian are variants that are often considered West Germanic dialects. These are only coded for a few features in the

WALS, and do not have much impact on the rarity measure. When these are removed, the remaining languages in this area, all with a relatively high coverage in the WALS data, are English, German, Dutch, Frisian, and French.

The pressing question now of course is what makes these languages so exceptional? To investigate which features caused the high rarity index for this group, I considered each feature individually. Depending on the values for each feature, I took the original rarity index, as shown in (2), for each value of each language in the area. Then the mean of these rarity indices was computed, and the features were ordered to this mean. This resulted in a list of most exceptional characteristics of this area. The top ten of this list is shown in Table 4 (the mean rarity of each feature for this area is shown in the first column).

Table 4. Top 10 of the rarest characteristics as found in northwestern Europe.

Rarity	Feature	Exceptional value present in Europe
8.39	Polar Questions	Interrogative word order
7.96	Uvular Consonants	Uvular continuants only
7.93	The Perfect	Perfect of the 'have'-type
7.56	Coding of Evidentiality	Modal morpheme
4.58	Demonstratives	No distance contrast
4.32	Negative Indefinite Pronouns	No predicate negation present
4.15	Front Rounded Vowels	High and mid
3.46	Relativization on Subjects	Relative pronoun
3.14	Weight-Sensitive Stress	Right-oriented, antepenultimate involved
2.86	Order of Object and Verb	Both orders, neither order dominant

This list of exceptional characteristics of northwestern European languages will be quickly reviewed here. For more details on the coding and the decisions to distinguish between various values, please refer to the relevant texts accompanying the maps in the WALS. A summary of the presence of these exceptional traits in northwestern European languages is given in Table 5, alongside the basic percentages of these exceptional features among all the world's languages.

The exceptional features of northwestern Europe are the following. First, the marking of polar questions is unusual. In most of the world's languages, polar questions are constructed by using a question particle. Two other major marking patterns are polar questions marked solely by use of intonation or by special verb morphology. The typical northwest European change in word order to mark polar questions is extremely uncommon worldwide, with only few attestations outside of Europe (Dryer 2005e).

Uvular consonants are not very widespread among the world's languages. Maddieson (2005d) finds them only in 17 % of the world's languages. Most of these languages have at least some kind of uvular stop – possibly alongside other

Table 5. Occurrence of rare characteristics in northwestern Europe compared to their worldwide frequency.

Unusual characteristic	French	English	German	Dutch	Frisian	World
Word order in polar questions	–	+	+	+	+	1.4 %
Uvular continuants only	+	–	+			2.1 %
Perfect of the 'have'-type	+	+	+			3.2 %
Modal morpheme for evidentiality	+	–	+	+		1.7 %
No distance contrast in demonstratives	+	–	+			3.0 %
No negation with negative indefinites	–	–	+	+		5.3 %
High and mid front rounded vowels	+	–	+			4.1 %
Relative pronoun	+	+	+			7.2 %
Right-oriented stress, antepenultimate	–	+	+	+		5.4 %
Both orders of object and verb	–	–	+	+	+	6.6 %
No productive reduplication	+	+	+			15.3 %
Comparative particle	+	+		+		13.2 %

[Note: Blank cells in this table are not coded in the data from the WALS. Informal inspection and personal knowledge of the present author indicates that they are almost all to be marked as present ('plus').]

kinds of uvular consonants. The situation found in northwestern Europe, namely the existence of uvular continuants (in the form of a voiceless fricative), without the existence of uvular stops as well, is highly uncommon. Outside Europe this is mainly attested in a few incidental languages scattered throughout central Asia.

A perfect (like in English *I have read the book*), defined as a construction combining resultative and experiential meanings, is reasonably widespread throughout the world's languages. Dahl and Velupillai (2005) find a construction with similar semantics in almost half of the world's languages. However,

the typical European perfect construction of the 'have'-type (derived from a possessive construction) is a European quirk, unparalleled elsewhere in the world.

Evidentiality is the marking of the evidence a speaker has for his/her statement. Grammatical devices to code this are reasonably widespread among the world's languages. De Haan (2005) finds some kind of evidentiality in slightly more than half of the world's languages. However, the usage of a modal verb for this means, as found in northwestern Europe (e.g. Dutch *het moet een goede film zijn*, French *il aurait choisi la mort*), is extremely uncommon worldwide.

Demonstratives are normally expected to have some distinctions as to distance, like English *this* vs. *that*. In a survey of such distance contrasts in adnominal usage, e.g. *this book* vs. *that book*, Diessel (2005) finds distance contrasts in almost all of the world's languages. However, there are a few languages that do not have such distance contrasts in adnominal usage. Some examples are found in western Africa and, somewhat surprisingly, in French (*ce*) and German (*dies-* or *das*; note that *jen-* does not mark a distance contrast in modern German, although it did in older stages of the language).

Negative indefinite pronouns, like *nobody, nothing* or *nowhere*, are in most of the world's languages accompanied by a regular predicate negation. Haspelmath (2005) finds predicate negations to be obligatorily present in 83 % of the world's languages. There are only very few languages in which a negative indefinite pronoun can occur (or even has to occur) without the predicate negation. This unusual phenomenon is mainly found in a few languages in Mesoamerica and in northwestern Europe.

Front rounded vowels, like high [y] or mid [ø], are highly unusual as phonemes in a language. Maddieson (2005e) finds them only in 7 % of the world's languages. Both the high and the mid front rounded vowels are mostly found in some languages of northern Eurasia, among them French and German. Related to this unusual characteristic are the exceptionally high number of vowel quality distinctions (Maddieson 2005b) and the low consonants to vowel ratio (Maddieson 2005c) of northwestern European languages. These two related characteristics just did not make it into the top ten of rare features of northwestern European languages.

Relative clauses are a much debated and widely investigated aspect of human language. It might come as a surprise to many linguists that the typical European usage of a relative pronoun is only highly sporadically found outside of Europe (Comrie and Kuteva 2005).

There is a large variety of stress-systems attested among the world's languages. The typical northwestern European system is a weight-sensitive stress system in which also the antepenultimate syllable is involved (Goedemans and

Van der Hulst 2005). Such a system is unusual, though it is also found in the near east and sporadically throughout the world's languages.

The last rare characteristic in the top ten of rarest traits in northwestern Europe is the variable order of verb and object (Dryer 2005c). This variability is paralleled in the likewise rare trait of having variable order of genitive and noun (Dryer 2005d), which, however, did not make it into the top ten of rare characteristics of northwestern Europe.

Finally, two interesting characteristics of northwestern European languages that also did not make it into the top ten of rarity deserve quick mention here. First, the languages of northwestern Europe are exceptional because they do not allow for productive reduplication (Rubino 2005) and, second, because they use a special particle for comparative constructions (Stassen 2005).

Going through this list of rare characteristics of northwestern European languages, it is important to realize that there are no worldwide correlations between any pair of these features. From a typological perspective, all these features appear to be independent parameters of linguistic variation. At least, I have not been able to find any clearly significant correlations between any two features in this list in the WALS data. Not even the presence of a 'have'-perfect and a 'have'-possessive correlate. This would mean that there are no internal linguistic reasons for these features to co-occur in northwestern Europe. It is probably an accidental effect of historical contingency that exactly these rare features are found in this area, and not others.

As can be seen from the summary in Table 5, the exceptional characteristics are basically found in Continental West Germanic, with English and French sharing these unusual traits in about half of the cases. This areal centre roughly coincides with the *Charlemagne* Sprachbund, or *Standard Average European* (SAE) as summarised in Haspelmath (2001). Some of the typical characteristics of SAE languages, as described by Haspelmath (2001), are also found in the present investigation. In particular, the word order in polar question, the perfect of the 'have'-type, no negation with negative indefinites, the special structure of the relative clause, and the usage of comparative particles are noted in both investigations. However, there are also clear differences between my claim for northwestern Europe to have many unusual characteristics and Haspelmath's claim that the European languages share many traits. For example, the existence of definite and indefinite articles is a clear case of a pan-European characteristic (Haspelmath 2001: 1494). This areality is also found in the WALS maps on articles (Dryer 2005a, 2005b). However, articles are not nearly as rare on a worldwide basis to show up in the present investigation. In contrast, the presence of the rare uvular continuants cannot be claimed to be a typical European characteristic. In fact, almost no European languages have such conso-

nants (except for Continental West Germanic and French), but their presence is exceptional enough from a worldwide perspective to end up as a rare trait of northwestern Europe. Summarising, the claims for SAE as a linguistic area and the presence of many exceptional characteristics in this area are supplementary claims, probably both to be explained by long-term influence between the languages in question.

There are a few words of caution to be added to these results. Matthew Dryer, one of the WALS editors, warns (in personal communication) that in some cases the exceptionality of northwestern Europe in the WALS data might have been enlarged by more or less deliberate decisions. He suggests that the WALS editors and authors might have included typical European oddities as separate values, thereby enhancing the exceptional profile of this area. This might indeed, to some extent, be the case for polar questions, modal evidentials, the 'have'-perfect, relative pronouns and particle comparatives. These characteristics are really European quirks. They are common in Europe, and any linguist with a training based on European languages (which means almost all linguists) will at first consider them to be the norm. While investigating the worldwide typological diversity, it will probably come as a surprise that European languages are exceptional in these respects. This might have raised the interest to investigate these characteristics of human language, eventually leading to their inclusion in the WALS. Though this process might have had some effect, there are still numerous rare features in Europe that do not seem to have been influenced by this bias.[7]

8. Conclusion

The usage and interpretation of large linguistic typological databases is still in its infancy. In this paper, I have laid out a first attempt to approach a new large-scale typological database, the *World Atlas of Language Structures*, using quantitative methods. As a showcase, I have taken the notion of rarity and investigated the distribution of rare characteristics among the world's languages.

7. In this same vein, it might also be speculated that the strong influence from Russian and North American linguists on the research in typology in recent decades has lead to the introduction of such features as to enlarge the exceptionality of the languages in the Caucasus and North America. However, even if true, the presence of these exceptional features is still highly interesting. And there are still other areas with high rarity that show up in the present investigation. Any scientific-historical influence is probably only a minor factor influencing the results as presented in this paper.

Individual languages and linguistic families were ranked according to their level of rarity. Rarity appears to be found rather evenly distributed throughout the world's languages, though there are, of course, some languages and groups of languages that have more of it than others. The remaining question, that has to be answered by future research, is whether these languages or language groups with relatively many rare features are really 'rare languages'. This would only be the case when in a completely different dataset the same languages would have a high level of rarity as well. Personally, I do not believe that this will be the case. Circumstantial evidence for this can be discerned in Figure 1, as with a rising number of characteristics considered, the mean rarity seems to approach normality. This might indicate that throughout all structures of a whole languages, rare and common characteristics are kept in balance.

Still, it is interesting to interpret the distribution of rare traits in the current data. The most fascinating result was that the northwestern European area, centred on Continental West Germanic, turned out to be one of the most linguistically unusual geographical areas word-wide. Many of the rare characteristics as attested in this area might have been considered the norm from a European perspective, though the typological data show that these characteristics are to be considered special structures of European languages, and not of human language in general.

References

Comrie, Bernard, and Tania Kuteva
2005 Relativization strategies. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 494–501. Oxford: Oxford University Press.

Cysouw, Michael, Jeff Good, Mihai Albu, and Hans-Jörg Bibiko
2005 Can GOLD "cope" with WALS? Retrofitting an ontology onto the World Atlas of Language Structures. *Proceedings of E-MELD workshop 'Linguistic Ontologies and Data Categories for Language Resources'.*

Dahl, Östen, and Viveka Velupillai
2005 Tense and Aspect. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 266–281. Oxford: Oxford University Press.

de Haan, Ferdinand
2005 Coding of Evidentiality. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 318–321. Oxford: Oxford University Press.

Diessel, Holger
2005 Distance constrasts in demonstratives. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 170–173. Oxford: Oxford University Press.

Dryer, Matthew S.
2005a Definite articles. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 154–157. Oxford: Oxford University Press.

Dryer, Matthew S.
2005b Indefinite articles. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 158–161. Oxford: Oxford University Press.

Dryer, Matthew S.
2005c Order of object and verb. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 338–341. Oxford: Oxford University Press.

Dryer, Matthew S.
2005d Order of genitive and noun. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 350–353. Oxford: Oxford University Press.

Dryer, Matthew S.
2005e Polar questions. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 470–473. Oxford: Oxford University Press.

Goedemans, Rob, and Harry van der Hulst
2005 Weight-sensitive stress. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 66–69. Oxford: Oxford University Press.

Haspelmath, Martin
2001 The European linguistic area: Standard Average European. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 1492–1510. Oxford: Oxford University Press.

Haspelmath, Martin
2005 Negative indefinite pronouns and predicate negation. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 466–469. Oxford: Oxford University Press.

Haspelmath, Martin, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible
2001 *Language Typology and Language Universals.* Vol. 2. (Handbooks of Linguistics and Communication Science 20.2) Berlin: Walter de Gruyter.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie
2005 *The World Atlas of Language Structures.* Oxford: Oxford University Press.

Maddieson, Ian
2005a Consonant inventories. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 10–13. Oxford: Oxford University Press.

Maddieson, Ian
2005b Vowel quality inventories. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 14–17. Oxford: Oxford University Press.

Maddieson, Ian
2005c Consonant-vowel ratio. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 18–21. Oxford: Oxford University Press.

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

Maddieson, Ian
2005e Front rounded vowels. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 50–53. Oxford: Oxford University Press.

Rubino, Carl
2005 Reduplication. In *The World Atlas of Language Structures,* Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 114–117. Oxford: Oxford University Press.

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

Wälchli, Bernhard
2005 Par tipoloģijas atlantu un latviešu valodas materiālu tajā. [About the typological atlas and the Latvian material in it]. Paper presented at Letonistu seminārs [Letonists' seminary], August 6–13, 2005, Mazsalaca, Latvia.