Rethinking Universals

Empirical Approaches
to Language Typology

45

Editors

Georg Bossong
Bernard Comrie
Yaron Matras

De Gruyter Mouton

Rethinking Universals

How Rarities affect Linguistic Theory

Edited by

Jan Wohlgemuth
Michael Cysouw

De Gruyter Mouton

ISBN 978-3-11-022092-6
e-ISBN 978-3-11-022093-3
ISSN 0933-761X

Library of Congress Cataloging-in-Publication Data is available at http://catalog.loc.gov.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Printing: Hubert & Co. GmbH & Co. KG, Göttingen
∞ Printed on acid-free paper

Printed in Germany

www.degruyter.com

Preface

About this book

The idea for this volume arose in the context of a lecture by Larry Hyman during the ALT Summer School in Linguistic Typology in Cagliari preceding the fifth meeting of the Association for Linguistic Typology (*ALT V*), 2003. Mentioning the unique case of “affixation by place of articulation in Tiene”,[1] Hyman argued that there should be a more consistent interest into rarities, as a counterpart to the widely practiced pursuit of broad-scale typological generalizations. In reaction, Jan Wohlgemuth, David Gil, Orin Gensler and Michael Cysouw organized an international conference around the topic of *rara* and *rarissima* which was held in Leipzig from 29 March to 1 April 2006. The present volume consists of a selection out of the fifty-two papers that were presented at that conference.

For the conference we invited papers dealing with the description and analysis of (apparently) rare features in individual languages. Additionally, we explicitly solicited papers dealing with the reflection and discussion of the impact of *rara* on linguistic theory and linguistic universals. The papers in this volume are of the latter kind: They deal with rare phenomena that do not seem to fit into received universals and discuss how linguistic theories should approach the existence of rare and unusual phenomena. Papers dealing with the former topic are collected in the companion volume “Rara & Rarissima: Documenting the fringes of linguistic diversity”, also published by Mouton de Gruyter.

Acknowledgments

We would like to thank Bernard Comrie and the staff at the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology, especially Claudia Schmidt and Julia Cissewski, for their support with the organization of the *Rara & Rarissima* conference and the preparation of the present volume.

For their assistance in the editing and proofreading process we wholeheartedly thank our (anonymous) reviewers and Eike Lauterbach and Orin Gensler. Furthermore, we are indebted to Patrick Schulz and Johannes Reese

for their invaluable part of the typesetting and layout, as well as Pavel Iosad and Harald Hammarström for additional help with LaTeX idiosyncrasies. Finally, we are grateful to the admirably persistent and indulgent staff of Mouton de Gruyter, especially Ursula Kleinhenz, Julie Miess, and Wolfgang Konwitschny.

Leipzig, Winter 2009 / 2010
JAN WOHLGEMUTH & MICHAEL CYSOUW

Notes

[1] cf. Hyman's contribution to the companion volume

Contents

List of contributors

Michael Cysouw
Department of Linguistics
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
04103 Leipzig
GERMANY
cysouw@eva.mpg.de

Harald Hammarström
Department of Computer Science
Chalmers University
412 96 Gothenburg
SWEDEN
harald2@chalmers.se

Thomas Hanke
Friedrich-Schiller-Universität Jena
Institut für Anglistik / Amerikanistik
Ernst-Abbe-Platz 8
07743 Jena
GERMANY
thhanke@gmail.com

Alice C. Harris
Department of Linguistics
226 South College
University of Massachusetts Amherst
150 Hicks Way
Amherst, MA 01003
U. S. A.
acharris@linguist.umass.edu

Eric W. Holman
Department of Psychology
University of California
Los Angeles, CA 90095
U. S. A.
holman@psych.ucla.edu

Pavel Iosad
University of Tromsø / CASTL
Center for Advanced Study in Theoretical Linguistics
University of Tromsø
Tromsø 9037
NORWAY
pavel.iosad@uit.no

Andrej Malchukov
Department of Linguistics
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
04103 Leipzig
GERMANY
andrej_malchukov@eva.mpg.de

Matti Miestamo
Helsinki Collegium for Advanced Studies
P.O. Box 4
00014 University of Helsinki
FINLAND
matti.miestamo@helsinki.fi

Frederick Newmeyer
University of Washington
University of British Columbia
and Simon Fraser University
fjn@u.washington.edu

Jan Rijkhoff
Department of Linguistics
Aarhus University
Bygning 1410
Bartholins Allé 16, 3
8000 Århus C
DENMARK
linjr@hum.au.dk

Søren Wichmann
Department of Linguistics
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
04103 Leipzig
GERMANY
wichmann@eva.mpg.de

Jan Wohlgemuth
Nürnberger Straße 22
04103 Leipzig
GERMANY
jan@linguist.de

The other end of universals: theory and typology of *rara*

Michael Cysouw & Jan Wohlgemuth

1 *Rara* and *Rarissima*

Universals of language have been studied extensively for at least the last four decades, allowing fundamental insights into the principles and general properties of human language. Only incidentally have researchers looked at the other end of the scale. And even when they did, they mostly just noted peculiar facts as "quirks" or "unusual behavior", without making too much of an effort at explaining them beyond calling them exceptions to various rules or generalizations.

Yet, *rara* and *rarissima*, features and properties found in very few languages, can tell us as much about the capacities and limits of human language(s) as do universals. Explaining the existence of cross-linguistically rare phenomena on the one hand, and the fact of their rareness or uniqueness on the other, should prove a reasonable and interesting challenge to any theory of how human language works. The current volume consists of papers dealing with such rarities, their analysis, and their impact on the study of human language in general.

A *rarum* (and its extreme case, a *rarissimum*) is not just something that is rare or infrequently attested. In the introduction to his "Raritätenkabinett",[1] Plank defines a *rarum* as

> "… a trait … which is so uncommon across languages as not even to occur in all members of a single … family or diffusion area … Diachronically speaking, a rarum is a trait which has only been retained, or only been innovated, in a few members of a single family or sprachbund or of a few of them."

With this definition, Plank very specifically delimits a *rarum* from other infrequent phenomena among the world's languages. Following Plank, a *rarum* should not just be infrequent, but its attestations should also be independent, i. e. it should also never occur locally spread out, forming either genealogical and / or geographical clusters.

A similar view of *rara* is formulated by Bickel and Nichols (2003: 3). They distinguish between two types of *rara* that are rather different in their quality. The first type, *absolute rara*, are those that are found rarely across language families and thus *rara* in Plank's sense. One example of this type of *rara* is found in the languages Pirahã and Kawi which have no number distinction in pronouns, thus effectively violating the Greenbergian universal 43 (cf. Frerick 2006: 41; Greenberg 1963: 113). The second type, *relative rara*, are those that are rare on a global scale but common within a geographical area or a language family. A prime example for this type are click phonemes: Their distribution is restricted to Southern and Eastern Africa, where they are common among several, yet not all, groups of languages, while clicks are essentially unattested in all other parts of the world — and thus relatively rare on a global scale (cf. Frerick 2006: 10, 68).[2]

Plank (2000) suggests a few other terms for talking about rare phenomena. He proposes the term *singulare* for features found in only one language, but this term has an inherent problem when used in English: the adjective derived from it is homophonous with the noun and adjective referring to grammatical number category SINGULAR (as opposed to e. g. PLURAL). In a similar vein, *nonesuch*, the alternative term for *singulare* suggested by Plank (2000), might evoke the false interpretation that there were *no* language with such a characteristic. Furthermore, this term bears the connotation of a value judgment since *nonesuch* also means 'someone or something that is better than all others'. To avoid homonymous or misleading terms, we prefer not to adopt these terms but suggest to use *unicale / unique* instead for such features that apparently are attested in only one language. Whatever term one prefers, it is of course to a large extent only of superficial interest that there is just and exactly *one* single known example of a particular phenomenon. The study yielding this one example will only have looked at a limited set of other languages — enlarging the sample of languages might very well turn up more cases. Absolute numbers of occurrence never tell very much about the prevalence of a characteristic among the world's languages.

For the sake of brevity some linguists use the collocations *"rare language(s)"* and *"unique language(s)"* to refer to languages *having* such rare or unique characteristics. This, however, seems inappropriate to us, especially in the context of language endangerment,[3] and given the fact that, by virtue of its specific combination of features and characteristics, *every* language is unique.

2 The study of *rara*

A central goal of investigating *rara* is to fathom the variability and limits of human language structure(s). Broad-scale typological research using samples of the world's languages will give an indication about what are the common kinds of linguistic structures. Yet, such studies will not be able to accurately depict the fringes of human languages, i. e. those structures that are only rarely attested. Far too often, these rare structures are hidden in a heterogeneous waste-basket category of unclassifiable 'other' structures in typological surveys.

Admittedly, the search for, and study of, *rara* is methodologically difficult. There is no principled method for studying objects that are only rarely attested, except for using extremely large samples (which is normally too labor-intensive to be practically feasible). The only option seems to be to rely on serendipitously noted cases — either as a by-product of large-scale typological surveys or stemming from specific descriptions of mystifying phenomena encountered by specialists of a particular language. Starting from such a nucleus of known cases, the search for similar phenomena can be continued through checking closely related languages and areally close languages. Still, such a search for *rara* inevitably takes time, and the research will often span many years (or even decades) as a side-track of other research activities.

On the basis of the current knowledge about the diversity of human languages it remains infeasible to decide whether unattested structures are absolutely impossible or simply highly improbable. We presently "only" have some knowledge about a few thousand languages, and the variability of these languages is highly constrained by genealogical and areal cohesion. The fact that something is not attested among the sufficiently described world's languages might thus just as well be the result of historical coincidences instead of a sign of limits on the structural possibilities of human language.

Explicitly studying rarities will present a much more detailed picture of what is linguistically possible. An excellent example of the importance of studying *rara* for the understanding of the limits of the structure of human language is the paper on the interaction between gender and number by Plank and Schellinger (1997). They start from the well known Greenbergian (1963) typological universals 37 and 45, which state that gender distinctions in the plural imply gender distinctions in the singular. However, Plank and Schellinger show that – on closer inspection – a large set of "counterexamples" exists. Instead of considering such counterexamples nuisance elements that spoil an otherwise nice theory or generalization, Plank and Schellinger argue that these counterexamples be taken as opportunities: by collecting and interpreting such "exceptional" examples, a deeper and more accurate understanding of the possible variability of human language can be reached.

A different goal of the study of *rara* and *rarissima* is to argue against widespread assumptions about the limits of possibilities of human language. Either some generalizations had been proposed to which "counterexamples" are attested (like in the case of the correlation between genders and numbers discussed above), or some phenomenon that was deemed to be completely impossible is shown to exist after all. A prominent example of this kind of study is the survey of the labial flap by Olson and Hajek (2003). This sound, the only non-rhotic flap, has long been thought to be non-existent or at least not to be a distinctive phonological unit in any language. Yet, as Olson and Hajek (2003) showed, the labial flap exists in about 70 languages of Africa and one in Indonesia and in 22 of these languages the sound is indeed a distinctive unit contrasting with other bilabials.

Yet another possible use of *rara* is in tracking historical connections between languages. If any set of languages shares a rare or unique feature or even a bundle of "shared quirks", this is a strong indicator for a shared history of ancient contact or common descent, making these occurrences a useful diagnostic in diachronic linguistics and typology. This has e. g. been illustrated by Gensler (1994, 1997, 2003) by using different syntactic parameters and constructions as evidence for ancient language contact. For example, the syntagm S-AUX-O-V-OTHER can be reconstructed for Proto-Niger-Congo and is common all over the family. The same sequence is, however, basically unattested outside the family apart from half a dozen languages of Sudan into which it must have diffused.

In general though, the main question raised by the existence of *rara* is how to deal with them in theoretical approaches to language. The fact that *rara* exist – and even stronger, that the existence of *rara* as such does not seem to be exceptional at all – suggest that a theory of linguistic structure should have some principled notion of dealing with the existence of rare traits of human languages. Cysouw (2005: 248) estimates for person-marking syncretisms that even when taking the somewhat more widespread *rara* into account in a theory, there still are about 16 % of the world's languages that possess some structure which is rare. Each of these cases in itself is a *rarum*, but all together they make up a sizable portion of the world's linguistic structures. So, it does not suffice to simply dismiss any *rara* as incidental aberrations in the space-time of linguistic structure, as “exceptions” or “historical coincidences”. The real challenge is to develop theoretical notions for human language that inherently can deal with rarity and other types of variation.

At any rate, the terms *rarum* and *rarissimum* are used to refer to grammatical characteristics found only in very few languages, where the latter term would be referring to characteristics found in even fewer of the world’s languages. For a more tangible quantification, a threshold of attestations in $\leq 5\%$ of the world’s languages for rara and in $\leq 1\%$ of the world’s languages for rarissima has been discussed by Frerick (2006: 65–67), noting that such quantification is rather arbitrary. One must bear in mind that $\leq 1\%$ of about 7,000 languages still amounts to approximately 70 languages on a worldwide basis. And, given that the current world’s languages can be grouped into about 350 different genera (Dryer 2005), the criterium of non-genealogical clustering of *rara* would result in each fifth genus having a language with the *rarissimum* in question. From this perspective, even the $\leq 1\%$ criterium does not seem that unusual after all.

A different take on defining *rara* is to try and establish the stability of a linguistic phenomenon through time. The underlying rationale of Plank’s definition of *rara* (viz. *absolute rara* in the Bickel and Nichols sense) is that a *rarum* is a phenomenon that could very well arise in a particular language (after all, languages allow all kinds of strange things to happen), but when this happens it should not be for too long. The *rarum* should be an ‘instable’ characteristic and quickly change again into something else. Reformulating this idea as a dynamic process, it suggests that the possibility of ‘change away’ from a *rarum* to something else should be much greater than the probability of the *rarum* arising in the first place. As a measure of rarity one could then use the quotient of these probabilities. In contrast, at least some *relative rara* appear to be extremely stable and can even be traced back to ancestral languages, as noted e. g. by Harris (this volume: 98). This question suggests that the study of *rara* should be of great interest to the investigation of the dynamics of language change and vice versa.

Compared to the ongoing research tradition on language universals, investigations dealing with (rare) varieties only arose relatively recently. First and foremost there is “*das grammatische Raritätenkabinett: a leisurely collection to entertain and instruct*” already mentioned above, which has been edited and published online for more than a decade now by Frans Plank. This easily searchable database is a good starting point for any investigation into rare or infrequent structures of human languages.

Furthermore, in the same time frame in which the *Rara & Rarissima* conference and this volume were prepared, Horst Simon and Heike Wiese organized a session during the 27th annual meeting of the *Deutsche Gesellschaft für Sprachwissenschaft* in Cologne (DGfS Jahrestagung 2005), entitled "Expecting the Unexpected — Exceptions in Grammar". This session will also result in a collection of papers (Simon and Wiese (eds.) forthc.). Although the topic of exceptions is not necessarily the same as the study of rarities, there is still a good chance that rarities will be unexpected and occasionally even overlooked exceptions with respect to many theoretical proposals about the structure of human language.

3 Survey of this book

This book consists of various papers dealing with the theory and / or typology of *rara* among the world's languages. There is also a companion volume to the present book dealing with the details of rare and unusual structures in individual languages, namely "Rara & Rarissima: Documenting the fringes of linguistic diversity" (Wohlgemuth and Cysouw (eds.) 2010).

The current volume starts with two papers dealing with numeral systems among the world's languages, the first by HARALD HAMMARSTRÖM "Rarities in numeral systems" and the second by THOMAS HANKE "Additional rarities in the typology of numerals". Numeral systems have a long history of typological investigations (see the references in these papers), so this domain of linguistic structure is a prime example in which the study of *rara* can supplement known general tendencies with lesser-known minor tendencies.

The paper by ALICE HARRIS "Explaining typologically unusual structures: The role of probability" is the first of various papers in this volume dealing explicitly with the challenge that *rara* pose for theoretical consideration of language structure (see also the papers by Malchukov, Newmeyer, and Rijkhoff). Harris argues that *rara* are rare because it is unlikely for them to arise. Specifically, she illustrates this by rare phenomena that only arise through a combination of various diachronic steps. Each change individually is not necessarily special in any sense, but the combination of all diachronic requirements makes the end result unusual from a world-wide perspective.

Taking Plank's definition of *rara* seriously, the paper by PAVEL IOSAD "Right at the left edge: initial consonant mutations in the languages of the world” is not really about a *rarum*. As he shows, initial consonant mutation is incidentally found throughout the world’s languages, but it is also a general trait of the Celtic languages. Such a consistent distribution throughout all members of a genealogical group shows that although the trait might be unusual from a worldwide perspective, it is still a stable possibility for a language to portray and does not count as a real *rarum*. The paper by Iosad can thus be read as (implicitly) arguing that initial consonant mutation is not a *rarum* in Plank’s sense after all, but rather a *relative rarum* in Bickel and Nichols’ sense.

Various possible explanations for rarities and rareness are presented by Andrej Malchukov in his paper “Quirky case: Rare phenomena in case-marking and their implications for a theory of typological distributions”. Malchukov describes a few unusual phenomena related to case marking. These examples illustrate three different reasons why a phenomenon might be a *rarum*. First, a rare pattern may result from a conflict between a grammaticalization path and a functional constraint. Second, a pattern may be rare as it requires the co-occurrence of several different conditions (cf. Harris’ paper in this volume). And third, functionally deviant cases may result from incomplete grammaticalization cycles.

In his paper “Negatives without negators” Matti Miestamo takes up the challenge of a long-known typological (relative) *rarum*: the marking of negation by the absence of linguistic marking in some Dravidian languages. He compares the situation in such languages to the world-wide diversity of the marking of negation, pointing out various partial parallels in other languages. By combining the typological survey with the study of a *rarum*, Miestamo is able to make some sense of the otherwise rather puzzling negation structure in Dravidian.

The next two papers take the central question of *rara* head-on: how should *rara* be treated by theoretical notions of language structure? Frederick J. Newmeyer notes in his paper “Accounting for rare typological features in formal syntax: Three strategies and some general remarks” that rarities present a particular challenge for the Principles & Parameters approach to language, given the central idea of this approach that seeming complexity and idiosyncrasy are purely epiphenomenal. He argues that the existence of a rare feature is derivable from the interaction of processes known to be motivated in the grammars of the world’s languages.

Jan Rijkhoff in his paper “Rara and grammatical theory” discusses various *rara* in the domain of noun phrase structure in the context of *Functional*

Discourse Grammar. More generally, though, he argues that *rara* play a crucial role in the validation of claims made by any theory.

The question how to quantify the overall level of rarity of a language is taken up by SØREN WICHMANN and ERIC W. HOLMAN in their paper "Pairwise comparisons of typological profiles". Using the *World Atlas of Language Structures* and computing degrees of (typological) difference between two languages at a time, they investigate the relation between genealogical relationship and typological profiles of languages.

Finally, the paper by JAN WOHLGEMUTH "Some reflections on the interrelation of language endangerment, community size and typological rarity" investigates the influence of non-linguistic characteristics of a speaker community on *rara*. Specifically, he argues that there is a relation between the overall rarity of a language and its endangerment status.

Notes

1. http://typo.uni-konstanz.de/rara/intro/index.php?pt=1
2. Clicks were, however, also attested independently in the extinct speech register Damin of Lardil in Australia (cf. Hale 1998: 204 *passim*)
3. cf. Wohlgemuth (this volume)

References

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. http://www.uni-leipzig.de/~autotyp/download/enclaves@ALT5-2003BB-JN.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.

Dryer, Matthew
2005 *Genealogical Language List*. In: Martin Haspelmath, David Gil, Matthew S. Dryer and Bernard Comrie (eds.): *The World Atlas of Language Structures*. Oxford etc.: Oxford University Press, 584–643.

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

Gensler, Orin
1994 On reconstructing the syntagm S-Aux-O-V-Other to Proto-Niger-Congo. In *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistic Society, February 18-21, Special Session on Historical Issues in African Linguistics (BLS 20-S)*, Kevin E. Moore, David A. Peterson, Comfort Wentum (eds.), 1–20. University of California, Berkeley.

Gensler, Orin
1997 Grammaticalization, typology, and Niger-Congo word order: Progress on a still-unsolved problem. Review article on "Die Stellung von Verb und Objekt in Niger-Kongo-Sprachen", by Ulrike Claudi. *Journal of African Languages and Linguistics* 18.1: 57–93.

Gensler, Orin
2003 Shared quirks: A methodology for "non-orthodox" historical linguistics. Paper presented at the 17th International Congress of Linguists, Prague, 29 July 2003.

Greenberg, Joseph H.
1963 Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of Language*, Joseph H. Greenberg (eds.), 73–113. Cambridge, Mass.: MIT 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.

Harris, Alice C.
this volume Explaining typologically unusual structures: the role of probability. In *Rethinking Universals: How rarities affect linguistic theory*. Jan Wohlgemuth and Michael Cysouw (eds.), 91–103. (Empirical Approaches to Language Typology; 45). Berlin / New York: Mouton de Gruyter.

Olson, Kenneth S. and John Hajek
2003 Crosslinguistic insights on the labial flap. *Linguistic Typology* 7.2: 157–186.

Plank, Frans and Wolfgang Schellinger
1997 The uneven distribution of genders over numbers: Greenberg Nos. 37 and 45. *Linguistic Typology* 1.1: 53–101.

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

Simon, Horst and Heike Wiese (eds.)
forthc. *Exception in Language*. 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.

Wohlgemuth, Jan

this volume Language endangerment, community size and typological rarity. In *Rethinking Universals: How rarities affect linguistic theory*, Jan Wohlgemuth and Michael Cysouw (eds.), 255–277. (= Empirical Approaches to Language Typology; 45). Berlin / New York: Mouton de Gruyter.

Rarities in numeral systems

Harald Hammarström

1 Introduction

The paper surveys rarities in numeral systems across the world. Space permits us only to look at the most conspicuous kinds of rarities that are featured in the vast set of languages in the world. The study aims at a high level of preciseness as to what counts as a numeral and what counts as rare, and doubtful cases will be treated pre-emptively in footnotes.

2 Numerals

2.1 What are numerals?

In this paper, we define numerals as:

1. *spoken*
2. *normed expressions* that are used to denote the
3. *exact number* of objects for an
4. *open class of objects* in an
5. *open class of social situations* with
6. *the whole speech community* in question.

With the first point we mean to disregard symbol combination systems, e. g., Roman numerals, that are confined to written communication (but, of course, essentially all of our primary data come from written representations of the spoken language).

The second point serves to exclude expressions that also denote exact numbers, but are not the normal or neutral way to say those numbers, e. g., ‘eight-times-nine-and-another-two’ for the normal ‘seventy-four’, but also to demarcate the area where the numeral system ends, which is, when there aren’t any normed expressions.

As for the third point, languages usually have a rich set of expressions for inexact quantities, ‘a lot’, ‘few’, ‘really many’, ‘about fifty’ (but hardly *‘about fifty-one’) that have relatively high frequency in discourse. These are interesting in themselves but will not be included here because of their different fuzzy nature compared to exact number expressions.

Concerning the fourth point, some languages have special counting systems for a restricted class of objects (e. g. in Wuvulu (Hafford 1999: 37–39) for counting coconuts). These can be quite idiosyncratic and since all languages which have exact enumeration must have a means for counting an open class of objects, it is preferable to study that, as it corresponds to a general kind of communicative need of a society.

The reason for the fifth point, the requirement on social situations, is to take a stand on so-called body-tally systems (cf. Lean 1992: 2.4–2.6). A body-tally-system may be defined as follows. Assume a sequence of body parts beginning with the fingers of one hand continuing with some points along the lower and upper arm, reaching one or more points of the head, then ending with the corresponding body-parts on the opposite arm and finally hand. A number *n* is then denoted by the *n*th body-part-term in the sequence, e. g., ‘nose’ or ‘elbow on the other side’. There are features that distinguish body-tally systems from other counting systems with etymologies from body parts. Non-body-tally systems use only fingers, toes, hands, occasionally eye and head, whereas body-tally systems always use some intermediate points, such as elbow, shoulder or nose, and let them form a sequential order from one side of the body to the other. Typically, body-tally systems are only used in special circumstances, such as bridal price negotiations, and in other cases you would use a different numeral system or not use exact enumeration at all. The information on the social status of the body-tally numeral systems is very incomplete; We can say that for the vast majority we do not have such information, but for those in which we do, the social situation restriction applies. Body-tallying has to be done on a physically present person and to understand what number is referred to the process must be watched, so, for instance, body-tallying numerals would be infelicitous when it is dark. For instance, de Vries (1998) found that body-tally numerals in a Bible translation could not be understood, i. e., were often mis-translated back to Indonesian by bilingual persons. Of course, there could be some other language(s), unknown to us at present, where body-tally numerals can be used in a fully open class of social situations; such a body-tally system would accordingly be included in the study. Body-tally systems are attested in abundance in Papua New

Guinea and Indonesian Papua, in a geographically continuous area centered at the Ok family and, even if in decline, are still used today. Although many writers have neglected to mention it, there are also indisputable attestations of long extinct body-tally systems from Kulin (Pama-Nyungan, Australia) varieties in southeast Australia (Howitt 1889: 317–318, Howitt 1904: 697–703).

Finally, regarding the sixth point, we are not interested in numeral systems which are particular to some small subsets of the speakers of the language in question (e. g., professional mathematicians) because such systems might not respond to the conditions and needs of the majority of a society.

Numerals provide a good testing bed for patterns across languages given their comparatively clear semantics and modularity. As to numeral semantics, languages may differ as to which quantificational meanings they express / lexicalize, notably in approximate numeration and whether a counted set of objects constitute a group or not, but these matters are minor compared to differences languages show, e. g., in verbal tense / aspect. Likewise, although not universally, numerals tend to have uniform, clearly identifiable, syntactic behaviour within a language. Also, if two languages have exact numeration for a certain range of numbers, one expects the two to give a similar functional load to these expressions, excluding possibilities such as numbers also being used for, say, colours or as metaphors significantly wider in one language or the other. This appears sound also in the light of the only corpus study of numeral frequencies in a language with a restricted numeral system – McGregor (2004: 204) – which shows that ‘one’ and ‘two’ in Gooniyandi (Bunaban, Australia) occur with comparable frequency to ‘one’ and ‘two’ in English.

2.2 Rareness

In this paper we present cases that are rare, either in that (a) they are present in few languages or in that (b) they are present in few geographical spheres. Most cases are of the (a)-kind, but for example, base-12 systems in northern Nigeria are present in relatively many languages, from several different families, but are confined to just this geographical sphere, so they are counted as rare in the sense of (b) only. Geographically separate instances are likely to be independent, and the bottom line is that we are interested in rare independent innovations – whether or not they have grown genetically or areally onto many languages.

2.3 Survey

Lots of data is available in one form or another for numerals. It seems that numerals together with pronouns, kinship terms, body part terms, and other basic vocabulary (sun, water, etc), and perhaps "sketchy" phonological inventory, are the parts of language where there exists empirical data for a really large subset of the world's known languages. One may legitimately ask just how large this subset is when it comes to numerals – for how many languages do we have data on numerals? Let's say we count about 7,000 attested native spoken languages for the world. A definite lower bound is 3,880, since we can produce a list of references to numeral data from 3,880 definitely distinct languages. An upper bound is harder to give. We entertain the rather time-consuming methodology of trying to obtain every first-hand descriptive data reference found in any handbook or relevant publication whatsoever. The survey in this paper is based on the data we have collected so far. We currently have about 13,500 references, some describing numeral systems of many languages in the same publication, and, with 7,000 languages in the world, many different publications describe the same language. (The fact that often there is more than one independent source for one and the same language helps us to determine the accuracy.) It is impossible at this point to say how many languages the sources account for since they attest dialectal varieties, varieties from the same location but different centuries, partial data, data of varying quality, duplicated data, etc. However, at least one language from every attested language family or isolate is included in the survey (if numeral data is at all attested for the family in question).

In addition to first hand sources, we have also drawn inspiration from the rich existing literature on numerals in general. The subject, in fact, goes back more than 200 years in time — the first major work being the remarkable *Aritmetica Delle Nazioni* by Hervás y Panduro (1786). Since then, our bibliography counts some 20 doctoral dissertations, over 100 further monographs and more than 700 articles to have appeared. These range from purely descriptive accounts to areal, comparative-historical, typological, and deep syntactic studies — solely devoted to spoken language numerals as defined above. (The literature on written symbol systems for mathematics is even more voluminous.) However, since most of the literature just re-hashes the same data, the recourse to first-hand sources is essential in order to understand the true diversity in numerals in the world's languages.

3 Rarities

3.1 Rare bases

Perhaps the most salient single characteristic of a numeral system is its base, or more correctly speaking, its set of bases. The *set of bases* of a natural language numeral system may be defined as follows.

The number n is a base iff

1. the next higher base (or the end of the normed expressions) is a multiple of n; and
2. a proper majority of the expressions for numbers between n and the next higher base are formed by (a single) addition or subtraction of n or a multiple of n with expressions for numbers smaller than n.

This assumes that, for any expression, the linguist can unambiguously analyze each numeral expression into its constituent parts (or analyze it as consisting of only one part). As an example, for Swedish we would begin by finding the biggest part of the highest normed expression, which according to our own knowledge is *miljard* (10^9). Thereafter we can find the next lower base by trying divisors x of 10^9 to see if the numbers between x and 10^9 are expressed in the required form. For example, $x = 5 \cdot 10^8$ is not, because we do not say **en-halv-miljard plus ett* (*half-a-billion plus one) or the like for $5 \cdot 10^8 + 1$ or any, let alone a majority, of the numbers between $5 \cdot 10^8$ and 10^9. However, 'miljon' (10^6) fulfils the requirements, and, continuing with the same analysis for lower and lower numbers, we arrive at the conclusion that Swedish has $\{10, 10^2, 10^3, 10^6, 10^9\}$ as its set of bases.

The definition of base as stated gives unambiguous decisions for formations which are sometimes (and sometimes not) called base by other authors; systematic subtractions, special lexemes for base-multiples, or isolated cases of addition, e. g., only $7 = 6 + 1$ but otherwise no additions involving 6. Examples of such cases and their systematic resolution with our definition are given in Table 1 on the following page. It is important here to note that there doesn't have to be a monomorphemic word for something that is a base. In the case of Kare, at least if we assume that the numbers above 20 are formed parallel to 30, then 20 is a base. Further, 10 or 15 are not bases even though the words for them are monomorphemic — the definition interprets them as special words for multiples of 5, just like some base-10 systems have monomorphemic words for 20, 30, . . . , 90.

Table 1. Examples of formation types and outcomes of the definition of base (see text).

	Lutuami		**Nyokon**		**Kare**		**Ainu**	
	Klamath-Modoc, USA		Bantoid / Atlantic-Congo, Cameroon		Bantu / Atlantic-Congo, Sudan		Isolate, Japan	
	(Dixon and Kroeber 1907: 673)		(Richardson 1957: 30)		(Dijkmans 1974: 147)		(Refsing 1986: 110)	
	Analysis	Expression	Analysis	Expression	Analysis	Expression	Analysis	Expression
1	1	nas	1	ámɔ̀	1	emotí	1	sine
2	2	lap	2	àfɔ́ɔ̀	2	ibili	2	tu
3	3	ndan	3	átár	3	etotu	3	re
4	4	umit	4	ı̨́nnı̨̀s	4	biu	4	ine
5	5	tunip	5	ı̨́tɔ́ɔ̀r	5	etano	5	asikne
6	5+1	nas-ksapt	6	átʃı̨́n	5+1	etano na emoti	10-4	iwan
7	5+2	lap-ksapt	6+1	ı̨́tʃı̨́n námɔ̀	5+2	etano na ibili	10-3	arwan
8	5+3	ndan-ksapt	?	ı̨́yáá nı̨̀ màn	5+3	etano na etotu	10-2	tupesan
9	10-1	nas-xept	8+1	ı̨́yáá nı̨̀ màn námɔ̀	5+4	etano na bînu	10-1	sinepesan
10	10	te-unip	10	àwát	10	la-ato	10	wan
11	10+1	taunep-anta nas	10+1	àwát ámɔ̀	10+1	laäto na emoti	10+1	sine ikasma wan
...	...	...	...	...	...	...	...	...
15	...	...	...	...	15	sanga		
16	...	...	...	...	15+1	sanga-na-emoti		
...	...	...	...	...	...	...		
20	2x10	lap-eni taunep	20	nı̨̀tʃı̨́n	2x10	atumbili	20	hot
21	2x10+1	lap-eni taunep-anta nas	20+1	nı̨̀tʃı̨́n ámɔ̀	...	...	20+1	sine ikasma hot
...	...	...	...	...	...	...	...	...
30	3x10	nda-ni taunep	3x10	àwát átár	2x10+10	atumbili na laato	20+10	wan e tu hot
...	...	...	...	...	...	...	...	...
40	...	...	...	...	...	...	2x20	tu hot
Base	**5-10**		**10**		**5-20**		**5-10-20**	

The expression 'base-x system' will be used to mean that 'x is in the set of bases' for the numeral system in question. Similarly, 'base-x_1-…-x_n' system will mean that all of x_i is in the set of bases, without any commitment that the $x_1, \ldots, x_n$ exhaust the set of bases.

3.1.1 No base

There are a number of languages for which there is an explicit statement in the descriptive literature that they lack (exact) numerals above one.

Nadëb (Nadahup, Brazil):

According to Weir (1984: 103–104), the words for 2 and 3 are inexact. The vocabulary of a closely related variety lists completely different words for 1–3 (Schultz 1959) and the study by Münzel (1972) lacks information on numerals (cf. Epps 2006: 263). We have not seen the wordlist collected by Natterer (Koch-Grünberg 1906: 881), though this might not include numerals anyway.

Pre-contact Jarawara (Arawán, Brazil):

According to Dixon (2004: 559) and indeed the only other published wordlists for Jarawara (and closely related varieties) show some overlap between forms for 2, 3, 'few' and 'many' (Anonby and Anonby 2007: 25).

Pre-contact Yuqui (Tupi-Guaraní/Tupí, Bolivia):

According to Villafañe (2003: 68). As far as we are aware, there are no other published descriptions of this language that include the numerals.

Canela-Krahô (Jê/Jê-Jabutí, Brazil):

According to Green (1997: 181). However, an early vocabulary shows a restricted system (Kissenberth 1912: 54).

Krenák (Aimoré, Brazil):

According to a synthesis of earlier data by Loukotka (1955: 125–126) which follows observations such as Renault (1903: 1111). Even if there were no normed oral expressions, small numbers could be communicated using fingers on the hand (Ehrenreich 1887: 41–46).

Parintintin (Tupí-Guaraní/Tupí, Brazil):

According to Nimuendajú (1924: 240–241). Indeed, the larger dictionary by Betts (1981) agrees that the word frequently glossed as ‘two’ (cf. Sampaio 1997: 57–58) actually has an inexact meaning.

Wari’ (Chapacura-Wanham, Brazil):

According to one vocabulary collected by Hanke (1956). A later, more extensive, description of a variety in the same dialect cluster does show a word for ‘two’ albeit glossed literally as ‘facing each other’(Everett and Kern 1997: 452–459). An attempt at documentation of the most closely related language, the moribund Oro Win, failed to uncover any number words (Popky 1999: 38).

Chiquitano (Isolate, Bolivia):

According to Adam and Henry (1880: 19) which is corroborated by d’Orbigny (1839: 163) and Clark (1937: 118–119,138) and several later attestations of Chiquitano dialects show Spanish (Nordenskiöld 1911: 232, Nordenskiöld n.d.; Tormo 1993: 15, 108) or Portuguese (Santana 2005: 94) loans for ‘two’ and above. However, there are also dialects where a native term for ‘two’ is attested (Montaño Aragón 1989: 335–400).

“All” Campa and Machigenga groups (Pre-Andine/Arawak, Peru):

According to Wise and Riggle (1979: 88). As far as we are aware, published vocabularies (too many to list) show little indication that the words given for ‘two’ (and sometimes above) are in reality inexact. However, Wise and Riggle (1979) did work with basic mathematics education among these groups and therefore their judgement is arguably deeper.

Culina (Arawán, Peru):

According to Wise and Riggle (1979: 88). Unfortunately, we have not had access to other materials on either Brazilian or Peruvian Culina to double check the claim.

Arabela (Zaparoan, Peru):

According to Wise and Riggle (1979: 88), although the later, quite extensive dictionary of Rich (1999) does show distinct expressions for ‘two’ and ‘three’. Possibly, Wise and Riggle (1979) who did work with basic mathematics education looked at these expressions and their meaning more closely.

Achuar (Jivaroan, Ecuador):

According to Wise and Riggle (1979: 88), though later more extensive descriptions show expressions for ‘two’ and higher numerals (Fast and Fast 1981: 58–59; Fast et al. 1996). It is possible that expressions for ‘two’ and higher numerals crystallized as a result of increased contact with a counting culture (Gnerre 1986) or even reflects normative rather than descriptive usage. Therefore, Wise and Riggle (1979) who did work with basic mathematics, could very well be descriptively more accurate for the traditional state of the language.

Fuyuge (Goilalan, Papua New Guinea):

One early description of Fuyuge says that the ‘two’ word is also used for a small number (Ray 1912: 313–314). However, there is a word listed as ‘three’ but no explicit statement to the fact that this, like ‘two’, also has an inexact meaning. A very small vocabulary, probably collected by the same person lists 1, 2, $2+1$ and no further comments (Fastre 1920: 116), and the later, more modern description by Bradshaw (2007: 45) attests a native 1, 2, $2+1$, $2+2, \ldots$ system.

Viid (Border, Indonesia):

In one wordlist (a.2) of Viid from Senggi (Smits and Voorhoeve 1994: 211–212), ‘tambla’ is listed both with the meaning 2 and 3, but this is not borne out in other early wordlists (Smits and Voorhoeve 1994: 211–212) or the more recent (Menanti forthc.), which have $3 = 2+1$.

Gedaged (Oceanic/Austronesian, Papua New Guinea):

Nikolaj von Miklucho-Maclay, a pioneer researcher on the Rai-coast of Papua New Guinea, reports that (von der Gabelentz and Meyer 1882: 503):

> Sehr viele Papuas kennen die Zahlwörter ihres eigenen Dialektes nicht. In Mitebog [a village speaking a dialect of Gedaged – HH] fragte ich fünf oder sechs Eingeborene, aber die Angaben waren widersprechend und jedenfalls unrichtig, nur olam (eins) konnte ich als sicher notiren.
>
> [Very many Papuans do not know the numerals of their own dialect. In Mitebog I asked five or six natives, but the information given was contradictory and, in any case, erroneous, I could only note down olam (one) as certain.]

One interpretation of this statement is that there was no normed expression for numerals above ‘one’ in the lect of Mitebog. A later, longer description of a different dialect shows monomorphemic numerals 1–5 inherited from Austronesian (Dempwolff n. d.: 36–37),

To lack numerals above one means that the normed expressions for the quantities above one are inexact. We may call such systems 1-few-many for the time being. In these languages, it may be possible to communicate a higher exact quantity successfully, perhaps using gestures, context, one-to-one pairings, repetition or a specialized lexical item e. g., 'twin' for a certain kind of exact quantity. However, in these languages, the normed expressions are still 'one', 'a few', 'many', ... when these quantities occur in discourse. In no case does it appear to be possible, or normed, to say $few + 1$, $1 + 1$ or $few + few$ to designate an $exact$ number, so there is no base.

From the above cases, one certainly gets the impression that there is a thin line between 1-few-many systems and 1-2-many systems. In some cases, different observers on the same language variety differ as to whether the 'two'-word is approximate or exact in meaning. In other cases, the speech community seems to have acquired norms for number expressions over time. One may then conjecture that many more 1-few-many systems would have been found if more languages had been documented in detail before extensive contact with modern society.[1] It is also apparent that questions on this level of granularity are almost beyond the scope of classical forms of language documentation. Of languages potentially showing 1-few-many systems or 1-2-many systems only two, Mundurukú (Mundurukú / Tupí, Brazil; Pica et al. 2004) and Pirahã (see below), have been subject to investigations approaching standards of experimental psychology.

There are two further languages in the Amazon, Pirahã (Mura-Pirahã, Brazil) and Xilixana (Yanomama, Brazil) that stand apart from the above 1-few-many systems in that they are argued to lack all exact numerals, i. e., there is no normed way to denote an exact quantity even for 'one'.

In Pirahã, there are two words which prototypically mean 'one' and 'a couple' respectively, but it has been checked fairly extensively that their meanings are fuzzy 'one' and 'two' rather than discrete quantities (Everett 2005, 2004; Frank et al. 2008). It is not possible to combine or repeat them to denote higher (inexact?) quantities either (Gordon 2004). The Pirahã have the same cognitive capabilities as other humans and they are able to perform tasks which require discerning exact numeration up to the subitizing limit, i. e., about 3 (Gordon 2004). They just do not have normed expressions even for low quantities, and live their life happily without paying much attention to exact numbers. It does not appear to be possible to express an exact quan-