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 Klamath-Modoc, USA (Dixon and Kroeber 1907: 673)		**Nyokon** Bantoid / Atlantic-Congo, Cameroon (Richardson 1957: 30)		**Kare** Bantu / Atlantic-Congo, Sudan (Dijkmans 1974: 147)		**Ainu** Isolate, Japan (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	ɟnnǹ̩s	4	biu	4	ine
5	5	tunip	5	ɟtɔ́ɔ̀r	5	etano	5	asikne
6	5+1	nas-ksapt	6	átʃɟ́n	5+1	etano na emoti	10-4	iwan
7	5+2	lap-ksapt	6+1	ɟtʃɟ́n námɔ̀	5+2	etano na ibili	10-3	arwan
8	5+3	ndan-ksapt	?	ɟyáá nǹ̩ màn	5+3	etano na etotu	10-2	tupesan
9	10-1	nas-xept	8+1	ɟyáá nǹ̩ 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	nǹ̩tʃɟ́n	2x10	atumbili	20	hot
21	2x10+1	lap-eni taunep-anta nas	20+1	nǹ̩tʃɟ́n 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, $2+1$, $2+2$, ... 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-

tity simply by repeating an expression the appropriate number of times, like one can and often does in, e. g., Sanuma (Yanomama, Brazil) for 2 and 3 (Borgman 1990: 152). If one says "I'll be back after it gets dark and it gets dark again" this might just as well be interpreted as two days or as three days (p. c. Daniel L. Everett 2005). It seems relevant to note that Pirahã grammar lacks singular-plural distinctions of any kind, even in pronouns (p. c. Daniel L. Everett 2008). A wordlist of the only known relative of Pirahã, the extinct Mura language,[2] features words glossed 'one' and 'two' (Nimuendajú 1932; Nimuendajú and do Valle Bentes 1923). The 'one'-word is an obvious cognate to the Pirahã fuzzy one, and the 'two'-word is an obvious loan from some Tupi language.

Xilixana is the language of a group which has been on the Mucujai river at least for the past century. In modern divisions, it is sorted as a dialect of Ninam, also known as Yanam or Central Ninam (superseding Southern Ninam in older terminology) (Migliazza 1972). Swain (2000)[3] describes Xilixana numerals as not even having an exact 'one':

'one'	*mõli*	Note: Means 'one or a few'.
'two'	*kup*; *yalukup*	Note: Means 'two or a few'.
'three'	*pək*	Note: Can refer to any number more than two or a few.

John Peters, the first missionary to live among the same group, also describes the same expressions as having inexact value and adduces that "exact numbers were not important" (Peters 1998: 52). The closest other Yanomami variety for which there is a grammar is the dialect Shiriana, of the Uraricoera, to the north (Gómez 1990). This describes the numerals 'one' and 'two' as exact, but the author only spent 14 weeks in the field. Also Migliazza (1972: 117–118, 422), who spent many years in all of the Yanomama territory, describes Shiriana lower numerals as exact in the numerals section of his thesis and, in fact, all other description of Yanomama languages we have been able to consult describe 'one' and 'two' as exact (Ramirez 1994a, 1994b; Zerries and Schuster 1974; Becher 1960; Knobloch 1967; Vinci 1956; Wilbert 1962; de Matallana and de Armellada 1943; Koch-Grünberg 1928; Mattei-Müller 2007). Also, most Yanomama varieties have singular, dual and plural but we do not know the precise status of Xilixana, and if so, if they are inexact as well. However, on one page (Migliazza 1972: 38) the #moli word is glossed as 'one, few' (in contrast to pages 117–118 and 422). This is significant because language descriptions rarely claim 'one' and 'few' overlap in meaning,

and now three independent observers do it for the same or nearly the same language. Swain was a UFM / MEVA missionary who lived with the Xilixana for very long periods of time in the 1970–1990s and therefore she is certainly not a superficial observer. The Xilixana were monolingual (except for an occasional captured Dekwana) and uncontacted by modern society up to at least 1957 (Early and Peters 2000).

3.1.2 Base-3

Base-3 appears to be rarer than base-4. We have found only a few cases[4], some of them somewhat sporadic within their respective dialect cluster:

Ambulas of Wingei (Ndu, Papua New Guinea):

An Ambulas dialect survey (Wilson 1976: 57) mentions that the variety of Wingei counts in units of three, and the actual forms can be found in Wilson (1989a: 16–17). The forms are reproduced in Table 2 on the next page. Presumably, this is the same case that Laycock (1970) refers to when speaking (without forms given) of base-3-6-24 system(s) in the Ndu family, citing personal communication from Anthony Forge. The etymology of the forms reveal that the system is much like a commonplace 5-10-20 or 5-20 system except that the hand is seen as having six features! At the time of elicitation only older people knew the indigenous system, whereas the young used Tok Pisin or English for higher numerals. Other, better described, varieties of Ambulas (Wilson 1976, 1980) show no base-3 and comparative evidence shows that the original Ambulas (1-3) and Ndu (1-2) system were restricted (Aikhenvald 2008: 595; Laycock 1965: 173–174).

Waimirí of Atroarí (North Amazonian Carib / Cariban, Brazil):

Base-3 counting could be used up to about 9 according to Green (1997: 6–7), who cites personal communication with Ana Carla de Bruno Santos. However, the more recent grammar by Bruno (2003: 140–142) states that Portuguese loans are used above 3 and is silent about a possible base-3 alternative.

Som (Finisterre-Huon / Trans New Guinea, Papua New Guinea):

According to Smith (1988: 29) base-3 counting can be used up to about 9. We know of no other description of this variety.

Table 2. Numerals in Wingei Ambulas (Wilson 1989a: 16–17), Maprik Ambulas (Wilson 1980), Wosera-Mamu Ambulas from around Serangwandu (Wilson 1989b: 15) and Wosera-Kamu-K from around Kunjingini (Wilson 1990: 15).

	Wingei	Maprik	Wosera-Mamu	Wosera-Kamu-K
1	nawurak	nakurak	vétik	vétik
2	vétik	vétik	vétik	vétik
3	kupuk	kupuk	kupuk	kupuk
4	kupukiva	nakwasa/wan vétik wan vétik	vétik vétik	vétik vétik
5	kupuk'etik	naktaba	taambak	taambak
6	taabak	naktaba sékét naktaba nakurak		
7	taabak kaayek	naktaba sékét naktaba vétik		
8	taabak kaayek vétik	naktaba sékét naktaba kupuk		
9	taabak kaayek kupik	naktaba sékét naktaba kupuk wan vétik wan vétik		
10	vétik taaba vétik	taaba vétik		
11	nawurak taaba vétik	taaba vétik sékérék maan-ba kayék nakurak		
12	taaba vétik	taaba vétik sékérék maan-ba kayék vétik		
20		maan vétik taava vétik	nakurak mi	nakurak dumi / maan vétik taaba vétik
24	nawura mi			

Etymologies of roots are as follows #maa is 'foot, leg', #taaba is 'hand, arm', #mi is 'tree' and #du is 'man'. Apparently, in Wingei counting, the hand is seen to have six features. The etymology of the expresssion *nawura mi* / *nakurak mi* / *nakurak dumi* is not clear but it may have to do with either tree (typologically unusual but matches *mi*) or man (typologically very common, but resembles only *dumi*).

Bine (Eastern Trans-Fly, Papua New Guinea):

In at least one vocabulary repoduced in Wolfers (1972: 218) and Wolfers (1971: 79), a variety of Bine is base-3 and reaches up to 9. However, all other attestations of Bine show only a restricted system and / or a body-tally system (Lean 1986d), including the lengthiest description (Fleischmann and Turpeinen 1975: 16). The base-3 vocabulary must therefore be considered somewhat dubious.

Bukiyip (Arapeshan[5], Papua New Guinea):

Fortune (1942: 58–60) describes the Rohwim dialect of Mountain Arapesh to have a base-3 system for counting some objects and a base-4 system for counting other objects, which seems to have reached up to 24. A later description of an inland Bukiyip (Conrad and Wogiga 1991: 73–76) variety shows a conflation of the two systems (with no indication of them being used for different objects). Conrad submitted the base-3 system for the entry on Bukiyip (dialect not indicated) on the *Numeral Systems of the World's Languages* website.[6] Available data on other Arapeshan languages, such as Abu' Arapesh (Nekitel 1985: 82–84) and Mufian (Conrad et al. 1978: 104), show base-5, at least from 7 and up.

3.1.3 Base-4

Base-4 systems are attested on four continents:

North America:

Some extinct Chumash languages (Chumashan, USA) show original base-4 systems, running up to 32 (Beeler 1967, 1963; Hughes 1974; Mamet 2005: 113–115). Base-4-8 is also documented with the older generation in the now extinct Yuki (Isolate, USA). For Yuki, Kroeber (1925) describes how base-4 is related to hand-counting by considering the spaces between the fingers (cf. Hinton 1994[7]). The Chumashan languages and Yuki are both in California but quite distantly apart, with Yuki in the north and Chumashan in the south, and other language families intervening.

South America:

The extinct Lule (Isolate, Argentina) of Clark (1937: 102) and Machoni de Cerdeña (1732: 84–86) as well as the poorly attested extinct Charrúa

(Charruan, Uruguay) reported in Ibarra Grasso (1939b: 202) appear to have had base-4 up to 10, at which point the system turns into a commonplace 5-10-20 system with hands and feet. It cannot be inferred from the data at hand that there was ever a true base-4 system here, beyond 10.

A couple of descriptions of a Guaraní variety in Paraguay (Tupí-Guaraní / Tupi, Paraguay) show base-4 up to 10, but the expressions for numbers above 10 are not shown (Ibarra Grasso 1938: 278, 1939a: 590). Other old and new descriptions of any varieties of Guaraní (too many to list) do not show any traces of base-4. Isolated vocabularies of Mocovi and Toba (Guaicuruan, Argentina) show base-4 up to 8 and 10 respectively (Koch-Grünberg 1903: 114–124), but the vast majority of vocabularies for these languages (too many to list) show no trace of this.

The extinct Payaguá (Isolate[8], Paraguay) has one attestation with alternative base-4 forms up to 20 (Koch-Grünberg 1903: 114–124). All these cases occur within a relatively small area of South America, but there is otherwise little evidence for an areal connection.

Oceania:

An indeterminate number of languages in the New Guinea highlands have a variations of a base-4 system (Lean 1986a: 13–86, 1986c: 15–59, 1992: Ch. 5), where at least one, Kakoli (Hagen / Trans New Guinea, Papua New Guinea) is attested with base 4-24 (Bowers and Lepi 1975). Kewa (Engan / Trans New Guinea, Papua New Guinea) has several parallel numeral systems, one of them being base-4 (Franklin and Franklin 1962) and goes at least up to 20, and beyond that it may be combined with a body-tally system to form higher numbers in units of four (Pumuge 1975). The word for '4' is 'hand', i. e., four fingers constitute one hand and the thumb is separate. The traditional counting system of Mbowamb (Hagen / Trans New Guinea, Papua New Guinea) near Mt. Hagen has been been described with a fair amount of detail. It is clearly a 2-4-8 system, for which Vicedom and Tischner (1948: 268–270) give expressions up to 24, and say the system can be used up to about 80. Another description seems to indicate that after 20, counting can be done in units of 20 (Strauss 1962: 315–318), cf. also Lancy and Strathern (1981). As in Kewa, the base-4 is connected with counting the fingers of one hand, the thumb counted separately. The origin of the highland base-4 system(s) has not been systematically investigated, but given the geographical proximity and the fact that the Engan and Hagen languages are not closely related,

an areal connection seems likely even if this is not directly observable in the forms in question.

On the north coast, around the border between Indonesian Papua and Papua New Guinea, base-4 is also present variously in the Sko languages (most of the data is collected in in Lean (1986b), but see Donohue (2008) for a good attestation of 4-12-24 in Skou) as well as 4-24 in Tobati (Sarmi-Jayapura Bay / Austronesian) for which the best attestation is Moolenburgh (1904). Given the proximity of the languages and the fact that they are genetically unrelated, there is almost certainly an areal connection between base-4 in Skou and the Sarmi-Jayapura Bay Oceanic languages.

Africa:

An indeterminate number of languages in the northeastern Democratic Republic of the Congo (DRC) have (traces of) a base-4 system. The first attestation appears to be a Nyali (Bantu, DRC) variety for which Stuhlmann (1894: 624) notes that $8 = 2*4$, $9 = 2*4+1$, $13 = 12+1$, $14 = 12+2$, $16 = 2*8$, $17 = 2*8+1$ but $20 = 2*10$. Later reports of related Bantu varieties show that there was / is a fully systematic 4-24 or 4-32 underlying these forms (van Geluwe 1960; Kalunga Mwela-Ubi 1999; Bokula and Ngandi 1985). Furthermore, thanks to Kutsch Lojenga (1994a: 353–357), we have a full attestation of almost obsolete Ngiti (Lendu / Central Sudanic, DRC) and Lendu (Lendu / Central Sudanic, DRC) 4-32 systems (p. c. Constance Kutsch Lojenga 2007). Various wordlists attest traces of the same base-4 systems in decay or amalgamation with base-10 and base-20 in closely related Bantu and Central Sudanic languages (Johnston 1922b; Struck 1910; Johnston 1904; Bokula 1970; Harries 1959; Kutsch Lojenga 1994b; Schebesta 1966, 1934; Asangama 1983; Czekanowski 1924; Stuhlmann 1917; and unpublished SIL survey lists).

Non-cases

In addition, there are a number of languages which have been claimed to be base-4 in the literature but which are not base-4 according to the definition used in this paper. We will mention a few of the most important ones here. The language called Āfúdu (Unassigned[9], West Africa) by Koelle (1854) uses some additions with 4 in the numbers below 10 but is decimal in the range 10–20. Bodo and Deuri (Bodo-Garo / Sino-Tibetan, India) have vestiges of base-4 counting extending higher than 20 and Bai (Bai / Sino-Tibetan, China) is documented with a base-4-16-80 system for shell money in medieval times

(Mazaudon 2007). Yiwom (West Chadic A / Afro-Asiatic, Nigeria) has 7–9 as $4+3$, $4+4$, $4+5$ but no other forms are based on 4 (Ibriszimow 1988). De Castelnau (1851a: 10–13) reports base-4 (actually base-2-4) in Apinayé (Jê / Jê-Jabutí, Brazil) but no actual forms are given (de Castelnau 1851b: 270–274) and is likely to be spurious in the absence of corroborating data in this rather well-documented language (too many references to list). Base-4 for counting special objects is widely attested in the Oceanic languages of Melanesia (Kolia 1975; Friederici 1912; Parkinson 1907).

3.1.4 Base-6

Base-6 systems are attested on Kolopom Island (formerly Frederik-Hendrik-Eiland) in southwest Indonesian Papua, as well as in the Kanum and Nambu languages in southern New Guinea around the Indonesian-Papua New Guinea border. Their origins have been discussed extensively (Donohue 2008; Evans 2009; Hammarström 2009; Plank 2009) and need not be repeated here.

In addition, there are a number of languages which have been claimed to be base-6 in the literature but which are not base-6 according to the definition used in this paper (cf. Plank 2009; Gamble 1980; Beeler 1961; Ibarra Grasso 1939b). A few require comment. One early attestation of Balanta (Northern Atlantic / Atlantic-Congo, Senegal / Guinea Bissau) has additions of 6 for the numbers 7–12 (Koelle 1854). But since we do not know the continuation beyond 12, it is unsure whether the 6:s generalize (cf. Wilson 1961a). Also, later attestations give different, non-base-6, forms (Wilson 1961b; Quintina 1961; Fudeman 1999). Similarly, Less Traditional Tiwi (Isolate, Australia) may have formed some numbers in the range 7–10 with 6 (Lee 1987: 96–100), but not further.

3.1.5 Base-8

Northern Pame (Otopamean / Otomanguean, Mexico), the sole case of a base-8 language (attested up to 32) which does not have 4 as a sub-base is presented and discussed in Avelino (2006), though 5–8 have etymologies which involve 5.

3.1.6 Base-12

Dhivehi (Indo-Aryan / Indo-European, Maldives) has an early attested (Gray 1878) but long extinct base-12 which is attested up to 96 thanks to the ef-

forts of Fritz (2002: 107–123).[10] Apart from that case, there are base-12 systems in the Plateau area of northern Nigeria. The first known attestations of such systems[11] come from the famous *Polyglotta Africana* by Koelle (1854) which includes numerals 1–20 in a number of West African languages and the first proclamation of duodecimality as a system appears to be Schubert's (1888). As shown in Table 3, we have tried to collect all independent attestations that have been published, or, are unpublished but available on the internet.[12] However, not all of them are necessarily independent as this information is not always deducible from the text. It is likely that there are a few more attestations in publications that we do not have access to. For many, if not all, other sources on the same varieties attest base-10 rather than base-12, which means that the base-12 systems are currently under pressure.

Table 3 shows published attestation of base-12 systems in the Plateau area. $12-144$ means that the attestation gives forms $\leq$ 12, forms $12+x$, multiples of 12, and a word for 144; 12+ means forms $\leq$ 12 and forms $12+x$ or multiples of 12; ≤ 12 means forms ≤ 12; "12" means that the source simply states that there was a "duodecimal system" but gives no forms; Cont.-10 means an attested 10-system contaminated by forms following a "duodecimal system" and Spec.-12 means that some duodecimal connection is speculated. Further half-attestations are as follows. Arago (base-10 in Judd 1923), Kagoma and Agatu were judged "uncertain" by Thomas (1920a). Gwara, a Margi variety (Biu-Mandara A / Afro-Asiatic, Nigeria) has monomorphemic 1–10 and forms 11–12 with formations that may include 1 and 2 – a bit like Germanic – but there is otherwise no reason to suspect base-12 counting (Wolff 1975).

Table 3. Published attestation of base-12 systems in the Plateau area.

Language	Source	Type	Family	Comment
Ake	Blench 2006a	≤ 12	Plateau	
Afo	Bouquiaux 1962	"12"	Plateau	
Afo (Apho)	Bouquiaux 1962	"12"	Plateau	
Afo (extinct Afu)	Thomas 1920a	"12"	Plateau	
Afo	Meek 1925: 142–143	12+	Plateau	
Afo (Eloyi)	Mackay 1964; Armstrong 1983	12+	Plateau	
Aten	Blench 2006d	≤ 12	Plateau	
Aten (Ganawuri)	Bouquiaux 1964, 1962	12–144	Plateau	
Aten (Ganawuri)	Meek 1925: 142–143	12+	Plateau	

Continued on next page

Language	Source	Type	Family	Comment
Birom	Bouquiaux 1970	12–144	Plateau	
Birom	Thomas 1920b	"12"	Plateau	
Birom (Tahoss)	Blench 2006g	≤ 12	Plateau	
Che (Rukuba)	Gerhardt 1987	Spec.-12	Plateau	Cites BCCWL.
Che (Rukuba)	Blench et al. 2006	≤ 12	Plateau	
Eggon	Blench and Hepburn 2006	≤ 12	Plateau	
Eggon	Gerhardt 1983: 47	"12"	Plateau	
Eggon	Gerhardt 1987	"12"	Plateau	Cites Gospel 1935 + Lukas 1952 field-notes
Eggon	Shimizu 1975	"12"	Plateau	
Hyam	de Castelnau 1851c: 59	≤ 12	Plateau	
Hyam (Jaba-Kwoi)	Meek 1931: 123	12–144	Plateau	Also base-10 forms
Hyam (Jaba)	Bouquiaux 1962	"12"	Plateau	
Hyam	Thomas 1920b	≤ 12	Plateau	
Hyam	Blench 2006f	≤ 12	Plateau	
Ikulu	Seitz 1993: 37–38	Spec.-12	Plateau	
Izere (Fobur)	Blench and Kaze 2006	≤ 12	Plateau	
Izere (Ganang)	Blench 2006c	≤ 12	Plateau	
Izere (Zarek-Gana)	Gerhardt 1987	"12"	Plateau	Citing BCCWL
Kaningkom	Gerhardt 1987	"12"	Plateau	
Koro	Thomas 1920b	12+	Plateau	
Koro	Williamson 1973: 453	12+	Plateau	
Koro (Idũ)	Blench 2009a	12+	Plateau	
Koro (Nyankpa)	Thomas 1920b; Gerhardt 2005; Blench 2009b	12+	Plateau	
Koro (Tinɔr)	Gerhardt 1973	"12"	Plateau	
Koro (Tinɔr)	Blench 2009c	≤ 12	Plateau	
Lungu	Gerhardt 1987	"12"	Plateau	
Mada	Blench and Kato 2006	≤ 12	Plateau	
Mada	Thomas 1920a	"12"	Plateau	
Mada (S. Mada)	Mathews 1917	12–144	Plateau	
Ninkyop	Blench 2006e	≤ 12	Plateau	
Ninzam	Mathews 1917	12–144	Plateau	
Ninzam	Thomas 1920a	"12"	Plateau	
Nungu	Mathews 1917	12–144	Plateau	
Nungu	Thomas 1920a	"12"	Plateau	
Rigwe	Bouquiaux 1962	"12"	Plateau	
Rigwe	Gerhardt 1987	"12"	Plateau	
Rigwe	Gerhardt 1969: 125–127	≤ 12	Plateau	

Continued on next page

Language	Source	Type	Family	Comment
Teria (Cara)	Blench 2006b	≤ 12	Plateau	
Teria / Fachara	Meek 1925: 142–143	12+	Plateau	
Tesu	Blench 2006f, 2006h	≤ 12	Plateau	
Tyap (Gworok)	Adwiraah 1989	"12"	Plateau	
Tyap (Gworok)	Gerhardt 1987	≤ 12	Plateau	Not confirmed in Gerhardt 1968
Amo	Luzio 1973	Cont.-10	E. Kainji	
Gure	Meek 1931: 203	≤ 12	E. Kainji	
Iguta	Shimizu 1979	12+	E. Kainji	
Janji	Meek 1931: 185–187	≤ 12	E. Kainji	
Janji	Shimizu 1979	≤ 12	E. Kainji	
Janji	Bouquiaux 1962	"12"	E. Kainji	
Jere	Shimizu 1982	≤ 12	E. Kainji	Not Sheni, Ziriya, Gana, Taura, Shau, Gyem, Gamo
Jere (Boze, Akwẹrẹ clan)	Nengel n. d., 1999	≤ 12	E. Kainji	
Kahugu	Meek 1931: 212	≤ 12	E. Kainji	
Lemoro	Shimizu 1979	≤ 12	E. Kainji	Not Cokobo
Piti	Meek 1931: 139	12+	E. Kainji	Switched to base-10
Piti	Matsushita 1998	"12"	E. Kainji	
Rop	Meek 1925: 142–143	12+	E. Kainji	
Sanga	Shimizu 1979	≤ 12	E. Kainji	
Dyarim	Blench 2007	Spec.-12	W. Chadic	Etymological Connection
Gwandara	Shimizu 1975	"12"	W. Chadic	Citing P. Newman p. c.
Gwandara (Nimbia)	Matsushita 1998	12–144	W. Chadic	
Mwaghvul	Jungraithmayr 1963	12+	W. Chadic	
Ron of Daffo	Seibert 1998	12+	W. Chadic	Not confirmed in Jungraithmayr 1970
Mumuye	Matsushita 1998	"12"	Adamawa	Not Zing Mumuye pace Shimizu 1983
Mama (Kantana)	Gerhardt 1987	"12"	Jarawan Bantu	
Mama	Thomas 1927	≤ 12	Jarawan Bantu	
Mama	Mathews 1917	12–144	Jarawan Bantu	
Mama	Thomas 1920a	"12"	Jarawan Bantu	

The base-12 systems occur only in languages in the area of Jos plateau of Nigeria, but which belong to different (sub-)families, namely Plateau (Atlantic-Congo), East Kainji (Atlantic-Congo), West Chadic (Afro-Asiatic), Adamawa (Atlantic-Congo) and Jarawan Bantu (Atlantic-Congo). A root resembling *#sok* for 12, with plausible sound correspondences (Gerhardt reconstructs **suak*), is widespread in Plateau, wherefore it is very likely that base-12 is old in Plateau. The same root occurs in Jarawan Bantu and Ron of Daffo, both of which are isolated instances of this root, or indeed base-12, in their respective families, so borrowing from (proto-southwest) Plateau is highly likely if not certain, as concluded by Maddieson and Williamson (1975: 136) and Gerhardt (1997: 140–141) for Jarawan Bantu. In East Kainji and the Beromic subgroup of Plateau, a root *#kuri* occurs for 12, which makes a borrowing in either direction likely. Furthermore, *#piri* is 12 in Gure and Kahugu (East Kainji) and *#zowa* is 12 in Ake and Koro (Plateau) and yet other roots for 12 appear in the remaining West Chadic cases. Since base-12 is so rare in the languages of the world, the variety of non-ancient roots suggest that a base-12 *system* may be borrowed even without key morphemes. The root for 12 in the alleged Mumuye variety with base-12 is not known.

There are no obvious clues as to the unusual choice of 12 as a base. A few of the base-12 languages in Meek (1931) have hand gestures that often are used accompanying the spoken expression. A combination of fingers and eyes make up 12 in at least one of these cases, but no traces of words meaning ‘eye’, ‘hand’ or ‘finger’ can be found in the corresponding spoken expressions. On the other hand, although not a base, 12 bears a special position in several modern European languages too, with a special word like ‘dozen’ and an elevated frequency (Dehaene and Mehler 1992). The reason(s) for this is not well-understood either.

3.1.7 Base-15

There appears to be only one case of a language attested as base-15, at least for a number of decades, namely Huli (East New Guinea Highlands / Trans New Guinea, Papua New Guinea) of the southern highland fringes. It is clearly an original body-tally system with a cycle of 29 – midway / centerpoint is thus 15 – which under influence from a Tok Pisin base-system turned into base-15 (Cheetham 1978; Lomas 1988).

3.1.8 Rare second bases

Some rarities in the next higher bases after 5, 10 or 20 are as follows:

10–40:

Pech (Paya / Chibchan, Honduras) as of Conzemius (1928: 264–265) and Hawaiian (Oceanic / Austronesian, USA) until it restructured to 10–100 under foreign pressure (von Chamisso 1837; Dwight 1848; Hughes 1982).

5-20-40:

Southwestern Pomo (Pomoan, USA) in one attestation (Closs 1986: 35–41).

10-60:

Attested (Drabbe 1952) in Ekagi (Paniai Lakes / Trans New Guinea, Indonesia) and Ntomba (Bantu / Atlantic-Congo, DRC) until it restructured to 10–100 under foreign pressure (Gilliard 1928, 1924).

5-10-20-60:

Famously known from the long extinct Sumerian (Isolate, Iraq), see, e. g. Powell (1972).

(5-)10-20-(60 /)80:

Attested in Mande (Monteil 1905; Dombrowski and Dombrowski 1991; Delafosse 1928; Hartner 1943), Dogon (Calame-Griaule 1968), Gur (Carlson 1994; Welmers 1950: 167–169) and Bangi Me (Blench 2005) languages in a relatively small area in West Africa, wherefore an areal connection is almost certain. In the Mande attestations, the systems vary between 60 and 80 as per a certain root that sometimes means 60 and sometimes 80.

5-25:

Gumatj (Yolngu / Pama-Nyungan, Australia) is described, with ample examples, to be 5-25 (up to 625). However, one would not usually use exact numbers for counting this high in this language and there is a certain likelihood that the system was extended this high only at the time of elicitation with one single speaker (Harris 1982; Sobek p. c.).

At least one speaker of Biwat (Yuat River, Papua New Guinea) appears to have made the same 5-25 innovation (McElvenny 2006), as two other earlier

attestations rather show a commonplace 5-20 system (Haberland and Seyfarth 1974; Mead 1932).[13]

It is remarkable that there is no incontestable attestation of a 5-25 system that extends to a whole speech community.[14] The contrast with 5-20 systems, which are ubiquitous, reveals much as to the evolution of normed number expression within a community.

3.1.9 Last notes

At least two cases of alleged base-11 exist, both of which appear to be mistaken. Pañgwa (Bantu / Atlantic-Congo, Tanzania) is presented with a base-11 vocabulary (Johnston 1922a: 477), but this cannot be corroborated in other attestations (Stirnimann 1983) so it is presumably an error.

A fairly early discussion of Māori (Balbi 1826: 256–257) likewise claims undecimality, but this was refuted already in the same century (Conant 1896: 122–123). One alleged case of counting in 30s is in Klingenheben (1927: 43) but this too has failed to be corroborated later.

3.2 Other rarities

Other than rare bases, there are a few very interesting rarities which we mention below.

3.2.1 Streak of unanalyzable forms

Several, but not all, of the base-12 languages have monomorphemic words for all of 1–12 as does, e. g., Chalchihuitán Tzotzil (Mayan, Mexico) (Hopkins 1967: 16). However, the record streak appears to be 15, as evidenced in Chocho of Santa Catarina Ocotlán (Popolocan / Oto-Manguean, Mexico) in Table 4 on the next page.[15]

A claim of monomorphemic 1–20 in Munda (subfamily of Austroasiatic, India) appears, on closer scrutiny, to be artificial or unsubstantiated.[16]

3.2.2 Order of additive units

As we have seen, all languages which have numerals above 20 form the higher numbers using addition and multiplication of integers (and occasion-

Table 4. The monomorphemic numerals up to 15 in Chocho of Santa Catarina Ocotlán. 15–19 are formed as 15+1 etc and 20 is a base (Veerman-Leichsenring 2000: 33–34), cf. also Mock (1977: 153–154).

1 ngū	6 šu̱	11 tǫ́
2 žú	7 žàadù	12 rxá
3 nīé	8 šį́	13 šé
4 ňu̱ų́	9 nīà	14 rxǫ̀
5 žų́	10 tè	15 rxǫ̀ʔ

ally subtraction as well multiplication with fractions). Both addition and subtraction are commutative operations so languages are free to change the order of the operands. Not surprisingly, the order of multiplier and multiplicand is usually the same as the order of numeral and noun in the language in question. For additive units the situation is more interesting. For expressions where the sum is less than, say, a 100, we find both smaller-precedes-larger and larger-precedes-smaller in the languages of the world. A lot of languages have one order for the teens and the opposite order for higher sums. For sums above 100, the situation is quite different. Almost all languages, and a multitude of the cases must be independent, show larger-precedes-smaller order. At least three ancient languages[17] – Classical Attic Greek, Classical Arabic, Sanskrit (as well as Vedic) – are attested with both orders possible. The only modern languages with invariable smaller-bigger order between additive units in numeral expressions ≥ 100 appear to be (certain dialects of) Malagasy (Barito / Austronesian, Madagascar), Chuj (Mayan, Guatemala) and Tzotzil (Mayan, Mexico),[18] see Daval-Markussen et al. (in press) for references.

3.2.3 Cardinal dominance?

In natural languages, it appears that cardinal numerals hold a primary position over other kinds of numerals, e. g., distributive numerals, and exact number marking in general, in the sense that the non-cardinals are morphosyntactically derived from the cardinals and that the cardinals run higher. The dominance appears to be exceptionless for all languages which have numerals above 3, but we will review two interesting challenges below.

One description of a Great Andamanese variety explicitly says that there are more ordinals than cardinals (Man 1883a: 100), or – to be more specific – that there are only two cardinals but six ordinals. But a closer inspection of

the forms reveals that the six "ordinals" are not true ordinals. 3–6 do not mean 'third'–'sixth' but 'in the middle', 'the next one', 'last' and so on. They only acquire the fixed ordinal meaning in the context of a game or the like when the number of participants is known (Man 1883b: 413).

One description (Mathews 1904) of Wuddyāwūrru (West Victoria / Pama-Nyungan, Australia) says that there are more grammatical numbers (singular, plural, *trial*, and plural) than cardinals (one, two). This is not contradicted by other sources on the same or related languages (too many to list). However, there is no linguistic data in this case to ascertain that the trial was a true trial (rather than a paucal) and Mathews has described many other Australian languages as having trials where this is questionable (p. c. Barry Blake 2005). We will never know for sure whether this language had a true trial or not, since the language is extinct.

4 Conclusion

This paper has surveyed rarities for a number of structural properties of numeral systems. We have given full primacy to data presentation rather than interpretation to make the factual status of the data maximally clear. With this, we hope to have set the stage for future generalizations and interpretations of rareness with a high level of empirical validity.

Acknowledgments

We wish to thank the following libraries for granting access and services: Centralbiblioteket (Göteborg), Institutionen för orientaliska och afrikanska språk (Göteborg), Etnografiska Muséet (Göteborg), LAI (Göteborg), Carolina Rediviva (Uppsala), NAI (Uppsala), Karin Boye (Uppsala), KB (Stockholm), SUB (Stockholm), LAI (Stockholm), Universiteitsbibliotheek (Leiden), KITLV (Leiden), Universiteitsbibliotheek (Amsterdam), Institute for Asian and African Studies (Helsinki), MPI-EVA (Leipzig), Universitätsbibliothek (Leipzig), Butler / Columbia University (New York City), IfA (Cologne), BNF (Paris), SOAS (London), ILPGA (Paris), ZAS (Zürich) and Völkerkundliche Bibliothek (Frankfurt).

We are also indebted to (in no particular order) Hein van der Voort, Lincoln Almir Amarante Ribeiro, Eduardo Rivail Ribeiro, Michael Cysouw, Nathan Hill, Jesús Mario Girón, Karsten Legère, Helene Fatima Idris, Ber-

nard Comrie, Pedro Viegas Barros, Lionel M. Bender, John Kalespi, Hilário de Sousa, Frank Seidel, Tom Güldemann, Lourens de Vries, Ian Tupper, Johanna Fenton, Randy Lebold, Willem Adelaar, Lyle Campbell, Norbert Cyffer, Maarten Mous, Thilo Schadeberg, Raoul Zamponi, Paul Whitehouse, Swintha Danielsen, Lauren Campbell, Dmitry Idiatov, Nick Evans, Matthew Dryer, Mark Donohue, Roger Blench and Peter Bakker for help with access to data. The bibliographies by Alain Fabre (for South America) and Jouni Filip Maho (for Africa) have been very helpful in bibliographical searching leading up to this study. William McGregor kindly corrected a mis-identification by me of Gumatj.

Notes

1. Further cases may include the following. **1.** Aikhenvald and Dixon (1999: 358) conjecture that Djeoromitxí (Jabutí / Jê-Jabutí, Brazil) "originally had no numbers" since the word *je-bo* for 'two' given by Pires (1992: 66) is from a root with the meaning to 'be equal'. However, an etymology for 'two', even if correct (cf. van der Voort 2004: 212; 2007: 162) does not automatically mean that there was no original word for 'two', nor that a present meaning of 'two' (Ribeiro 2008: 42) is somehow subordinate to the etymological meaning. Also, an early attestation of *yawo yawo* $(2+2)$ for 4 in (Loukotka 1963: 50) speaks against an inexact meaning for 'two'. **2.** Barriga Puente (1998: 132, 263) reports that Esmeraldeño (Isolate, Ecuador) has a limit of counting at one, based on a misreading of Lehmann (1920: 37). There is only one vocabulary of the now extinct Esmeraldeño which has been reprinted a number of times (Adelaar 2004: 155–161). However, the earliest of these publications (Wolf 1892: 528) is clear that the lack of native Esmeraldeño words above one could be due to the memory of the last speaker(s). So we are not in a position to assert that Esmeraldeẽno ever lacked numerals above one. **3.** Koch-Grünberg (1928: 316) describes numerals in Sapé (Isolate, Venezuela) as 1 'me̥yakán' and 2 'me̥yakán' and remarks

> "Die Kaliána haben nur ein Zahlwort und gebrauchen stets denselben Ausdruck, in dem sie dabei an den Fingern und dann an den Zehen weiterzählen".
>
> [The Kaliána have but one numeral and always use the same expression while they continue to count using their fingers and toes.]

It's not clear what to make of this, but, in any case, the only other two published vocabularies do show distinct words for lower numerals (de Matallana and de Armellada 1943; Migliazza 1978) and Koch-Grünberg's vocabularies contain other cases of puzzling numeral elicitations (Zerries and Schuster 1974: 56). **4.** It is hard to know whether the Guayakí variety recorded from two youths by Vogt extended to a whole community of speakers (Vogt 1903: 861) and another attestation from roughly the same time appears to give forms for exact 1 and 2 (Mayntzhusen 1920: 20), though it may be that these forms are etymologizable (Vellard 1935). **5.** On the grounds that the present-day numerals can be etymologized to 'that', 'pair / couple', 'few' and 'another', Proto-Tupi

(Schleicher 1998: 12–13) may be argued to lack numerals. **6.** A vocabulary of Ofaye has 1 hœhá, 2 ñokoádi, 3 ñokoádn 4 ñokoádi (Hanke 1964: 29), i. e., 2 is the same as 4. A good guess, following more recent documentation (das Dores de Oliveira 2006: 109–110), is that the 4 in this earlier vocabulary is simply an error of some kind. **7.** Bernatzik (1942) claims that Yumbri lacked numerals above one. There is no further material on this variety but the closely related Minor Mlabri (Rischel 1995) has numerals up to three. Bernatzik's account has a sweeping and condescending flavour, and also has other doubtful claims of the same kind, e. g., lack of fiction which does not hold for Minor Mlabri either (cf. Velder 1963: 15). Another complicating factor is that he is able to discuss twin births at length with the people he says cannot comprehend any more distinctions than 'one' and 'many'. **8.** The oft-repeated claim (Parker 1909: 85) of lack of numerals in Vedda (Unclassified, Sri Lanka) appears, on closer scrutiny of the underlying sources, to be hearsay (Seligmann and Seligmann 1911: 33, 412). The only thing we can say is that no native term above two could be collected from the memories of the descendants, which does not necessarily mean that none existed. **9.** The first record of the language of Utanata (Asmat-Kamoro / Trans New Guinea, Indonesia) indicates counting inability on the part of the inhabitants (Earl 1837). However, lower numerals are attested in all subsequent descriptions – especially the most extensive piece (Drabbe 1953) – and have cognates in other Asmat-Kamoro languages (Galis 1955). Therefore, the counting inability reported probably reflects some kind of misunderstanding in the midst of the very difficult communication circumstances. **10.** Grondona (1998: 91) conjectures that pre-contact Mocoví (Guaicuruan, Argentina) lacked numerals above one as 2 and above are Spanish loans ("It seems that Mocoví lacked numeral forms, and has borrowed all its numerals from Spanish"). While it is true that Mocoví borrowed 2 and above from Spanish (cf. Gualdieri (1998: 211–212) and for the related Pilagá (Vidal 2001: 129)), it does not necessarily follow that Mocoví lacked 2 and above, before the borrowing. Older sources do, in fact, consistently attest a specific form for 2, see Koch-Grünberg (1903: 112–124) as well as Lafone Quevedo (1893: 244 and 1892: 410) **11.** Paiconeca (Bolivia-Parana / Arawakan, Bolivia) is a poorly attested extinct language of presumed Arawakan affiliation (Montaño Aragón 1989: 161–173). The naturalist d'Orbigny (1839: 191) travelled through the area in the 19th century and is the only source for numerals in the language. Since this is the only source, we can neither confirm or deny his report of lack of numerals:

> Il n'y a, dans cette langue, aucun système de numération, qu'y remplacent à peine quelques termes de comparaison, eux-mêmes, très-bornés.
>
> [In this language, there is no numeral system, in the place of which they are only just able to substitute some terms for comparison, which are themselves very narrow-minded.]

12. In all descriptive publications, Khoedam (Khoe / Khoe-Kwadi, Namibia) *|úí* and *|áḿ* are glossed as 'one' and 'two' respectively, but closer inspection reveals that these are really meanings accustomed to linguistic elicitation, and 'singleness' and 'dualness' are more appropriate glossings. There is a subtle difference between 'dualness' and 'twoness' in that dualness implies an association between the items in question. So "*|áḿ* children" would mean 'twins' rather than 'two children'. If this difference is deemed significant, then there was no word for 'two' in traditional Khoedam (Brenzinger 2009).

2. We regret that we have not been able to access two relevant-looking publications on the Mura language (Hanke 1950, 1952).

3. Swain has also submitted the same information for the Ninam entry for the *Numeral Systems of the World's Languages* website at http://lingweb.eva.mpg.de/numeral/Ninam.htm, accessed 1 July 2009.
4. Ross and Paul (1978: 60) give expressions for 1–8 in Waskia (Adelbert Range / Trans New Guinea, Papua New Guinea) with the structure 1, 2, $2+1$, $2+2$, $2+2+1$, $(2+1)+(2+1)$, $(2+1)+(2+1)+1$, $(2+1)+(2+1)+2$, that is, 6–8 are formed with additions based on $(2+1)+(2+1)$ for 6. This comes close, but does not count as base-3 according to the definition used in this paper.
5. Due to lack of data, we cannot confirm that the Arapeshan langages are related to Kombio or other groups usually subsumed under Torricelli.
6. Shown at http://lingweb.eva.mpg.de/numeral/Bukiyip.htm, accessed 1 July 2009.
7. We wish to thank Peter Bakker for highlighting this reference to us.
8. Payaguá, though poorly attested, is often counted as related to (at least) the Guaicuruan languages (Viegas Barros 2004) but we do not think the evidence is conclusive.
9. This language has not yet been identified with any modern variety (p. c. Jouni Filip Maho 2004; p. c. Roger Blench 2009).
10. With some speculative etymologizing, Chepang (Mahakiranti / Sino-Tibetan, Nepal) may have had 12 atoms and duodecimal counting up to 50, for a counting system associated with hunting (Caughley 1988, 1972; Hale 1973). One synopsis of Brúnkajk (Talamancan / Chibchan, Costa Rica) says that "también se cuenta por medio de docenas" (Arroyo Soto 1972: 32), but it is not clear on what this statement is based. It is not corroborated by a ten or so other descriptions of Brúnkajk, and it was not normed anyway, so it does not count as a base-12 system. In a modern description of Kinikinau (Bolivia-Parana / Arawakan, Brazil) higher numbers may be expressed using (dúzias) dozens (de Carvalho Couto 2005: 51), but this does not appear to be normed for exact enumeration of quantities that are not exact multiples of twelve.
11. However, vocabularies including monomorphemic 1–12 are listed for Hyam (there called 'Java') a few years earlier (de Castelnau 1851c: 59).
12. We wish to thank Roger Blench for help with sorting out various Plateau language identifications and classification questions.
13. We wish to thank James McElvenny for access to archival material on Biwat.
14. The extinct Saraveka has 'five hands' attested for 25 but no numerals 20–24 nor above 25 are recorded (de Créqui-Montfort and Rivet 1913). The 5-25-50 counting system in Kikongo (Bantu / Atlantic-Congo, DRC) referred to in Schmidl (1915: 181) was for counting pearls only (Laman 1968, 1912, 1936).
15. We wish to thank Thomas Hanke for bringing this case to our attention.
16. Sharma (2003: 63) claims that

    > We may say Munda speakers are the earliest known people who practised this system of counting which had monomorphemic units of counting up to twenty.

    but gives no source and no forms. Monomorphemic 1–20 forms cannot be found in the monograph on Munda numerals by Zide (1978) nor in any published description of Kharia or any other Munda language we have been able to consult. Nevertheless, a recent unpublished description of Kharia (Peterson 2006: 138–139), a set of monomorphemic 11–19 are recorded as alternative forms alongside a set of composite forms. Peterson notes, however, that the monomorphemic forms were given to him by youths who all confirmed that they had be taught them in school (and themselves used Sadani loans for

the numbers in question). Further inquiries by Peterson with experienced local teachers also point towards an "artificial" origin of the 11–19 forms (p. c. John Peterson 2008).

17. A modern example may be the recently innovated Palikúr (North Arawak / Arawak, Brazil-Guyana) numeral system, but it is not fully clear what the norms are Green (1994); Launey (2003).
18. We wish to thank Aymeric Daval Rasmussen for bringing the Mayan cases to our attention.

References

Adam, Lucien and Victor Henry

1880 *Arte y Vocabulario de la Lengua Chiquita con algunos textos traducidos y explicados compuestos sobre manuscritos inéditos del XVIII* (Bibliothèque Linguistique Américaine VI). Paris: Librairie-Éditeur J. Maisonneuve.

Adelaar, Willem F. H.

2004 *The Languages of the Andes* (Cambridge Language Surveys). Cambridge University Press.

Adwiraah, Eleonore

1989 *Grammatik des Gworok (Kagoro): Phonologie, Tonologie, Morphologie und Textanalyse* (Europäische Hochschulschriften: Reihe XXI: Linguistik 71). Frankfurt am Main: Peter Lang.

Aikhenvald, Alexandra and R. M. W. Dixon

1999 Other Small Families and Isolates. In R. M. W. Dixon and A. Aikhenvald (eds.), *The Amazonian Languages* (Cambridge Language Surveys), 341–383. Cambridge University Press.

Aikhenvald, Alexandra Y.

2008 *The Manambu language of East Sepik, Papua New Guinea*. Oxford University Press.

Anonby, Stan and Sandy Anonby

2007 A Report on Three Arauan Speech Varieties (Jamamadi, Jarawara, and Banawá) of the Amazon. SIL International, Dallas. SIL Electronic Survey Reports 2007-022 http://www.sil.org/silesr/abstract.asp?ref=2007-022.

Armstrong, Robert G.

1983 The Idomoid Languages of the Benue and Cross River Valleys. *Journal of West African Languages* XIII (1): 91–147.

Arroyo Soto, Víctor Manuel

1972 *Lenguas indígenas costarricenses*. 2nd edn. San José, Costa Rica.

Asangama, Natisa

1983 *Le Budu: langue bantu du nord-est du Zaïre: Esquisse Phonologique et grammaticale*. Paris: Université de la Sorbonne doctoral dissertation.

Avelino, Heriberto

2006 The typology of Pame number systems and the limits of Mesoamerica as a linguistic area. *Linguistic Typology* 10 (1): 493–513.

Balbi, Adrien
1826 *Discours préliminaire et introduction* (Atlas Ethnographique du Globe I). Paris: Rey et Gravier. [Also titled "Introduction à l'Atlas Ethnographique du Globe"].

Barriga Puente, Francisco
1998 *Los Sistemas de Numeración Indoamericanos: un enfoque areotipológico* (Colección Lingüística Indígena 7). México: Universidad Nacional Autónoma de México.

Becher, Hans
1960 *Die Surára und Pakidái: Zwei Yanonámi-Stämme in Nordwestbrasilien, mit anhang über die Sprache der Surára und Pakidái von Aryon D. Rodrigues* (Mitteilungen aus dem Museum für Völkerkunde in Hamburg XXVI). Hamburg: Kommissionsverlag Cram, De Gruyter & Co.

Beeler, Madison S.
1961 Senary Counting in California Penutian. *Anthropological Linguistics* 3 (6): 1–8.

Beeler, Madison S.
1963 Ventureño Numerals. In W. Bright (ed.), *Studies in Californian Linguistics* (University of California Publications in Linguistics 34), 13–18. Berkeley / Los Angeles: University of California Press.

Beeler, Madison S.
1967 *The Ventureño Confesario of José Señán, O.F.M.* (University of California Publications in Linguistics 47). Berkeley: University of California Press.

Bernatzik, Hugo Adolf
1942 *De Gula Bladens Andar: Forskningsresor i Bortre Indien*. Stockholm: Bokförlaget Natur och Kultur.

Betts, LaVera
1981 *Dicionário parintintín-português português-parintintín*. Brasília: Summer Institute of Linguistics.

Blench, Roger
2005 Baŋgi Me, a language of unknown affiliation in Northern Mali and its affinities. Draft Manuscript March 18, 2005.

Blench, Roger
2006a The Ake language of Central Nigeria and its affinities. Draft Manuscript January 2, 2006.

Blench, Roger
2006b The Cara language of Central Nigeria and its affinities. Draft Manuscript January 3, 2006.

Blench, Roger
2006c The Ganang Language of Central Nigeria and its Affinities. Draft Manuscript January 3, 2006.

Blench, Roger
2006d Iten-English Dictionary. Draft Manuscript January 2, 2006.

Blench, Roger
2006e The Ninkyop Language of Central Nigeria and its Affinities. Draft Manuscript January 7, 2006.

Blench, Roger
2006f Prospecting Proto-Plateau. Draft Manuscript January 3, 2006.

Blench, Roger
2006g The Tahoss Dialect of the Berom Language of Central Nigeria and its Affinities. Draft Manuscript January 3, 2006.

Blench, Roger
2006h The Təsu language of Central Nigeria and its affinities. Draft Manuscript September 16, 2006.

Blench, Roger
2007 The Dyarim language of Central Nigeria and its affinities. In Henry Tourneux (ed.), *Topics in Chadic Linguistics III: Historical Studies: Papers from the 3rd Biennial International Colloquium on Chadic Languages, Villejuif, November 24–25, 2005* (Tschadistik / Linguistique Tchadique 4), 41–59. Köln: Rüdiger Köppe.

Blench, Roger
2009a The Idũ language of Central Nigeria: Phonology, wordlist and suggestions for orthography changes. Draft Manuscript May 20, 2009.

Blench, Roger
2009b The Nyankpa [= Yeskwa] language of Central Nigeria. Draft Manuscript May 20, 2009.

Blench, Roger
2009c The Tinɔr [= Koro Waci] language of Central Nigeria and its affinities. Draft Manuscript May 20, 2009.

Blench, Roger, Ruth Adiwu and Gideon Asukutu
2006 The Ce [Rukuba] language of Central Nigeria and its affinities. Draft Manuscript January 4, 2006.

Blench, Roger and I. D. Hepburn
2006 A Dictionary of Eggon. Draft Manuscript January 2, 2006.

Blench, Roger and Barau Kato
2006 A Dictionary of Mada, a Plateau Language of Central Nigeria based on the Rija Dialect: Mada-English with an English-Mada finderlist. Draft Manuscript January 4, 2006.

Blench, Roger and Bitrus Bulus Kaze
2006 A Dictionary of the Izere Language of Fobur. Draft Manuscript January 6, 2006.

Bokula, F.-X.
1970 La Langue Bodo: Formes Nominales. *Africana Linguistica* IV: 63–84.

Bokula, M. and L. Ngandi
1985 Numération Cardinale dans les Langues Bantu du Haut-Zaire. *Annales Æquatoria* 6: 189–196. [Also in *Annales de l'I.S.P.-Kisangani* 13 (July): 700–707, 1984.]

Borgman, Donald M.
1990 Sanuma. In Desmond C. Derbyshire and Geoffrey K. Pullum (eds.), *Handbook of Amazonian Languages* volume II, 15–248. Mouton de Gruyter.

Bouquiaux, Luc
1962 A propos de numération: L'emploi du système décimal et du système duodécimal dans la langue Birom (Nigéria septentrional). *Africana Linguistica* I: 7–10.

Bouquiaux, Luc
1964 A Word List of Aten (Ganawuri). *Journal of West African Languages* I (2): 5–26.

Bouquiaux, Luc
1970 *La langue Birom (Nigeria Septentrional): phonologie, morphologie, syntaxe* (Bibliothèque de la Faculté de Philosophie et Lettres de l'Université de Liège – Fascicule CLXXXV). Paris: Société d'Édition "les belles lettres".

Bowers, Nancy and Pundia Lepi
1975 Kaugel Valley Systems of Reckoning. *Journal of the Polynesian Society* 84: 309–324.

Bradshaw, Robert
2007 *Fuyug grammar sketch* (Data Papers on Papua New Guinea Languages 53). Ukarumpa, Papua New Guinea: SIL-PNG Academic Publications.

Brenzinger, Matthias
2009 Documenting concepts in contact. Paper presented at the CIPL Conference on the Sociolinguistics of Language Endangerment, 27 June 2009, School of Oriental and African Studies, London.

Bruno, Ana Carla
2003 *Waimiri Atroari Grammar: Some Phonological, Morphological, and Syntactic Aspects*. University of Arizona doctoral dissertation.

Calame-Griaule, Geneviève
1968 *Dictionnaire Dogon Dialecte Tørø: Langue et Civilisation* (Langues et Littératures de l'Afrique Noire IV). Paris: Librairie C. Klincksieck.

Carlson, Robert
1994 *A Grammar of Supyire* (Mouton Grammar Library 14). Mouton de Gruyter.

Caughley, Ross C.
1972 *A vocabulary of the Chepang language*. Kirtipur: Summer Institute of Linguistics and Institute of Nepal Studies, Tribhuvan University.

Caughley, Ross C.
1988 Chepang: A Sino-Tibetan language with a duodecimal numeral base? In David Bradley, Eugénie J. A. Henderson and Martine Mazaudon (eds.), *Prosodic analysis and Asian linguistics: To honour R. K. Sprigg* (Pacific Linguistics: Series C 104), 197–199. Canberra: Australian National University.

Cheetham, B.
1978 Counting and Number in Huli. *Papua New Guinea Journal of Education* 14: 16–27.

Clark, Charles U.
1937 Jesuit Letters to Hervás on American Languages and Customs. *Journal de la Société des Américanistes* XXIX: 97–145.

Closs, Michael P.
1986 Native American Number Systems. In Michael P. Closs (ed.), *Native American Mathematics*, 3–44. Austin: University of Texas Press.

Conant, Leonard Levi
1896 *The Number Concept: its origin and development*. New York: MacMillan.

Conrad, Robert J., Joshua Lukas and John Alungum
1978 Some Muhiang grammatical notes. In Richard Loving (ed.), *Miscellaneous papers on Dobu and Arapesh* (Workpapers in Papua New Guinea Languages 25), 89–130. Ukarumpa: Summer Institute of Linguistics.

Conrad, Robert J. and Kepas Wogiga
1991 *An Outline of Bukiyip Grammar* (Pacific Linguistics; C-113). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Conzemius, Eduard
1927, 1928 Los Indios Payas de Honduras: Estudio geográfico, histórico, etnográfico y lingüístico. *Journal de la Société des Américanistes* XIX, XX: 245–302, 253–360.

Czekanowski, Jan
1924 Sprachenaufnahmen. In *Forschungen im Nil-Kongo-Zwischengebiet: Zweiter Band: Ethnographie Uele / Ituri / Nil-Länder* (Wissenschaftliche Ergebnisse der deutschen Zentral-Afrika-Expedition 1907–1908: Ethnographie-Anthropologie VI:2), 575–714. Leipzig: Klinkhardt & Biermann.

das Dores de Oliveira, Maria
2006 *Ofayé, a língua do povo do mel: Fonologia e Gramática*. Maceió: Universidade Federal de Alagoas doctoral dissertation.

Daval-Markussen, Aymeric, Peter Bakker and Harald Hammarström
in press On the Origins of the Malagasy Numeral System. Submitted.

de Carvalho Couto, Valéria Guimarães
2005 *A Língua Kinikinau: Estudo do Vocabulário e conceitos Gramaticais*. Três Lagoas: Universidade Federal de Mato Grosso do Sul, Câmpus de Três Lagoas doctoral dissertation.

de Castelnau, Francis
1851a *Histoire du Voyage* (Expédition dans les parties central de l'Amérique du Sud III). Paris: P. Bertrand.

de Castelnau, Francis
1851b *Histoire du Voyage* (Expédition dans les parties central de l'Amérique du Sud V). Paris: P. Bertrand.

de Castelnau, Francis
1851c *Renseignements sur l'Afrique Centrale et sur une Nation d'Hommes à queue qui s'y trouverait*. Paris: P. Bertrand.

de Créqui-Montfort, G. and Paul Rivet
1913 Linguistique Bolivienne: La Langue Saraveka. *Journal de la Société des Américanistes* X: 497–540.

de Matallana, Bernardo and Cesareo de Armellada
1943 Exploración del Paragua. *Boletín de la Sociedad Venezolana de ciencias naturales* VIII (53): 61–110.

de Vries, Lourens J.
1998 Body part tally counting and Bible translation in Papua New Guinea and Irian Jaya. *The Bible Translator (Practical Papers)* 49 (4): 409–415.

Dehaene, Stanislas and Jacques Mehler
1992 Cross-linguistic Regularities in the Frequency of Number Words. *Cognition* 43: 1–29.

Delafosse, Maurice
1928 La numération chez les Nègres. *Africa, Journal of the International Institute of African Languages* 1 (3): 387–390.

Dempwolff, Otto
n.d. *A Grammar of the Graged Language*. Narer, Karkar Island: Lutheran Mission.

Dijkmans, Joseph J. M.
1974 *Kare-Taal: Lijst van woorden gangbaar bij het restvolk Kare opgenomen in de jaren 1927–1947*. Sankt Augustin: Anthropos-Institut – Haus Völker und Culturen.

Dixon, Roland B. and Alfred L. Kroeber
1907 Numeral Systems of the Languages of California. *American Anthropologist* 9 (4): 663–689.

Dixon, R. M. W.
2004 *The Jarawara Language of Southern Amazonia*. Oxford University Press.

Dombrowski, Franz A. and Bruno W. W. Dombrowski
1991 Numerals and numeral systems in the Hamito-Semitic and other language groups. In Alan S. Kaye (ed.), *Semitic Studies in Honor of Wolf Leslau* volume 1, 340–381. Wiesbaden: Otto Harrassowitz.

Donohue, Mark
2008 Complexities with restricted numeral systems. *Linguistic Typology* 12 (3): 423–429.

d'Orbigny, Alcide Dessalines
1839 *L'homme américain (de l'Amérique Méridionale): considéré sous ses rapports physiologiques et moraux* volume 2. Paris: Pitois-Levrault et C.

Drabbe, Peter
1952 *Spraakkunst van het Ekagi: Wisselmeren Ned. N. Guinea*. 'S-Gravenhage: Martinus Nijhoff.

Drabbe, Peter
1953 *Spraakkunst van de Kamoro-Taal*. 'S-Gravenhage: Martinus Nijhoff.

Dwight, T.
1848 Sketch of the Polynesian Language. *Transactions of the American Ethnological Society* II: 223–234.

Earl, George W.
1837 Review of Verhaal nan eene Reize naar en langs de zuid-west kust van Nieuw Guinea, gedaan in 1828, door Z. M. Corvet Triton, en Z. M. Coloniale schoener de Iris, door J. Modera, Lieut. ter Zee, van de tweede Klasse by Z. M. Corvet Triton; J. Modera. *Journal of the Royal Geographical Society of London* 7: 383–395.

Early, John D. and John F. Peters
2000 *The Xilixana Yanomami of the Amazon: history, social structure, and population dynamics*. University Press of Florida.

Ehrenreich, Paul
1887 Ueber die Botocudos der brasilianischen Provinzen Espiritu Santo und Minas Geraes. *Zeitschrift für Ethnologie* XIX: 1–46, 49–82.

Epps, Patience
2006 Growing a numeral system: The historical development of numerals in an Amazonian language family. *Diachronica* 23 (2): 259–288.

Evans, Nicholas
2009 Two *pus* one makes thirteen: Senary numerals in the Morehead-Maro region. *Linguistic Typology* 13 (2): 321–335.

Everett, Daniel L.
2004 The Absence of Numerals in Pirahã. Presentation at the Workshop on Numerals, Max Planck Institute for Evolutionary Anthropology, Leipzig, 29–30 of March, 2004.

Everett, Daniel L.
2005 Cultural Constraints on Grammar and Cognition in Pirahã: Another Look at the Design Features of Human Language. *Current Anthropology* 46 (4): 89–130.

Everett, Daniel L. and Barbara Kern
1997 *Wari': the Pacaas Novos language of Western Brazil* (Descriptive Grammars Series). London / New York: Routledge.

Fast, Daniel, Ruby Fast and Gerhard Fast
1996 *Diccionario Achuar-Shiwiar – Castellano* (Serie Lingüística Peruana; 36). Yarinacocha: Ministerio de Educación and Instituto Lingüístico de Verano.

Fast, Gerhard and Ruby Fast
1981 *Introducción al idioma achuar* (Documento de Trabajo 20). Lima: Ministerio de Educación and Instituto Lingüístico de Verano.

Fastre, P.
1920 Vocabulary: Name of Tribe, Fujuge (Mafulu), Name of Village, Sivu. *Commonwealth of Australia. Papua: Annual Report for the Year* 1918–19: 116–116.

Fleischmann, Lillian and Sinikka Turpeinen
1975 Bine Grammar Essentials. Ukarumpa, Papua New Guinea: Unpublished Typescript, The Summer Institute of Linguistics.

Fortune, Reo F.
1942 *Arapesh* (Publications of the American Ethnological Society XIX). New York: J. J. Augustin Publisher.

Frank, Michael C., Daniel L. Everett, Evelina Fedorenko and Edward Gibson
2008 Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition* 108: 819–824.

Franklin, Karl and Joyce Franklin
1962 The Kewa Counting Systems. *Journal of the Polynesian Society* 71 (2): 188–192.

Friederici, Georg
1912 *Wissenschaftliche Ergebnisse: einer amtlichen Forschungsreise nach dem Bismarck-Archipel im Jahre 1908: II Beiträge zur Völker- und Sprachenkunde von Deutsch-Neuguinea* (Ergänzungsheft der Mitteilungen aus den Deutschen Schutzgebieten 5). Berlin: Ernst Siegfried Mittler und Sohn.

Fritz, Sonja
2002 *The Dhivehi Language* (Beiträge zur Südasienforschung 191). Würzburg: Ergon.

Fudeman, Kirsten Anne
1999 *Topics in the Morphology and Syntax of Balanta, an atlantic language of Senegal*. Cornell University doctoral dissertation.

Galis, Klaas Wilhelm
1955 Talen en dialecten van Nederlands Nieuw-Guinea. *Tijdschrift Nieuw-Guinea* 16: 109–118, 134–145, 161–178.

Gamble, Geoffrey L.
1980 The "old time" Chunut Count. In Kathryn Klar, Margaret Langdon and Shirley Silver (eds.), *American Indian and Indoeuropean Studies: Papers in Honor of Madison S. Beeler*, 51–55. The Hague: Mouton de Gruyter.

Gerhardt, Ludwig
1968 Analytische und Vergleichende Untersuchungen zu einigen zentralnigerianischen Klassensprachen: Teil 1. *Afrika und Übersee* LI: 161–198.

Gerhardt, Ludwig
1969 Analytische und Vergleichende Untersuchungen zu einigen zentralnigerianischen Klassensprachen (fortsetzung). *Afrika und Übersee* LII: 23–57, 125–143, 207–242.

Gerhardt, Ludwig
1973 Abriß der nominalen Klassen im Koro, North Central State, Nigeria. *Afrika und Übersee* LVI (4): 245–266.

Gerhardt, Ludwig
1983 The Classification of Eggon: Plateau or Benue Group?. *Journal of West African Languages* XIII (1): 37–50.

Gerhardt, Ludwig
1987 Some Remarks on the Numerical Systems of the Plateau Languages. *Afrika und Übersee* 70: 19–29.

Gerhardt, L[udwig]
1997 Die Jarawan-Bantu-Sprachen und ihr linguistisches und kulturelles Umfeld. *Frankfurter Afrikanistische Blätter* 9: 129–146.

Gerhardt, Ludwig
2005 Some notes on Yeskwa (North-Western Plateau, Nigeria) with comments on Koelle's Polyglotta Africana. *Hamburger Afrikanistische Arbeitspapiere* 3: 35–52.

Gilliard, L.
1924 La numérotation des Ntomba, riverains du Lac Léopold II. *Congo* 5 (II (3)): 374–378.

Gilliard, L.
1928 *Grammaire synthétique de Lontomba: suivie d'un vocabulaire* (Bibliothèque-Congo 20). Bruxelles: Éditions de l'Essorial.

Gnerre, Mauricio Covaz
1986 Some notes on quantification and numerals in an Amazon Indian language. In Michael P. Closs (ed.), *Native American Mathematics*, 71–92. Austin: University of Texas Press.

Gómez, Gale Goodwin
1990 *The Shiriana Dialect of Yanam (Northern Brazil)*. Columbia University doctoral dissertation.

Gordon, Peter
2004 Numeral Cognition Without Words: Evidence from Amazonia. *Science* 306: 496–499.

Gray, Albert
1878 The Maldive Islands with a vocabulary taken from Farancois Pyrard de Laval, 1602–1607. *Journal of the Royal Asiatic Society of Great Britain and Ireland, N.S.* 10: 173–209.

Green, Diana
1994 O Sistema Numérico da Língua Palikúr. *Boletim do Museu Paraense Emílio Goeldi, Série Antropologia* 10 (2): 261–303.

Green, Diana
1997 Diferenças entre termos numéricos em algumas línguas indígenas do Brasil. *Boletim do Museu Paraense Emílio Goeldi, Série Antropologia* 13 (2): 179–207.

Grondona, Verónica M.
1998 *A Grammar of Mocovi*. University of Pittsburgh doctoral dissertation.

Gualdieri, Beatriz C.
1998 *Mocoví (Guaycurú): Fonologia e Morfossintaxe*. Universidade Estadual de Campinas doctoral dissertation.

Haberland, Eike and Siegfried Seyfarth
1974 *Die Yimar am Oberen Korowori (Neuguinea)* (Studien zur Kulturkunde 36). Wiesbaden: Franz Steiner.

Hafford, James A.
1999 *Elements of Wuvulu Grammar.* Masters thesis, University of Texas at Arlington.

Hale, Austin
1973 *Clause, sentence, and discourse patterns in selected languages of Nepal 4: Word lists* volume 40:4. Summer Institute of Linguistics: Publications in Linguistics.

Hammarström, Harald
2009 Whence the Kanum Base-6 Numeral System?. *Linguistic Typology* 13 (2): 305–319.

Hanke, Wanda
1950 Vocabulário e idioma Mura dos índios Mura do rio Manicoré. *Arquivos* (Manaus) 12: 3–8.

Hanke, Wanda
1952 *O idioma Mura* (Documentação do Amazonas 1). Manaus.
Hanke, Wanda
1956 Beobachtungen über den Stamm der Huari (Rio Corumbiara) Brasilien. *Archiv für Völkerkunde* 11: 67–82.
Hanke, Wanda
1964 Verstreute Indianerdörfer im Südosten des Mato Grosso. In *Völkerkundliche Forschungen in Südamerika* (Kulturgeschichtliche Forschungen 11), 9–33. Braunschweig: Albert Limbach.
Harries, Lyndon
1959 Nyali, a Bantoid Language (Belgian Congo). *Kongo-Overzee* XXV: 174–205.

Harris, John
1982 Facts and Fallacies of Aboriginal Number Systems. In Susanne Hargrave (ed.), *Language and Culture*, 153-182. Darwin: Summer Institute of Linguistics.
Hartner, W.
1943 Zahlen und Zahlensysteme bei Primitiv- und Hochkulturvölkern. *Paideuma* 6 / 7: 286–326.
Hervás y Panduro, Lorenzo
1786 *Aritmetica Delle Nazioni e Divisione del fra L'Orientali* (Idea dell'Universo XIX). Cesena: Gregorio Biasini.
Hinton, Leanne
1994 10: California Counting. In *Flutes of Fire: Essays on California Indian Languages*, 113–121. Berkeley, California: Heyday Books.
Hopkins, Nicholas A.
1967 A Short Sketch of Chalchihuitán Tzotzil. *Anthropological Linguistics* 9 (4): 9–25.
Howitt, Alfred William
1889 Notes on Australian Message Sticks and Messengers. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* XVIII: 314–332.
Howitt, Alfred William
1904 *The Native Tribes of South-East Australia*. New York: MacMillan.
Hughes, Barnabas B.
1974 The earliest known record of California Indian numbers. *Historia Mathematica* 1: 79–82.
Hughes, Barnabas
1982 Hawaiian Number Systems. *The Mathematics Teacher* 75 (3): 253–256.
Ibarra Grasso, Dick Edgar
1938 Las Numeraciones Indígenas Americanas. *Boletín de la Academia Argentina de Letras* VI (23–24): 397–417.
Ibarra Grasso, Dick Edgar
1939a Las Numeraciones Cuaternarias. *Boletín de la Academia Argentina de Letras* VII (28): 585–606.
Ibarra Grasso, Dick Edgar
1939b Las Numeraciones Senario-decimales en Sudamérica. *Boletín de la Academia Argentina de Letras* VII (25–26): 187–213.

Ibriszimow, Dymitr
1988 Some Remarks on Chadic Numerals. In W. J. G. Möhlig (ed.), *Afrikanistische Beiträge zum XXIV. Deutschen Orientalistentag 26.–30. September 1988* (Afrikanistische Arbeitspapiere: Sondernummer), 67–74. Köln: Universität zu Köln.

Johnston, Harry
1904 *The Uganda Protectorate* volume 2. 2nd edn. London: Hutchinson.

Johnston, Harry H.
1922a Chapter XIII: The Bantu and Semi-Bantu Numerals. In *A Comparative Study of the Bantu and Semi-Bantu Languages* volume II, 463–482. Oxford University Press.

Johnston, Harry H.
1922b *A Comparative Study of the Bantu and Semi-Bantu Languages*. Oxford University Press. [1919]

Judd, A. S.
1923 Notes on the Language of the Arago or Alago Tribe of Nigeria. *Journal of the African Society* XXIII (LXXXIX): 30–38.

Jungraithmayr, Herrmann
1963 Die Sprache der Sura (Maghavul) in Nordnigerien. *Afrika und Übersee* XLVII: 8–89, 204–220.

Jungraithmayr, Herrmann
1970 *Die Ron-Sprachen: Taschadohamitische Studien in Nordnigerien* (Afrikanistische Forschungen III). Glückstadt: J. J. Augustin.

Kalunga Mwela-Ubi, Marcel
1999 *Le Numéral en Bantu: Considérations Typologiques*. Université de Lubumbashi doctoral dissertation.

Kissenberth, Wilhelm
1912 Bei den Canella-Indianern in Zentral-Maranhão (Brasilien). *Baessler Archiv* II: 45–54.

Klingenheben, August
1926–1927 Zu den Zählmethoden in den Berbersprachen. *Zeitschrift für Eingeborenensprachen* XVII: 40–51.

Knobloch, Franz
1967 *Die Aharaibu-Indianer in Nordwest-Brasilien* (Collectanea Instituti Anthropos 1). St. Augustin bei Bonn: Anthropos Institut.

Koch-Grünberg, Theodor
1903 Die Guaikurú-Gruppe. *Mitteilungen der Anthropologischen Gesellschaft in Wien* XXXIII: 1–128.

Koch-Grünberg, Theodor
1906 Makú. *Anthropos* I: 877–906.

Koch-Grünberg, Theodor
1928 *Sprachen* (Von Roroima zum Orinoco: Ergebnisse einer Reise in Nordbrasilien und Venezuela in den Jahren 1911–13 4). Stuttgart: Strecker und Schröder.

Koelle, Sigismund W.
1854 *Polyglotta Africana or Comparative Vocabulary of Nearly Three Hundred Words and Phrases in more than One Hundred Distinct African Languages*. London: Church Missionary House.

Kolia, J. A.
1975 A Balawaia grammar sketch and vocabulary. In T. E. Dutton (ed.), *Studies in languages of central and south-east Papua* (Pacific Linguistics: Series C 29), 107–226. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Kroeber, Alfred L.
1925 Yuki: Culture. In *Handbook of the Indians of California* (Bulletin of the Bureau of American Ethnology 78), 169–181. St. Clair Shores, Michigan: Scholarly Press.

Kutsch Lojenga, Constance
1994a *Ngiti: A Central-Sudanic Language of Zaire* (Nilo-Saharan: Linguistic Analyses and Documentation 9). Köln: Rüdiger Köppe.

Kutsch Lojenga, Constance
1994b Kibudu: A Bantu Language with nine Vowels. *Africana Linguistica* XI: 127–133.

Lafone Quevedo, Samuel A.
1890–1893 Notas ó sea principios de gramática mocoví segun ellos se desprenden de los trabajos de Tavolini, Dobrizhoffer, Barcena y otros. *Revista del Museo de La Plata* 1: 111–144, 1: 305–328, 2: 241–272, 2: 289–352, 2: 393–424, 3: 129–167, 4: 257–287.

Lafone Quevedo, Samuel A.
1892 Introducción al "Arte Mocoví" del Padre Tavolini: Estudio de Gramática Comparada. *Revista Museo de la Plata* 4: 369–432.

Laman, Karl Edvard
1912 *Grammar of the Kongo Language (Kikongo)*. New York: The Christian Alliance Publishing Co.

Laman, Karl Edvard
1936 *Dictionnaire kikongo-français avec une étude phonétique décivant les dialectes les plus importants de la langue dite kikongo* (Mémoires de l'Institut Royal colonial Belge: Section des sciences morales et politiques: Collection in-8o). Bruxelles: Institut Royal colonial Belge.

Laman, Karl Edvard
1953–1968 *The Kongo* (Studia Ethnographica Upsaliensia). Uppsala. 4 vols.

Lancy, David F. and Andrew J. Strathern
1981 "Making Twos": Pairing as an Alternative to the Taxonomic Mode of Representation. *American Anthropologist, N.S.* 83 (4): 773–795.

Launey, Michel
2003 *Awna Parikwaki: Introduction à la langue Palikur de Guyane et de l'Amapá* (Didactiques). Paris: IRD.

Laycock, Donald C.
1965 *The Ndu language family (Sepik District, New Guinea)* (Linguistic Circle of Canberra Publications: Series C, Books 1). Canberra: Australian National University.

Laycock, Donald C.
1970 Eliciting Basic Vocabulary in New Guinea. In Stephen A. Wurm and Donald C. Laycock (eds.), *Pacific linguistic studies in honour of Arthur Capell* (Pacific Linguistics: Series C 13), 1127–1176. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Lean, Glendon A.
1986a *Enga, Western Highlands, Simbu* (Counting Systems of Papua New Guinea 9). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Lean, Glendon A.
1986b *Sandaun Bay Province* (Counting Systems of Papua New Guinea 7). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Lean, Glendon A.
1986c *Southern Highlands* (Counting Systems of Papua New Guinea 10). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Lean, Glendon A.
1986d *Western Province* (Counting Systems of Papua New Guinea 12). Port Moresby: Papua New Guinea University of Technology. Draft Edition.

Lean, Glendon A.
1992 *Counting Systems of Papua New Guinea and Oceania*. Doctoral dissertation, Papua New Guinea University of Technology.

Lee, Jennifer
1987 *Tiwi Today: A Study of Language Change in a Contact Situation* (Pacific Linguistics: Series C 96). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Lehmann, Walther
1920 *Zentral-Amerika: Die Sprachen Zentral-Amerikas* volume I. Berlin: Dietrich Reimer.

Lomas, Gabe
1988 *The Huli language of Papua New Guinea*. Macquarie University doctoral dissertation.

Loukotka, Čestmír
1955 Les Indiens Botocudo et leur Langue. *Lingua Posnaniensis* V: 112–135.

Loukotka, Čestmír
1963 Documents et vocabulaires inédits de langues et de dialectes sud-américains. *Journal de la Société des Américanistes* LII: 7–60.

Luzio, Aldo Di
1972 / 1973 Preliminary Description of the Amo Language. *Afrika und Übersee* LVI: 3–61.

Machoni de Cerdeña, Antonio
1732 *Arte y vocabulario de la lengua Lule o Tonocoté*. Buenos Aires: Coni. [Reprint 1877].

Mackay, Hugh D.
1964 A Word List of Eloyi. *Journal of West African Languages* I (1,2): 5–12, 60.
Maddieson, Ian and Kay Williamson
1975 Jarawan Bantu. *African Languages / Langues Africaines* I: 124–163.
Mamet, Ingo
2005 *Die Ventureño-Chumash-Sprache (Südkalifornien) in den Aufzeichnungen John Peabody Harringtons* (Europäische Hochschulschriften: Reihe 19, Volkskunde / Ethnologie: Abteilung B: Ethnologie 67). Frankfurt am Main: Peter Lang.
Man, E. H.
1883a On the Aboriginal Inhabitants of the Andaman Islands (Part I.). *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 12: 69–116.
Man, E. H.
1883b On the Aboriginal Inhabitants of the Andaman Islands (Part III.). *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 12: 327–434.

Mathews, H. F.
1917 Notes on the Nungu Tribe, Nassawara Province, Northern Nigeria, and the Neighboring Tribes which use the Duodecimal System of Numeration. *Harvard African Studies* I: 83–93.
Mathews, R. H.
1904 Language of the Wuddyāwūrru Tribe, Victoria. *Zeitschrift für Ethnologie* 36: 729–734.
Matsushita, Shuji
1998 Decimal vs. Duodecimal: An interaction between two Systems of Numeration. Paper presented at the 2nd Meeting of the AFLANG, October 1998, Tokyo.
Mattei-Müller, Marie-Claude
2007 *Lengua y Cultura Yanomamɨ: Diccionario ilustrado Yanomamɨ-Español / Español-Yanomamɨ*. Caracas: Consejo Nacional de la Cultura.
Mayntzhusen, F. C.
1919-1920 Die Sprache der Guayaki. *Zeitschrift für Eingeborenensprachen* X: 2–22.
Mazaudon, Martine
2007 Number building in Tibeto-Burman languages. Paper presented at the North-East India Languages Symposium 2 (NEILS-2), Gauhati (Assam, India), 5-9 February 2007.
McElvenny, James
2006 Draft sketch grammar of Biwat (Mudukumo / Mundugumor), a Yuat language of East Sepik Province, PNG. Available on request from the author http://www.pfed.info/james accessed 13 May 2009.
McGregor, William B.
2004 *The Languages of the Kimberley, Western Australia*. London and New York: Routledge.
Mead, Margaret
1932 Mundugumor Linguistics. Manuscript typed up for Donald C. Layock 1973.
Meek, Charles K.
1925 *The Northern Tribes of Nigeria*. Oxford University Press.

Meek, Charles K.
1931 *Tribal Studies in Northern Nigeria* volume 2. London: Kegan Paul, Trench, Trübner.

Menanti, Jackie
forthc. Laporan Sosiolinguistik Bahasa Viid di Kampung Senggi, Kabupaten Keerom, Papua – Indonesia. To appear in the SIL Electronic Survey Reports.

Migliazza, Ernest
1972 *Yanomama grammar and intelligibility*. Indiana University doctoral dissertation.

Migliazza, E. C.
1978 Maku, Sape and Uruak Languages: Current Status and Basic Lexicon. *Anthropological Linguistics* XX (3): 133–140.

Mock, Carol
1977 *Chocho: Santa Catarina Ocotlán, Oaxaca* (Archivo de Lenguas Indígenas de México 4). Mexico: Centro de Investigación para la Integración Social.

Montaño Aragón, Mario
1989 *Tribus de la Selva II* (Guía etnográfica lingüística de Bolivia). La Paz, Bolivia: Don Bosco.

Monteil, Charles
1905 Considérations générales sur le nombre et la numération chez les Mande. *L'Anthropologie* 16: 485–502.

Moolenburgh, P. E.
1904 Extract uit een verslag der Noord Nieuw-Guinea expeditie. *Tijdschrift voor Indische Taal-, Land- en Volkenkunde (TBG)* 47: 168–188, 381–385.

Münzel, Mark
1969–1972 Notas preliminares sobre os Kabori (Makú entre o río negro e o japura). *Revista de Antropologiía* 17–20: 137–181.

Nekitel, Otto
1985 *Sociolinguistic Aspects of Abu', a Papuan Language of the Sepik Area, Papua New Guinea*. Canberra: Australian National University doctoral dissertation.

Nengel, John G.
1999 *Precolonial African Intergroup Relations in the Kauru and Pengana Polities of Central Nigerian Highlands 1800-1900* (European University Studies: Series III; 814). Frankfurt: Peter Lang.

Nengel, John G.
n.d. Endangerment and Survival Prospects of the ɛBoze Language (Northern Jos Plateau of Central Nigeria). Manuscript, University of Jos. [ca. 2004]

Nimuendajú, Curt
1924 Os Indios Parintintin do Alto Madeira. *Journal de la Société des Américanistes* XVI: 201–278.

Nimuendajú, Curt
1932 Wortlisten aus Amazonien. *Journal de la Société des Américanistes* XXIV: 93–119.

Nimuendajú, Curt and E. H. do Valle Bentes
1923 Documents sur quelques langues peu connues de l'Amazone. *Journal de la Société des Américanistes* XV: 215–222.

Nordenskiöld, Erland
1911 *Indianer och Hvita i Nordöstra Bolivia*. Stockholm: Bonniers.

Nordenskiöld, Erland
n.d. Anteckningar om sydamerikanska infödingsspråk: Bolivia- och Chacostammar. Gothenburg: Manuscript B5858, Världskulturmuséet.

Parker, H.
1909 *Ancient Ceylon: An account of the aborigines and of part of the early civilisation*. New Delhi: Marwah Publications.

Parkinson, Richard
1907 *Dreißig Jahre in der Südsee*. Stuttgart: Strecker & Schröder.

Peters, John F.
1998 *Life among the Yanomami: the story of change among the Xilixana on the Mucajai River in Brazil*. Peterborough, Ontario: Broadview.

Peterson, John
2006 *Kharia: A South Munda Language*. Habilitationsschrift, Universität Osnabrück.

Pica, Pierre, Cathy Lemer, Véronique Izard and Stanislas Dehaene
2004 Exact and Approximate Arithmetic in an Amazonian Indigene Group. *Science* 306: 499–503.

Pires, Nádia N.
1992 Estudo da gramática da língua Jeoromitxi (Jabuti). Universidade Estadual de Campinas masters thesis.

Plank, Frans
2009 Senary summary so far. *Linguistic Typology* 13 (2): 337–345.

Popky, Donna
1999 Oro Win: A Descriptive and Comparative Outlook of an Endangered Language. University of Pittsburgh masters thesis.

Powell, Marvin A. Jr.
1972 The Origin of the Sexagesimal System: The Interaction of Language and Writing. *Visible Language* 6 (1): 5–18.

Pumuge, Hilary Manda
1975 The Counting System of the Pekai-Alue Tribe of the Topopul Village in the Ialibu Sub-district in the Southern Highlands District, Papua New Guinea. *Science in New Guinea* 3 (1): 19–25.

Quintina, Fernando R.
1961 Conhencimento da língua Balanta. *Boletim Cultural da Guiné Portuguesa* XVI (64): 737–768.

Ramirez, Henri
1994a *Iniciação à língua Yanomama: Dialetos do Médio Rio Catrimani e de Xitei: Curso de língua Yanomama*. Brasil: Diocese do Roraima, Boa Vista.

Ramirez, Henri
1994b *Le Parler Yanomami des Xamatauteri*. Aix-en-Provence: Université de Provence doctoral dissertation.

Ray, Sidney H.
1912 A Grammar of the Fuyuge Language. In Robert W. Williamson (ed.), *The Mafulu: Mountain People of British New Guinea*, 307–331, 336–344. London: MacMillian and Co.

Refsing, Kirsten
1986 *The Ainu Language: The Morphology and Syntax of the Shizunai Dialect*. Aarhus: Aarhus University Press.

Renault, Pedro Victor
1903 Exploração dos rios Mucury e Todos os Santos e seus affluentes – feita por ordem do governo da Provincia pelo engenheiro Pedro Victor Renault. [Coleccionada e organizada por Léon Renault]. *Revista do Archivo Publico Mineiro* VIII: 1049–1115.

Ribeiro, Michela Araújo
2008 Dicionário Djeoromitxi-Português: registro da língua do povo Jabuti. Universidade Federal de Rondônia, Guajará-Mirim masters thesis.

Rich, Rolland
1999 *Diccionario Arabela-Castellano* (Serie Lingüística Peruana 49). Lima: Instituto Lingüístico de Verano.

Richardson, Irvine
1957 *Linguistic Survey of the Northern Bantu Borderland* (Linguistic Survey of the Northern Bantu Borderland 2). Oxford University Press.

Rischel, Jørgen
1995 *Minor Mlabri: A Hunter-Gatherer Language of Northern Indochina*. Museum Tusculanum Press, University of Copenhagen.

Ross, Malcolm and John Natu Paul
1978 *A Waskia Grammar Sketch and Vocabulary* (Pacific Linguistics: Series B 56). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Sampaio, Wany Bernadete de Araujo
1997 Estudo comparativo sincrônico entre o Parintintin (Tenharim) e o Uru-eu-uau-uau (Amondava): contribuições para uma revisão na classificação das línguas Tupi-Kawahib. Universidade Estadual de Campinas masters thesis.

Santana, Áurea Cavalcante
2005 Transnacionalidade lingüística: a língua Chiquitano no Brasil. Goiânia: Universidade Federal de Goiás masters thesis.

Schebesta, Paul
1934 *Vollblutneger und Halbzwerge: Forschungen unter Waldnegern und Halbpygmäen*. Salzburg / Leipzig: Anton Pustet.

Schebesta, Paul
1966 Die Süd-Nyali oder bafuaNuma am Albertsee. *Wiener Völkerkundliche Mitteilungen* VIII: 37–54.

Schleicher, Charles O.
1998 *Comparative and Internal Reconstruction of the Tupi-Guarani Languages Family*. Madison: University of Wisconsin doctoral dissertation.

Schmidl, Marianne
1915 Zahl und Zählen in Afrika. *Mitteilungen der Anthropologischen Gesellschaft in Wien* XXXXV: 165–209.

Schubert, H.
1888 Das Zählen. *Dr. Neumayer's Anleitung zu Wissenschaftlichen Beobachtungen auf Reisen* II: 288–294.

Schultz, Harald
1959 Ligeiras notas sobre os Maku do paraná Boá-Boá. *Revista do Museu Paulista, N. S.* 11: 109–132.

Seibert, Uwe
1998 *Das Ron von Daffo (Jos-Plateau, Zentralnigeria): morphologische, syntaktische und textlinguistische Strukturen einer westtschadischen Sprache* (Europäische Hochschulschriften: Reihe XXVII: Asiatische und Afrikanische Studien 66). Frankfurt am Main: Peter Lang.

Seitz, Gitte
1993 Ikulu – Untersuchengen zu einer zentralnigeriansichen Klassensprache. Universität Hamburg masters thesis.

Seligmann, C. G. and Brenda Z. Seligmann
1911 *The Veddas*. Cambridge University Press.

Sharma, D. D.
2003 *Munda Sub-Stratum of Tibeto-Himalayan Languages* (Studies in Tibeto-Himalayan Languages VII). New Delhi, India: Mittal Publications.

Shimizu, Kiyoshi
1975 A Lexicostatical Study of Plateau Languages and Jukun. *Anthropological Linguistics* 17: 413–418.

Shimizu, Kiyoshi
1979 Five Wordlists with Analyses from the Northern Jos Group of Plateau Languages. *Afrika und Übersee* LXII: 253–271.

Shimizu, Kiyoshi
1982 Ten More Wordlists with Analyses from the Northern Jos Group of Plateau Languages. *Afrika und Übersee* LXV: 97–134.

Shimizu, Kiyoshi
1983 *The Zing Dialect of Mumuye: A Descriptive Grammar with a Mumuye-English Dictionary and an English-Mumuye index*. Hamburg: Helmut Buske.

Smith, Geoffrey P.
1988 Morobe Counting Systems. In *Papers in New Guinea Linguistics 26* (Pacific Linguistics: Series A 76), 1–132. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Smits, L. and C. L. Voorhoeve
1994 *The J. C. Anceaux collection of wordlists of Irian Jaya languages B: Non-Austronesian (Papuan) languages (Part I)* (Irian Jaya Source Material No. 9 Series B 3). Leiden-Jakarta: DSALCUL / IRIS.

Stirnimann, Hans
1983 *Praktische Grammatik der Pangwa-Sprache (SW-Tansania) / Indaki cha luchovo lwa vaPANGWA*. Freiburg: Universitätsverlag Freiburg.

Strauss, Hermann
1962 *Die Mi-Kultur der Hagenberg-Stämme im Östlichen Zentral-Neuguinea* (Monographien zur Völkerkunde / Hamburgisches Museum für Völkerkunde III). Kommissionsverlag Cram, de Gruyter & Co., Hamburg.

Struck, Bernhard
1910 Vokabularien der Bakondjo-, Baamba-, Bambuba-, Babira-, Balega-, Lendu- und Banyarisprachen aus dem linguistischen Nachlaß Emin-Paschas. *Mittheilungen des Seminars für Orientalische Sprachen* XIII: 133–165.

Stuhlmann, Franz
1894 *Mit Emin Pascha ins Herz von Afrika*. Berlin: Dietrich Reimer.

Stuhlmann, Franz
1916-1917 Wortlisten zentralafrikanscher Stämme. *Zeitschrift für Kolonialsprachen* VII: 257–308.

Swain, Carole
2000 Chapter 13: Quantity and Number: Xilixana. In Mary Ritchie Key (ed.), *South American Indian Languages* (Intercontinental Dictionary Series 1). Irvine: Computer Database on CD-ROM, University of California.

Thomas, Northcote W.
1920a Duodecimal Base of Numeration. *Man* 20 (32): 56–60.

Thomas, Northcote W.
1920b Duodecimal Base of Numeration. *Man* 20 (14): 25–29.

Thomas, Northcote W.
1927 The Bantu Languages of Nigeria. In *Festschrift Meinhof*, 65–72. Hamburg: L. Friederichsen.

Tormo, Jesús Galeote
1993 *Manitana Auqui Besüro: Gramática Moderna de la lengua Chiquitana y Vocabulario Básico*. Santa Cruz de la Sierra, Bolivia: Los Huérfanos.

van der Voort, Hein
2004 Review of The Amazonian Languages (R. M. W. Dixon and Alexandra Y. Aikhenvald, editors). *Anthropological Linguistics* 46 (2): 210–215.

van der Voort, Hein
2007 Proto-Jabutí: um primeiro passo na reconstrução da língua ancestral dos Arikapú e Djeoromitxí. *Boletim do Museu Paraense Emílio Goeldi: Ciências Humanas* 2 (2): 133–168.

van Geluwe, Harry
1960 *Les Bali et les Peuplades Apparentées (Ndaka-Mbo-Beke-Lika-Budu-Nyari)* (Ethnographic Survey of Africa: Central Africa: Belgian Congo: Part V). London: International African Institute.

Veerman-Leichsenring, Annette
2000 *Gramática del Chocho de Santa Catarina Ocotlán, Oaxaca*. Netherlands: Research School of Asian, African and Amerindian Studies (CNWS), Universiteit Leiden.

Velder, Christian
1963 Die Geister der gelben Blätter – ein Urvolk Thailands? Überblick über fünfzig Jahre Phi-Tong-Lüng Forschung. *Zeitschrift für Ethnologie* 88: 10–23.

Vellard, Jehan
1934, 1935 Les Indiens Guayakí. *Journal de la Société des Américanistes* XXVI: 223–292, XXVII: 175–246.

Vicedom, G. F. and H. Tischner
1943-1948 *Die Mbowamb: Die Kultur der Hagenberg-Stämme im Östlichen Zentral-Neuguinea* (Monographien zur Völkerkunde I). 3 vols. Hamburg: Kommissionsverlag Cram, de Gruyter & Co.

Vidal, Alejandra
2001 *Pilagá Grammar (Guaykuruan Family, Argentina)*. University of Oregon doctoral dissertation.

Viegas Barros, José Pedro
2004 Guaicurú no, macro-guaicurú sí. Una hipótesis sobre la clasificación de la lengua guachí (Mato Grosso do Sul, Brasil). Manuscript.

Villafañe, Lucrecia
2003 *Descripción de la lengua yuki*. Katholieke Universiteit Nijmegen doctoral dissertation.

Vinci, Alfonso
1956 *Samatari*. Bari: Leonardo da Vinci.

Vogt, P.
1902, 1903 Material zur Ethnographie und Sprache der Guayaki-Indianer. *Zeitschrift für Ethnologie* 34: 30–45, 35: 849–874.

von Chamisso, Adelbert
1837 *Über die Hawaiische Sprache*. Leipzig: Weidmannische Buchhandlung. [Also published in Philosophische, philologische und historische Abhandlungen der königlichen Akademie der Wissenschaften zu Berlin aus dem Jahre 1837: 1–79, 1839.]

von der Gabelentz, Georg. and A. B. Meyer
1882 *Beiträge zur Kenntnis der melanesischen, mikronesischen und papuanischen sprachen* (Abhandlungen der Philologisch-Historischen Klasse der Königlich-Sächsischen Gesellschaft der Wissenschaften 8 (4)). Leipzig: S. Hirzel.

Weir, Helen E. M.
1984 A Negação e outros Tópicos da Gramática Nadëb. Universidade Estadual de Campinas masters thesis.

Welmers, William E.
1950 Notes on Two Languages in the Senufo Group II: Sup'ide. *Language* 26 (4): 494–531.

Wilbert, Johannes
1962 Notes on a Sanema Vocabulary. *Journal de la Société des Américanistes* LI: 83–101.

Williamson, Kay
1973 *Benue-Congo Comparative Wordlist* volume II. University of Ibadan, Nigeria: West African Linguistic Society.

Wilson, Patricia
1989a Ambulas-Wingei Statement. Unpublished Manuscript.

Wilson, Patricia
1989b Brief Ambulas-Wosera-Mamu Statement. Unpublished Manuscript.

Wilson, Patricia
1990 Ambulas-Wosera-Kamu-K Statement. Unpublished Manuscript.
Wilson, Patricia R.
1976 Abulas dialect survey. In Richard Loving (ed.), *Surveys in five P.N.G. languages* (Workpapers in Papua New Guinea Languages 16), 51–79. Ukarumpa: Summer Institute of Linguistics.
Wilson, Patricia R.
1980 *Ambulas Grammar* (Workpapers in Papua New Guinea Languages 26). Ukarumpa, Papua New Guinea: Unpublished Typescript, The Summer Institute of Linguistics.
Wilson, W. A. A.
1961a Numeration in the Languages of Guiné. *Africa* 31 (4): 372–377.
Wilson, W. A. A.
1961b Outline of the Balanta Language. *African Language Studies* 2: 139–168.
Wise, Mary Ruth and Eduardo Riggle
1979 Terminología matemática y la ensenanza de conocimientos básicos entre los grupos étnicos de la Amazonía Peruana. *Lenguaje y Ciencias, Trujillo-Peru* 19 (3): 85–103.
Wolf, Teodoro
1892 *Geografía y geología del Ecuador*. Leipzig: F. A. Brockhaus.
Wolfers, Edward P.
1971 The Original Counting Systems of Papua and New Guinea. *The Arithmetic Teacher* February 1971: 71–83.
Wolfers, Edward P.
1972 Counting and Numbers. In Peter Ryan (ed.), *Encyclopedia of Papua and New Guinea* volume 1, 216–220. Carlton: Melbourne University Press.
Wolff, Ekkehardt
1974 / 1975 Sprachwandel und Sprachwechsel in Nordostnigeria. *Afrika und Übersee* LVIII: 187–212.
Zerries, Otto and Meinhard Schuster
1974 *Mahekodotedi: Monographie eines Dorfes der Waika-Indianer (Yanoama) am oberen Orinoco (Venezuela)* (Ergebnisse der Frobenius-Expedition 1954 / 55 nach Südost-Venezuela II). München: Klaus Renner.
Zide, Norman H.
1978 *Studies in the Munda Numerals* (CIIL Occasional Monographs Series II). Central Institute of Indian Languages: Grammar Series.

Additional rarities in the typology of numerals

Thomas Hanke

1 Introduction

This article deals with semantic and formal properties of numeral addition. Among the various types of quantifiers, there exists

> "the particularly privileged category [...] of *cardinal numerals* – quantifiers which refer to natural numbers, for example *one*, *two*, *thirteen*, *one hundred and twenty seven* [...]." (Gil 2001: 1278; emphasis in the original).

Numeral systems can be restricted to a few values, e. g. 1 to 3 or 4, or be practically unlimited: in languages like English, if the available numerals do not suffice to express a value, a new term is easily available, as seen in the growing use of *billion* and *trillion* in financial and other contexts.

Typological and other linguistic investigations of numerals have dealt with a wide range of topics, as evidenced by Greenberg (1978), Hurford (1975, 1987), Comrie (1997, 1999), and Gil (2001). A basic dichotomy distinguishes the external behavior of numerals and their internal make-up, i. e. the composition of numerals out of other numeral expressions. There exist some languages that lack complex numerals at all. Leaving that aside, higher numeral values are expressed by combining items of lower values.

The most widespread semantic relation between parts of numerals is addition, closely followed by multiplication. — Subtraction and division, which both employ higher values, are rarely employed. In fact, sums are in some languages the only semantic relation employed to construct complex numerals out of components.

Only recently, studies on numerals have employed worldwide samples that are explicitly described. As part of *The World Atlas of Language Structures*, Comrie (2005) offers a worldwide overview of "one aspect of the mathematical structure of linguistic expressions of numerals, namely the arithmetic base that is used in constructing numeral expressions" (2005: 530). In cases of doubt, "[...] preference has been given to the base that is most productive in the range 20–400" (Comrie 2005: 531).

In comparison to Comrie's survey, this article focuses on literally minor issues: paradigmatic structures which only occur below 10, and the conceptual origin of markers of additions. The database of this article is an exploratory sample of 281 languages (cf. Hanke 2005 and the overview on page 84 in the Appendix to this paper). In accordance with the overall theme of this volume, this article concentrates on relatively rare phenomena. I intend to show that they are important for a comprehensive typological overview of numerals, and that rare phenomena of all kinds may result from the same general principles as more frequent phenomena.

Section 2 presents basics of the linguistic investigation of numerals, and introduces the two main issues by means of English numerals. English and other European numerals are exemplary for the most common type of paradigmatic structure, the additive series, and the two most common sources of additive markers, namely conjunctive coordination and juxtaposition without any additive marker.

Section 3 deals with paradigmatic structures in numeral systems (or briefly: *numeral patterns*). The most common type of pattern in both addition and multiplication is the *series*, i. e. a sequence of numerals in which a fixed value – the *additive* or *multiplicative base* – is added to or multiplied by a sequence of values. An example is the additive series *twenty-one* to *twenty-nine*, where the base 20 is combined with the sequence 1 to 9. Serial patterns can be labeled *asymmetric* because they assign a privileged status to the base. Series are the predominant type of numeral patterns and in most languages the only one. Still, there exist two other, much less frequent types of additive patterns. Section 3 will show that those non-serial patterns can be explained in terms of general principles, too.

Section 4 deals with the origin of additive markers, with emphasis on less frequent conceptual sources. This domain has not received much typological attention since Greenberg (1978), who cites data from approximately 80 languages in the whole article.[1]

To offer an outlook, a major result is that additive markers are not only simple morphemes, as Greenberg's term *additive link* suggests, but also derived from complex constructions consisting of several morphemes. The attested range of sources is more varied, mainly due to a broader range of spatial source concepts in addition to Greenberg's *superessive links* ('on top of, above'). Furthermore, the role of comitative and coordinative constructions for the most common type of additive marker, "comitative links" in Greenberg's terms, is investigated. The distinction of these two related source types

is important for the identification of additive constructions as either *symmetric* or *asymmetric*: with the exception of juxtaposition and coordination, all means of addition are asymmetric, because their original use assigns different status to the involved entities. To illustrate this, 'X and Y' is equivalent to 'Y and X', but 'X on Y' differs from 'Y on X'.[2]

The extended classification of sources of addition enables to thoroughly test a hypothesis, which was first formulated by Greenberg (1978: 266): asymmetric means of addition are expected to occur only in asymmetric additive patterns, i. e. series.

The concluding remarks (Section 5) emphasize the typological relevance of less frequent phenomena and the role of sufficiently large exploratory samples in minimizing the risk of missing those rare phenomena. Altogether, the present results concerning rare phenomena in fact strengthen the role of common human practices, namely expressing abstract notions by concrete situations and, more specific to numerals, counting by hand.

A final remark: this article does not discuss the *order* of numeral constituents as a topic of its own. The order of numeral parts results from the interplay of the employed extra-numeral constructions (cf. Section 4) and tendencies of use: in sums, constituents with larger values tend to precede those with smaller values. This practice allows one to identify the magnitude as early as possible, cf. *two thousand and nine* in contrast to **nine and two thousand*. See Section 3.2.2 of Hammarström (this volume) for a brief discussion of that topic.

2 The linguistic analysis of numerals

After brief comments on my exploratory sample, Section 2.2 briefly discusses basic properties of whole numeral systems. Section 2.3 explains the analysis of single numerals and the recognition of numeral patterns, illustrated by English numerals.

First of all, a short comment on specific conventions employed in this article. In examples, the meaning of a numeral – its value – precedes it without quotation marks, e. g. 13 *thirteen*. Divergent forms of value expressions like *ten*, *-teen* and *-ty* are not a central issue of this article and therefore not distinguished in glosses.

For convenience, I employ Arabic numerals, but it is important to keep in mind that they are only a convenient means for naming values. The formal makeup of Arabic numerals, namely their regular base 10 system, has

no significance for the structure of the analysis of numerals. I occasionally use square brackets to illustrate the internal make-up of numerals, e. g. to clarify that *twen-ty* to *nine-ty* '2-10 … 9-10' has the serial structure '[2…9]-10'.

2.1 The sample

This article relies on a database for a fine-grained typological analysis of the internal structure of numerals (cf. Hanke 2005). It is combined with examples from previous works and some recent data. Some languages are not included in the sample because the available descriptions showed no relevant distinction to another language in the same genealogical group, an example for this being Dutch and German.

Australian languages are not included in the sample, because most systems have not been described adequately and are no longer in use now; the descriptions I am aware of conform to types attested elsewhere, mainly pair constructions and additive series based on hands (cf. Sections 3.1 and 4.6; Lynch 1998: 249; William MacGregor, p. c. March 2004).[3]

To my knowledge, the numeral systems of Pidgins and Creoles are usually more regular than those of the respective lexifier languages. This article does not deal with Sign languages either because I have not collected systematic data.

The sample is explicitly exploratory with the aim of providing maximal variation. It does not claim representativeness or unbiasedness in a statistical sense. As seen by Comrie's (2005) survey of numeral bases, multiples of 5 and 10 as additive and multiplicative bases show an overwhelming worldwide preponderance. This can be related to cultural and economic influences, namely the reduced need to express larger values in cultures with comparatively simple economic systems — in this respect, it is worth to mention the readiness to invent or borrow new numerals with higher values as soon as any need for them arises.

For additive constructions, the bias is arguably smaller. Still, it is likely that certain conceptual or formal types are less common and widespread in today's worldwide linguistic situation than in previous eras. A probable example are additive constructions of high complexity, namely those with reference to counting by hand. They were likely more widespread some centuries or millennia ago, while today they are mainly attested in cultures with

reduced needs for higher numerals. For this reason, all statements on the relative frequency of phenomena mainly refer to the perceived distribution rather than the exact number of languages or families.

2.2 Basic properties and a context-neutral definition

Entire numeral systems exhibit some properties which are nearly ubiquitous:

Continuity of the expressed meanings: the numerals of a system form almost always a gapless sequence from 'one' to the upper limit of a numeral system. For both, *gaps* and the concept of an *upper limit*, Comrie (1997: 42–45) offers an exemplary discussion.

One-to-one correspondence of form and meaning: numeral polysemy is nearly non-existent, i. e. a given expression has only a single numeral meaning. Synonymy between unrelated numerals – i. e. beyond minor variations such as *one hundred (and) one* is more common than polysemy but still rare.

Uniformity of series with different functions: numerals exhibit the same internal make-up in different syntactic functions.

These properties can be related to the especially simple semantics of numerals: gaps would hinder the practical use of numerals, and there is not much need for variation. As useful as these properties may be, for each of them a few exceptions are attested (cf. Comrie 1997; Hanke 2005: 23–29).

A handy guide to exclude alternative expressions like *a pair*, *a dozen*, *two times six* is that only numerals proper are normally used in the counting sequence.

To abstract away from the different form of numerals in different contexts, it is advisable to employ an operational definition of the numeral '*N*' in a given numeral system as the construction with the meaning '*N*' (cf. Hanke 2005: 46–47). This construction is the common component of all constructions expressing '*N*', regardless of their respective contextual properties. Such an operational definition renders the problematic term *cardinal* unnecessary.[4] In the remainder of this article, I employ the term *numeral* and disregard any context-dependent constituents such as gender or case markers.

2.3 From single numerals to patterns

The analysis of numerals operates on three interrelated levels:

- the formal structure of single numerals,
- the semantic relation between the value of a numeral and its parts,
- patterns that generalize over the structures of single numerals.

The identification of numeral constituents (Section 2.3.1) is comparable to other linguistic domains.

The semantic analysis of numerals (Section 2.3.2) is peculiar because the relations are nothing but exact basic arithmetic operations. Nevertheless, some systems show additive patterns that are more than minor deviations from the well-known base systems of English and most other languages. Section 2.3.3 illustrates the recognition of patterns. The mentioned overviews contain more information, e. g. on the interaction of additive and multiplicative patterns in base patterns.

2.3.1 The internal structure of single numerals

While every numeral system has simple members, only a small minority lacks complex members such as *fifty* or *two hundred (and) one*. Even systems that are limited to values below 10 usually use complex numerals, as seen in Section 3 below. In principle, numerals consist of just two types of constituents:

- elements expressing values such as *seven*, *fif-*, *-ty*, or *two hundred*
- elements expressing relations between values: e. g. English *and* as mentioned above, or the German plural suffix *-en* in 5,000,000 *fünf Millionen*, which may be considered an exponent of multiplication.

Value elements are not necessarily used as free numerals themselves, as shown above by *fif-* and *-ty*. The immediate constituents of a numeral are not necessarily simple. For example, [*two hundred*] *(and)* [*twenty-three*] is a sum consisting of a product (*two hundred*) and another sum (*twenty-three*), optionally supplemented by an additive marker.

The semantic analysis of numerals is based on the values of sub-constructions and their relations. For semantic analyses, markers such as *and* are generally not required. This enables an independent investigation of both domains, and a subsequent search for correlations.[5]

Due to their simple semantics, it is sufficient to compare numerals among themselves. In the vast majority of cases, a given meaning is either expressed by the same form or just slightly divergent forms such as *three* and *thir-* or *five* and *fif-* (it is worth to note Indo-European numerals show higher irregularity than average in several respects):

> The identification of phonologically different stretches as representing the same number is of three different kinds. One of these is agglutinal-fusional [...]. A second is suppletion [...]. A further type, portmanteau expression, can be illustrated by RUSSIAN sorok 40. (Greenberg 1978: 263–264; original emphasis)

Russian 40 *sorok* is a portmanteau, because it bears no similarity to the word for '4' nor to the other decades. Portmanteaus are not dealt with in this article, since they are only relevant in additive-multiplicative base patterns: 40 *sorok* is used as part of a series of additive bases from 10 up to 90 to add 1 to 9. Aside from a very small number of portmanteaus, the identification of components of complex numerals can rely on the simple semantic relations between them.

2.3.2 Semantic relations

The semantic structure of numerals is of a unique simplicity because they conform to arithmetical relations, mainly addition and multiplication. Subtraction and division occur only in a few numeral systems, which can be related to their inverted direction from larger to smaller values. The status of exponentiation as a semantic relation is controversial. The transparent base sequence *million*, *billion*, *trillion*, ... is a rarity, which spread from Romance and Germanic languages.[6] Other languages employ base expressions which originate in non-numeral expressions of huge quantities, without such a regular pattern. According to the topic of this article, the remainder of this article concentrates on additive patterns.

Semantic analyses of numerals can rely exclusively on the meaning of value-bearing elements, whose vast majority has identical or just slightly modified forms such as *fif-*. At the same time, values of components can be cross-checked by arithmetical analysis: *five hundred* means 500, which is 5 times 100, *twenty-five* means $25 = 2 \times 10 + 5$. The identification of suppletive or divergent forms is simplified by such interrelations; many instances of suppletion can be explained as the use of different forms for different func-

tions. One example is *ten*, its additive allomorph *-teen* and its multiplicative allomorph *-ty*. For instance, 10 is identified as the meaning of *-teen* by equations like *four-teen*: $14 = 4 + 10$, and *fif-teen*: $15 = 5 + 10$.

Since the arithmetical analysis does not depend on the occurrence of morphemes expressing the semantic relations of values, these two domains can be analyzed separately. Accordingly, the discussion of additive patterns in Sections 2.3.3 and 3, and the discussion of additive markers in Section 4 are independent of each other. The following Subsection and Section 3 will show that the attested variation of complex numerals built up by addition can be well described by no more than three types of semantic patterns. To give an example, the value 8 is expressed by the following arithmetic relations: $4+4$, $5+3$, $10-2$, $2+2+2+2$ or 'double 4'. In contrast, some other conceivable structures are not attested, e. g. $*1+1+1+1+1+1+1+1$, $*6+1+1$, or $*9-1$.

2.3.3 Additive patterns

In general, most structures of complex numerals are not isolated but belong to patterns, e. g.
$\ldots\ 5+1, 5+2, 5+3, 5+4, \ldots$ or
$\ldots\ 3+3, 4+3, 4+4, \ldots$

Two or more numerals with the same make-up can be described by a superordinate construction filled by several values, e. g. '$5+[1\ldots4]$'. Nearly all complex numerals in nearly all languages may be subsumed under just three types of patterns: *serial patterns* and the two types of *non-serial patterns* described in Section 3 (in fact, even subtraction and division can be subsumed under serial patterns of addition and multiplication, respectively).

The most prominent additive pattern is the additive series, which occurs in the overwhelming majority of languages across the world. An example of an additive series is *twenty-one* to *twenty-nine*. An additive series consists of an *additive base* (or *augend* in Greenberg's terms) to which a sequence of numbers is added. In the example above, 20 is the additive base, expressed by *twenty*. In nearly all cases, the sequence of numerals starts with '1'. In principle, the minimal range of a series is just two adjacent numerals, but in practice the smallest series consist of 5, 10 or multiples of these numbers. The apparent reason for the ubiquity of those values is the readily available model of the five digits of human hands.[7]

Additive series are nearly always embedded in patterns with wider range. The second major type of numeral pattern – nearly as common as additive series – is the *multiplicative series*, in which a *multiplicative base* (or *multiplicand*) is multiplied by a sequence of numbers, as in English *twenty* to *ninety*.[8] The two types of series are very often combined, as in the range from *ten* to *ninety-nine.* For this range, 10 is the *additive-multiplicative base* (often simply called *base*, cf. Comrie 2005), because 10 and its multiples are the augends of the succeeding additive series.

A good account of additive-multiplicative patterns with one or more bases is Hurford's *Packing Strategy* (Hurford 1975; Hurford 1987: 242–250). The Packing Strategy relies on given base numbers like 10, 100 etc. A "non-technical version" (i. e. without relying on 1970s phrase structure rules):

> When two numbers are added or multiplied to express a higher number, the resulting construction is usually markedly unbalanced, in the sense that one of the numbers is much greater than the other, and languages tend strongly to maximize this kind of imbalance. [. . .] The principle involved, which seems intuitively natural until one ponders it, could be expressed as 'When forming an expression for a high number, pick the highest-valued expression available as a starting point, and then build on that.' (Hurford 1987: 242–243)

This description neatly captures the asymmetric nature of serial patterns: a fixed additive base establishes the magnitude of a numeral, while one or another addend provides relatively minor details.

3 Additive patterns outside of series

While additive series are very common in the languages of the world, they are not the only type of additive pattern. As a reminder, an additive series is an additive construction with an open position that is filled by expressions for a sequence of values. The values are most often an uninterrupted sequence starting with 1, such as English *twenty-one* to *twenty-nine*. (Greenberg 1978: 265–266, 269).[9]

The following two subsections deal with two types of non-serial patterns, which have been largely left out in previous accounts of numerals. For instance, Greenberg (1978: 270) briefly speaks about "pairing systems" with a "pseudo-base" 2.[10] Hurford (1987: 250–251) refers to neo-2 structures, but only as counterexamples to his Packing Strategy. In another article, Hurford (1999: 13) records the frequent occurrence of pair sums in South Amer-

ica and New Guinea. As the following discussion will show, both types of non-serial additive patterns have characteristic features and can be compared to similar multiplicative patterns and to isolated numerals outside of patterns.

3.1 Pair sums

Pair sum constructions are defined as including no value higher than 2, which is the reason why some authors label them *binary*. In some languages, 4 is the only value expressed by a pair sum, as in Gapapaiwa (Austronesian, Oceanic): "The numerals up to five are unique forms, except 'four' which is 'two and two'." (McGuckin 2002: 301).

A comparable type of numeral construction is multiplicative 'pairing': '4' is sometimes expressed as a 'doubling' of '2', e. g. by 'dual' or 'pair'.

A remarkable phenomenon are systems in which '4' apparently is a reduplicated form, but obviously not of '2':

(1) Takia (Austronesian, Oceanic) (Ross 2002a: 226)
2 *uraru*
3 *utol*
4 *iwoiwo*

In other languages, pair sums constitute patterns of two or more numerals. A typical example for this is given in example (2):

(2) Amanab (Border; Gerstner-Link 2004: 2)

3 *sabaga mungu*
2 1

4 *sabaga sabaga*
2 2

5 *sabaga sabaga mungu*
2 2 1

6 *sabaga sabaga sabaga*
2 2 2

Quite often, pair sums are followed by additive series. An example is the following pattern in Koiari, which is followed by a series with the augend 5.

(3) Koiari (Trans-New Guinea, Koiarian) (Dutton 1996: 53)

3 *igau=ta abuti=ta*
1=and 2=and

4 *abuti=ta abuti=ta*
2=and 2=and

As in Koiari and Gapapaiwa, many numeral systems use pair sums only below 5, the base of the first additive series, cf. the overview of Papuan languages in Gerstner-Link (2004: 1).

The underlying principle of pair sums can be described as: "express a value out of bundles of two". It is viable to relate this to manual counting: either hand takes a counted item, this is repeated as often as necessary.

An alternative analysis treats pair patterns as additive series with a base 2. In any case, pair sums show a distinctive restriction to the values 1 and 2, which diminishes the asymmetry between the base 2 and the addends 1 and 2. For this reason, I regard pair sums as a separate pattern type.

3.2 Neo-2 sums

The so-called neo-2 sums consist of two summands which express the same value or differ by 1. I adopt this term from Hurford (1987: 250–251) and ultimately from Seidenberg (1960). While discussing the rareness of counterexamples to his Packing Strategy, Hurford quotes Seidenberg (1960: 227):

> Any system in which we find (with minor variations)
> (Type I): $6 = 2X3$, $7 = 6 + 1$, $8 = 2X4$, $9 = 8 + 1$
> or $6 = 2X3$, $7 = 8 - 1$, $8 = 2X4$, $9 - 10 - 1$ [sic! intended: $9 = 10 - 1$; TH]
> or (Type II): $6 = 3 + 3$, $7 = 4 + 3$, $8 = 4 + 4$, $9 = 5 + 4$
> . . . we will refer to as a Neo-2 system.

They are attested in some Niger-Congo languages, e. g. Ekoi:

(4) Ekoi (Niger-Congo, Bantoid) (Zaslavsky 1999: 48)

6 *esa-r-esa*
3-PLUS[11]-3

7 *eni-r-esa*
4-PLUS-3

8 *eni-r-eni*
4-PLUS-4

9 *elon-eni*
5-4

Another language group in which they are found are the Romani languages (Bakker 2001). What is more common than whole neo-2 patterns is an expression of '8' as '4 (plus) 4', with or without additional markers. Another type of neo-2 constructions is *doubling* to express 6 and 8. A doubling pattern may cover the whole range between 5 and 10 when 7 and 9 are expressed as '6 + 1' and '8 + 1', respectively, e. g. in Eastern Kayah Li (Sino-Tibetan, Karen; Solnit 1997: 204–205).

In any case, I have no evidence for neo-2 structures above 10, which is used as the first additive base.[12] This restricted range and the principle that neo-2 sums consist of two values with as little difference as possible can be related to manual counting: neo-2 sums express values as split up between two hands.

3.3 Finding underlying principles

All neo-2 structures and pair structures can be subsumed under *pairing*:

> The second [conceptual principle] is what Schmidt [1926] calls pairing ("das Paarsystem"), whereby the smallest quantity serves as a base for further counting: '3' is expressed as '2 + 1' [...]. Note, however, that addition is not the only arithmetic operation underlying pairing; pairing is also said to be present in languages that express, for instance, '6' as '2 times 3' [...] (Heine 1997: 19)

In fact, those binary patterns can be understood in a more physical way: while neo-2 sums split the value between two hands, pair sums can be seen as the verbalization of taking two items by both hands. This means that the rare additive patterns add to the evidence by the strong preference for multiples of 5 and 10 as additive bases: verbal counting has very often, if not always, its origin in physical, or rather manual counting.

Future research needs to establish the distribution of non-serial patterns in a larger sample of languages. There remain also some issues in the paradigmatic structure of numeral systems, first of all the few scattered instances of other additive patterns, e. g. *trinary* patterns with 3 as the essential number (e. g. 7 expressed as '3 + 3 + 1'). My database does not allow more specific generalizations, but even these patterns can be described by simple general rules, such as "bundle and stick to small values".

Altogether, owing to their exact meaning nearly all numerals are open for a very tight paradigmatic classification. The next section shows that this simplicity does not extend to the formal side.

4 Sources of additive expressions

My sample confirms that most sum constructions among the world's languages bear no marker of addition, but simply juxtapose the numeral components. As mentioned in the introduction, Greenberg's classification of additive markers (or *links* in his terminology) has been generally accepted without further research (cf. Hurford 1987: 237; Heine 1997: 33–34; Greenberg 2000: 777). This is the reason why the following sections are mainly a critical acclaim of Greenberg's insights. Greenberg (1978: 264–265) discusses the following types of additive markers as listed in Table 1.

Table 1. Greenberg's types of additive links

Type	Frequency
"Comitative" links ('and, with')	very common
Superessive links ('on, over')	less common, but in a wide variety of languages
Possessive links	rare
Expressions for 'extra'	rare
Expressions for 'left (over)'	rare

The given frequencies are derived from Greenberg's comments. "Comitative" is his cover term for conjunctive coordination and comitative constructions — the quotation marks around *comitative* point to the problematic status of this label (more on this issue in Section 4.1). This source is found in a wide range of families. The other types are dealt with in the subsequent subsections. Superessive links are assumed to be more common than the last three types, and are clustered in some areas like the Balkans. For the last three types, only isolated examples are given.

Before dealing with the different source types, my exploratory sample confirms Greenberg's frequency estimates, but reveals a much broader range of additive constructions. A major point is that Greenberg's typology does not mention complex constructions aside from multiple coordinators, which combine with more than one coordinated item. Nevertheless, a mere classifi-

cation can be supplemented by functional explanations and relations to other domains. This section concludes with an evaluation of Greenberg's generalization that asymmetric source constructions are only employed in serial addition.

A usual caveat applies to the typological classification and the generalizations based upon it. Like most large-scale studies, this study relies on grammatical descriptions. The sample includes relational elements whose original meaning is unknown or not mentioned in the description. They are treated with caution: I consider glosses like "plus" or "added" only as a reference to the numeral function, provided there is no further evidence that they are in fact e. g. derived from a verb meaning 'add'.

Fortunately, in practice this problem is restricted to one type of label: sometimes, components of numerals are labeled as 'and' to describe their additive function, but this does not relate to an origin in coordination. This is more obvious with items labeled as 'plus'.

I have tried to resolve all apparent issues by checking grammatical descriptions and, as a last resort, leaving out additive markers which lack sufficient evidence concerning their origin. It is sure that further data will be able to enhance the typology of additive constructions.

4.1 Coordination and comitatives

Greenberg characterizes this source type as follows:

> Almost as common [as no overt expression] is a formative, often affixed meaning 'and' or 'with.' In many languages a single morpheme means both 'and' and 'with.' We may call this a comitative link. Since the basic meaning is association, it may, on occasion, go with either of the two addends. (Greenberg 1978: 264)[13]

My sample confirms that coordinative constructions – quite often derived from a comitative – are indeed a common source of addition. Furthermore, Greenberg (1978: 276–277) compares multiple addition and complex coordination. Greenberg's main result that additive constructions show a similar distribution of coordinators in the same language remains valid.

Another issue is more interesting for the typology of additive constructions: during the data collection I realized that typological accounts of numerals do not cite any construction which is limited to accompaniment and addition, with the exclusion of coordination.

To date, the relation of coordination and comitative constructions has been examined in more detail (cf. Haspelmath 2004 (ed.), to name a whole volume devoted to coordination). A defining feature of (conjunctive) coordination is that it assigns the same syntactic and semantic status to two or more items. Comitative markers express accompaniment. They are a common source of coordinators. For distinguishing constructions with identical markers, symmetry vs. asymmetry of form and especially meaning is helpful, e. g. if you *go with someone to the super-market*, she need not walk herself, but may be carried by you.

Against this background, I tried to answer the following question: does any sum use a 'pure' comitative construction, i. e. a comitative construction that is not used for coordination? To answer this question, my analysis employs a simple procedure: additive constructions with an element glossed as 'with' or 'COM' or the like were checked to see whether they can also be used in nominal or verbal coordination.

This method can be demonstrated by data from So (Nilo-Saharan, Kuliak; Carlin 1993: 109–111). Carlin illustrates addition with examples such as the following (all glosses of *ka* are original):

(5) 6 *tud ka ni ɛdɛs*
5 with REL:SG 1

11 *tud enek ige ka in ɛdɛs*
5 hand:PL all with REL:PL 1

The relative construction is an indicator of the verb-like syntactic behavior of property expressions and numerals. In other contexts, *ka* is clearly recognized as a marker of coordination, e. g. for nominal properties:

(6) *tiisa lɔin ʃɛʃ ka in iman*

tiisa lɔ in ʃɛʃ ka in iman
have:1SG cattle REL:PL white and REL:PL black
'I have white cows and black cows.'

After applying this procedure, there were only two cases of a marker glossed as 'with' left which are not used for coordination. One of them is used in Northern Paiute (Uto-Aztecan, Numic; Thornes 2003: 212–213):

(7) 45 *watsikwɨʔ-yu manigi-ma tsɨpugi-dɨ*
4-ATTR 5-with release-NMLZ

A closer look reveals that *-ma* is an instrumental marker — the only occurrence of such a marker in my whole sample. In combination with 'released', this is rather like 'more' or a spatial source.

In Namia (Sepik, Yellow River; Gerstner-Link 2004: 6) the structure of the series 6…9 is only partially transparent, cf. example 8.

(8) 6 *napei-ran-waki-tija*
5-ACCOM-?-INDEF

7 *napei-ran-wal-pli*
5-ACCOM-place-2

8 *napei-ran-wal-mani*
5-ACCOM-place-3

9 *napei-ran-lwa*
5-ACCOM-4

The meaning of *waki* is unknown. '7' and '8' include the locative element *-wal*. Finally, Namia '9' is the only case of a "true" comitative in my whole sample of 281 languages. Even if, its form may originate in a shortened form, leaving out *wal* 'place'.

Since the *Rara & Rarissima* conference in 2006, I found another isolated occurrence in Hill Mari (Uralic, Finnic). The series 11…19 has the usual form '10-and-[1…9]'. For 15, there exists an alternative form *luckə*. Alhoniemi (1993: 92) explicitly mentions the uniqueness of the numeral 15, and proposes that its original form was *lu-βəc-ke* '10-5-COM'.

The high frequency of coordination and the extreme rarity of comitatives lead to the generalization that the latter are in nearly all cases only an indirect source of additive constructions. The result is a well-behaving extension of Haspelmath's (2004: 24) semantic map of coordination and related meanings: two unidirectional arrows lead from verbal and nominal coordination to addition, without any direct link to comitatives.

4.2 Superessive and other spatial constructions

Greenberg (1978: 265) identifies a single spatial source of addition:

> A third type is far less popular [than juxtaposition and "comitative links"; TH], but is still widespread: This is a word or affix meaning 'upon'. It will be called a superessive link. By its very meaning it would seem to go with the augend. If we add three items to ten, then the three are put on the heap of ten and not vice versa. 'Under' never occurs as a link.

Some examples are Eudeve (Uto-Aztecan, Cahita), Logbara (Nilo-Saharan, Moru-Madi), Welsh, Slavonic and other languages of the Balkan.

Greenberg adds the claim that "[b]y its very meaning it [a superessive link] would seem to go with the augend".

In fact, my sample reveals a variety of additive constructions with other spatial sources. To begin with, some Uralic languages actually employ a construction which refers to the additive base as 'under' the added values:

> The names of the teens (11–19) in northern, western, and eastern Vogul and in western Ostyak [Uralic, ObUgrian] were based on the model 'X (and) lying ten', e. g. So Vogul *kit + xujp + luw*, Kaz Ostyak *kat + xos^{j} + jaN* both TWO+LYING+TEN '12' [...]. (Honti 1998: 352)

Other constructions have a complex structure, including some which describe movements. The range found in the sample is illustrated by the following examples:

(9) Kresh (Nilo-Saharan, Kresh) (Santandrea 1976: 78–79)

16 *kpu-(i)ʃi-sal-uje-lem-'bala*
10-PLUS-5-place-ADD-1

(10) Kaulong (Austronesian, Oceanic) (Ross 2002b: 394)

12 *eip ponval u ponval me sup*
5 2 and_then 2 come inside

(11) Namia (Sepik, Yellow River) (Gerstner-Link 2004: 14)

260 *lulama pli powe popo ija lula napeiranwaki*
[200] after and 10 [6]

While in the present database these constructions are unique instances of a type each, they still conform to a general principle for spatial sources: constructions with spatial source always code the additive base like a reference object for position or movement.

Greenberg's observation that superessive constructions treat the addend as a physical object which is positioned relative to another, larger object is a special case of this generalization. The extended version holds for spatial expressions in combination with bodily expressions, too (cf. Section 4.6 below).

4.3 Possessive constructions

Greenberg's (1978: 265) examples for this additive source are Quechua (Quechua) and "the far-off Mountain Nubian" (Nilo-Saharan, Nubic), both with 11 expressed as "10 1-having". My sample contains a few constructions with similar meanings, including more complex constructions as in (12):

(12) Masalit (Nilo-Saharan, Maban) (Edgar 1989: 68–69)

26 *utuk utuk mbara gan toola iti*
10 10 2 ? it took 6

Utuk utuk mbara expresses 20, the part *gan toola* expresses that the base 20 'takes' the addend 6, although the origin of the first element of this part is unidentified.

4.4 'More'

Two rare types of links are mentioned by Greenberg (1978: 264) in a footnote: 'extra, added' and 'left'.[14] One example for the first type is 16 *tekvsmet'i* '10-6-more' in Georgian (South-Caucasian; Aronson 1991: 263–264). The sample includes a few similar cases, e. g.:

(13) Arosi (Austronesian, Oceanic) (Capell 1971: 50)

12 *e ta'i tangahuru ma adara rua*
ART 1 10 CONJ exceeding 2

(14) Iaai (Austronesian, Oceanic) (Lynch 2002: 780)

13 *li benyi-ta ke nua kun*
ART:DL hand-1INCL.POSS and again 3

4.5 'Left over'

Links with an original meaning 'left' (as in 'left over', not the opposite of 'right') are attested in a small number of languages. For this type of additive marker, Greenberg (1978: 265) made the strong claim that they always replace the additive base. His examples are *eleven*, *twelve* and cognate Germanic forms. They originally derive from a construction with the structure '1,2-left'. This claim is supported by evidence from the neighboring Indo-European group of Baltic languages, in which cognate forms are used up to 19, e. g. Lithuanian 19 *devynió-lika* '9-left'.

The exploratory sample includes negative evidence in Songhay languages:

(15) Koyraboro Senni (Heath 1999: 105)

84 *woy -yaaaha -tʃindi -taatʃi*
10 -8 -remain -4

This additive construction is used from 11 up to 99. If the additive base, e. g. 80 *woy-yaaaha*, were left out, **tʃindi-taatʃi* could express each value from 14 up to 94. Another counterexample is Ket which uses an expression for 'left over' in a similar way with (Bernard Comrie, p. c. March 2006).

This shows that it is doubtful to formulate strong generalizations on the basis of a few cases, especially if their independence is questionable and / or there are simpler explanations. The Germanic and Baltic numerals without base expression are just instances of one type of unexpressed values: the additive or multiplicative base of a series can be left unexpressed when its value can be inferred by the remaining elements, namely the addend or multiplicator and an expression of the semantic relation like 'left'. What Greenberg's original observation for Germanic 11 and 12 also illustrates is that only the smallest base of a series is ever left out, in this case 10.

4.6 Additive constructions bound to *bodily* patterns

Bodily augend expressions relate to the hands and often also to the feet. The augend values are nearly always 5, 10 and 15. Such series employ additive constructions from diverse sources. In addition, there exist specific constructions that are – still – common in South America and New Guinea, but also found in other regions, e. g. in Africa (e. g. So, cf. Section 4.1). Due to the limited length of this article, I illustrate the attested range of constructions with some examples from New Guinea:[15]

(16) Alamblak (Sepik, Sepik Hill) (Gerstner-Link 2004: 2)

6 *tir yohtt-i anakor tirt-ho ɾpat*
hand whole-CONJ other_side hand:3SG:F-POSS 1:3SG:F

(17) Moni (Trans-New Guinea) (Gerstner-Link 2004: 12)

11 *bado hago*
foot 1

16 *amo bado hago*
other foot 1

(18) Hatam (West Papuan, Hatam) (Reesink 1999: 59)

11, 12, . . . *simnai -bri -mig -[1,2,. . .]*
10 -go_along -leg -1,2,. . .

(19) Hua (Trans-New Guinea) (Haiman 1980: 314)

7 *dza' hatgi-di' bgo' rori hufu-na aidi'*
my:hand finish-INF more 2 jump-3SG:ANT hit:INF

4.7 The (more) complete list of additive sources

Table 2 summarizes the occurrence in my exploratory sample. As mentioned in the introduction, the given frequencies are tentative.

Table 2. Additive sources with approximate frequency

Additive source	Frequency
Juxtaposition	very common in all arithmetical types
Coordination	common in all arithmetical types
Comitative	extremely rare (2 isolated instances)
Superessive ('on, over')	scattered, clustered in several areas
Bodily constructions	common in some areas
'Extra, more'	rare (no known areal clusters)
'Left' and similar concepts	rare (no known areal clusters)
Possessive	rare (no known areal clusters)
Other spatial relations	rare tokens, altogether not very rare
Movements	rarer than stative spatial relations
'Under'	unique (a single instance)

Aside from juxtaposition and coordination, all source types share two properties which are logically independent: (i) exclusive use in serial sums and (ii) an origin in an asymmetric construction, which describes one entity in relation to another.

4.8 Correlations of sources and semantic patterns

The restriction of asymmetric source constructions to serial addition was first spelt out by Greenberg (1978: 266):

> Augends identified by serialization for any particular sum never disagree with those identified by superessive, possessive or other links which lend themselves to interpretation in this regard. The first notion is, however, the broader one. Further superessive or possessive links never occur in sporadic sums. This probably also holds for the other methods mentioned in footnote 4. ['extra, left over'; TH].

The first part of the quote repeats the insight that the additive base is treated like a physical anchor or ground for the smaller addend. This contains the conceptual explanation for the observed confinement of asymmetric means of addition to series as asymmetric patterns: addition is in principle symmetric, since the order of summands does not matter for the resulting sum: $5+3 = 3+5$. Only a component with a privileged position in a pattern invites an asymmetric expression.

According to my database, this generalization holds for all semantic types of "sporadic", i. e. non-serial, sums and all sources of additive constructions. Only coordinative markers and, of course, juxtaposition are found in all semantic types of sum, which can be related to the lack of asymmetry between the two value components.

There exists a type of exception that needs special consideration: a construction with an asymmetric source may become isolated by historical accident, e. g. if *eleven* (or *twelve*) were a single isolated case. By definition, it is not possible to identify the augend of a single numeral on its own. In any case, even in this hypothetical case, *eleven* could still be seen as part of the conceptual series 11 to 19, or as an isolated simple numeral.

An open question is the issue of correlations between additive sources and levels of base patterns, e. g. decades versus hundreds and further levels.[16] One obstacle to finding an answer is the rarity of most types, another the possible variation between coordination and sum constructions without a relational element. In my database, non-symmetric sources tend to occur more often with relatively low bases. In contrast, Hurford (2003: 48–49) observes for European languages, among several other trends, that juxtaposition is more common with smaller augends and that

> [s]ometimes, but not often, the presence or absence of a connective depends on the value of the lower conjunct. [...] There is a slight tendency across languages for a word for '1', as opposed to other 'digits', to trigger the use of an overt connective [...]

A well-known example is Standard French: 21 *vingt-et-un*, 31 *trente-et-un* ... with the coordinator *et* vs. juxtaposition in 22 *vingt-deux*, 32 *trente-deux*, ... According to my database, this is very rare outside of Europe.

5 Concluding remarks

Section 3 has shown that human languages employ only a few types of additive patterns. In comparison, the list of additive sources in Table 2 on page 80 may appear lengthy and open-ended. In fact, there are advantages beyond a mere listing of all attested phenomena.

First of all, this extended classification allows one to establish and test typological generalizations. As usual, this led to the generalizations being maintained, e. g. the existence of 'more' and coordination as sources of addition, or falsified, namely Greenberg's claim on left out bases with 'left over'. A third and prominent alternative is the modification of generalizations, as seen in the establishment of a more general spatial type subsuming Greenberg's superessive type and the specification of Greenberg's *comitative link* as markers originating in coordination. Based on the present extended classification, it is not only possible to confirm the restriction of asymmetric source constructions to serial sums, but Greenberg's idea gets a well-defined fundament: if comitatives were employed in non-serial addition (as assumed by Greenberg), their lack of symmetry would do away with the confinement of asymmetric expressions to series. Indeed, the two isolated cases of comitative sources in my sample occur in serial patterns and therefore fit into this general pattern.

Second, the general linguistic insight holds that abstract grammatical notions are expressed time and again by means of a limited range of concepts. It is possible to maintain a small number of classes by extending Greenberg's superessive type to a more general spatial type. More revealing is the recourse to the practice of counting, as discussed by Hurford (1987) and Heine (1997: 18–34).[17] These accounts consider only the sources of addition mentioned by Greenberg (1978): in series, the smaller amount is seen as put on the larger one, following it, or being left over after counting a "round number". The same reasoning applies to even the most complex constructions, whose translations sound like – admittedly elaborate – counting procedures, as illustrated in Section 4.

The rarity of "true" comitative sources of additive markers is remarkable from the perspective of grammaticalization research: Heine (1997: 34) in-

tends to use numerals as exemplary for the fact that spatial and comitative constructions are part of a small group of source concepts that are commonly employed for more abstract linguistic relations. For numerals, Heine relies on Greenberg's (1978) term "comitative" for the most common type of link.

Interestingly, nominal coordination is not generally known as an origin of grammatical constructions (cf. Heine and Kuteva (2002: 83–84), who only mention the pathway comitative > NP-and > S-and).[18] In any case, addition is a domain where a closer look reveals a secondary distinction between coordination and "pure" comitatives. The best explanation I can come up with is that a comitative construction is not a likely description of counting objects.

It is no new observation that the preponderance of serial patterns with bases 5, 10, 20 and further multiples relies on the omnipresence of the body, namely the hands and fingers. To complete the picture, there exists another pattern type based on counting by hand, the so-called *body-part counting* (body-tally-system). It is only known from languages of New Guinea, and is an extension of finger counting up the arm and around the upper body. This pattern type is not restricted to fingers, but includes a sequence of sources like 'thumb', 'middle finger', 'elbow', and 'nose'. In contradiction to widespread beliefs, there exists "no reason to suppose that the expressions of the body-part system are in any sense 'extra-linguistic'." (Comrie 1999: 83). They can even be combined with series to reach 100 and more, cf. Comrie (1999: 83–84), Gerstner-Link (2004: 17–18), and Hanke (2005: 83) for an apparent outlier on the Andamans).

The extension from 5 to 10 and from 10 to 20 is grounded both in the analogy of hands and feet and the general cognitive preference of binary structures. In serial patterns, the binary principle plays the same role as for any other hierarchical linguistic structure: there exists a strong preference for oppositions and pairs of items. A different instantiation of the binary principle are the non-serial patterns. In particular, neo-2 patterns that are only found in the range below 10 show another way that the binary principle and human physical conditions literally work hand in hand.

Altogether, this article has shown that exploring rare phenomena is worth the effort to get a (more) complete picture of a particular domain. As rare as some phenomena may be, it is worthwhile to consider their relation to more general principles. The present results concerning rare phenomena in fact strengthen the general role of common human practices, namely expressing abstract notions by means of concrete situations, in the case of numerals counting concrete objects.

Abbreviations

This article uses "1", "2", "3", etc. as numeral values. The only exception is the combination with number markers, such as '1SG'.

ACCOM = accompany; ANT = anterior; ART = article; ATTR = attributive; COM = comitative; CONJ = conjunction; DL = dual; F = feminine; INCL = inclusive; INDEF = indefinite; NMLZ = nominalizing; PL = plural; PLUS = additive marker of unknown origin; POSS = possessive; REL = relative marker; SG = singular

Acknowledgments

I want to thank the editors and Amy Budrikis for helpful comments on style and argumentation of previous versions of this article. All remaining mistakes are of course my own.

Part of the data reported in this article were collected during an internship at the Max Planck Institute for Evolutionary Anthropology and at the "Workshop on Numerals in the World's Languages", March 2004. I take this opportunity to express my gratitude for this opportunity!

Appendix

Sample of 281 languages, ordered by regions and language groups (full data in Hanke 2005, explanations in Section 2.3).

Language group	**Subgroup(s)**	**Language(s)**
Africa and Eurasia		
Afro-Asiatic	2	3
Kadugli	Kadugli	Krongo
Khoisan		!Xu
Niger-Congo	10	16
Nilo-Saharan	8	9
Altaic	3	11
Austro-Asiatic	3	3
Chukotko-Kamchatkan		2
Dravidian		6
Indo-European	4	15
Sino-Tibetan	4	4
Uralic	Finno-Ugric	7

Yeniseian	North	2
Northeast Caucasian		10
Northwest Caucasian		5
South Caucasian		3
Basque		
Japanese		
Nahali		
Oceania and New Guinea		
Andamanese, South		Jarawa
Austronesian	5	64
Border		3
Creoles	English Creole	Tok Pisin
Sepik	Middle Sepik	4
Solomons East Papuan		Lavukaleve
Trans-New Guinea	11	25
West Papuan		4
Others	10	10
The Americas		
Chibchan		3
Mayan		5
Misumalpan		2
Mixe-Zoque	Mixean	2
Oto-Manguean		6
Uto-Aztecan		8
Other (North)	11	13
Arawak		5
Aymaran		2
Guaviaré-Japurá		2
Huitoto		2
Other (South)	12	14

Notes

1. This is the main reason why Section 4 is first of all a detailed critique of Greenberg's seminal article, which was a major source of inspiration for my work, by means of more recent typological methods and a much larger sample.
2. In this respect, it is important to keep in mind that symmetry is in fact a criterion for coordination in contrast to comitatives.

3. Unfortunately, it is not rare to find remarks like "language X has no real numerals above five, just counting with the fingers" without any further elaboration or examples.
4. *Cardinal* is usually employed for attributive and / or counting numerals as the supposedly unmarked primary class of numerals (e. g. Greenberg 2000: 772). An issue is that both classes are often marked for some category such as noun class, case, person, or counting itself, for instance German *ein-s* 'one-COUNT'.
5. An exception is complex numerals in which a value is left unexpressed (cf. the survey in Hanke 2005: 53–55).
6. For opposing analyses of *million*, *billion*, etc. as base expressions and productive pattern, I refer to Comrie (1997) and Hurford (1975: 52–60), respectively. The expressions of different languages vary both in form and semantics above *million*, as evidenced by German *Milliarde* for 1,000,000,000 and the somehow dated British *thousand million* and *million million* (cf. Comrie 1997).
7. In fact, a few numeral systems use other base numbers, cf. Comrie (2005) and Hammarström (this volume).
8. *Ten* does not belong to this multiplicative series since it is neither constructed as a product (cf. **one-ty*) nor does it contain the same expression for 'ten' (cf. the hypothetical **ty*). The latter would suffice since 'one' deletion is quite common in products, cf. the variation of *(one) thousand*. Nevertheless, *ten* to *nineteen* belong to the additive-multiplicative pattern with base 10, which reaches up to *ninety-nine*.
9. An example of a series starting with 3 is the Germanic cognates of *thirteen* to *nineteen*: *eleven* and *twelve* are today best seen as simple numerals, while their historical origin is an additive construction with 'left over' (cf. Section 4.5).
10. From the examples, it is clear that Greenberg refers to neo-2 products like '2 times 3'.
11. 'PLUS' is used to gloss presumably additive markers whose origin is unknown.
12. Obviously, numerals like *nineteen* and *one hundred ninety-nine* are not analyzed as isolated neo-2 sums but as part of a series.
13. Greenberg diverges here from his own definition that 'addend' is the variable part of additive series opposed to 'augend'.
14. As mentioned before, I consider the gloss 'added' without further evidence as no more than a *description* of the numeral function.
15. Iaai (example (14) on page 78) shows that bodily sources are not limited to non-Austronesian languages.
16. An exception are body-related means that are apparently confined to the range below 20.
17. The conventionalization of numeral expressions (numeralization) is a good example that lexicon and grammar must not be separated (Hanke 2005: 111–119). Numeral systems consist of single items and paradigms, but there exists no a priori distinction between (more) concrete and (more) abstract meanings. Systematic structures are just the result of attempts to express the sequence of natural numbers, just like simple numerals.
18. Of course, this may be revised by further research on grammaticalization, especially with regard to more complex constructions. One interesting issue I am aware of are bipartite reciprocal constructions such as *isa-t isa* 'one-and one' in Tagalog (Austronesian, Meso-Philippine).

References

Alhoniemi, Alho
1993 *Grammatik des Tscheremissischen (Mari)*. Hamburg: Buske.

Aronson, Howard I.
1991 Modern Georgian. In *The Kartvelian Languages,* Alice C. Harris (ed.), 219–312. Delmar: Caravan.

Bakker, Peter
2001 Typology of Romani numerals. *Sprachtypologie und Universalienforschung* 54: 91–107.

Capell, A.
1971 *Arosi*. Canberra: Pacific Linguistics.

Comrie, Bernard
1997 Some problems in the theory and typology of numeral systems. In *Proceedings of LP '96, Prague*, Palek, B. (ed.), 41–56. Prague: Karolinum.

Carlin, Eithne
1993 *The So Language*. Cologne: Institut für Afrikanistik.

Comrie, Bernard
1999 Haruai numerals and their implications for the history and typology of numeral systems. In *Numeral Types and Changes Worldwide*, Gvozdanovic, Jadranka (ed.), 81–94. Berlin / New York: Mouton de Gruyter.

Comrie, Bernard
2005 Numeral Bases. In *The World Atlas of Language Structures*, Martin Haspelmath, Matthew Dryer, David Gil, Bernard Comrie, (eds.), 530–535, Oxford: Oxford University Press.

Dutton, Tom
1996 *Koiari*. München / Newcastle: Lincom Europa.

Edgar, John
1989 *A Masalit grammar: with notes on other languages of Darfur and Wadai*. Berlin: Reimer.

Gerstner-Link, Claudia
2004 *Typology of numeral systems in Papuan languages*. Paper presented at the Workshop on Numerals in the World's Languages, March 2004. Leipzig.

Gil, David
2001 Quantifiers. In *Language Typology and Language Universals / La typologie des langues et les universaux linguistiques / Sprachtypologie und sprachliche Universalien. An International Handbook / Manuel international / Ein internationales Handbuch*, Volume 2, Martin Haspelmath, Ekkehard König, Wulf Oesterreicher and Wolfgang Raible (eds.) (= Handbücher zur Sprach- und Kommunikationswissenschaft (HSK); 20.2), 1275–1294, Berlin / New York: Mouton de Gruyter.

Greenberg, Joseph H.
1978 Generalizations About Numeral Systems. In: *Universals of Human Language*, Joseph H. Greenberg (ed.), *Vol. 3: Word Structure*, 249–295. Stanford: Stanford University Press.

Greenberg, Joseph H.
2000 Numeral. In *Morphology. An International Handbook on Inflection and Word-Formation / Ein internationales Handbuch zur Flexion und Wortbildung*, Volume 1, Geert Booij, Christian Lehmann, Joachim Mugdan, in collaboration with Wolfgang Kesselheim, and Stavros Skopeteas (eds.) (= Handbücher zur Sprach- und Kommunikationswissenschaft (HSK); 17.1), 770–783. Berlin / New York: Mouton de Gruyter.

Haiman, John
1980 *Hua: A Papuan Language of the Eastern Highlands of New Guinea*. Amsterdam: Benjamins.

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

Hanke, Thomas
2005 *Bildungsweisen von Numeralia. Eine typologische Untersuchung*. (= Berliner Beiträge zur Linguistik; 3). Berlin: Weißensee.

Haspelmath, Martin (ed.)
2004 *Coordinating constructions*. (= Typological Studies in Language; 58.) Amsterdam: Benjamins.

Haspelmath, Martin
2004 Coordinating constructions: An overview. In: *Coordinating constructions*, Martin Haspelmath (ed.), 3–39. (= Typological Studies in Language; 58). Amsterdam: Benjamins.

Heath, Jeffrey
1999 *A Grammar of Koyraboro (Koroboro) Senni*. Cologne: Rüdiger Köppe.

Heine, Bernd
1997 *Cognitive foundations of grammar*. New York / Oxford: Oxford University Press.

Heine, Bernd, and Tania Kuteva
2002 *World Lexicon of Grammaticalization*. Cambridge, etc.: Cambridge University Press.

Honti, László
1998 ObUgrian. In *The Uralic languages*, Daniel Abondolo (ed.), 327–357, London: Routledge.

Hurford, James R.
1975 *The linguistic theory of numerals*. Cambridge, etc.: Cambridge University Press.

Hurford, James R.
1987 *Language and Number. The Emergence of a Cognitive System*. Oxford / New York: Oxford University Press.

Hurford, James R.
1999 Artificially growing a numeral system. In *Numeral Types and Changes Worldwide*, Gvozdanovic, Jadranka (ed.), 7–41. Berlin / New York: Mouton de Gruyter.

Hurford, James R.
2003 The interaction between numerals and nouns. In *Noun Phrase Structure in the Languages of Europe*, Frans Plank (ed.), 561–620. Berlin / New York: Mouton de Gruyter.

Lynch, John
1998 *Pacific languages: an introduction*. Honolulu: University of Hawai'i Press.

Lynch, John
2002 Iaai. In *The Oceanic Languages*, John Lynch, Malcolm Ross, Terry Crowley (eds.), 776–791. Richmond: Curzon.

Mcguckin, Cathrin
2002 Gapapaiwa. In *The Oceanic Languages*, John Lynch, Malcolm Ross, Terry Crowley (eds.), 297–321. Richmond: Curzon.

Reesink, Ger P.
1999 *A grammar of Hatam. Bird's Head Peninsula, Irian Jaya*. Canberra: Pacific Linguistics.

Ross, Malcolm
2002a Takia. In *The Oceanic Languages*, John Lynch, Malcolm Ross, Terry Crowley (eds.), 216–248. Richmond: Curzon.

Ross, Malcolm
2002b Kaulong. In *The Oceanic Languages*, John Lynch, Malcolm Ross, Terry Crowley (eds.), 387-409. Richmond: Curzon.

Schmidt, P. W.
1926 *Die Sprachfamilien und Sprachkreise der Erde*. Heidelberg: Winter.

Santandrea, Stefano
1976 *The Kresh Group, Aja and Baka Languages (Sudan). A linguistic contribution.* Napoli: Istituto universitario orientale.

Seidenberg, A.
1960 The Diffusion of Counting Practices. In *University of California Publications in Mathematics*; Vol. 3, Berkeley, 215–300.

Solnit, David B.
1997 *Eastern Kayah Li: grammar, texts, glossary*. Honolulu: University of Hawai'i Press.

Thornes, Timothy Jon
2003 *A Northern Paiute Grammar With Texts*. Doctoral dissertation. Eugene: University of Oregon.

Zaslavsky, Claudia
1999 *Africa counts: number and pattern in African cultures*. 3rd edition. Chicago: Lawrence Hill.

Explaining typologically unusual structures: the role of probability

Alice C. Harris

1 The problem

Among languages of the world, some morphological structures are common and expected, others are rare and unexpected. Still others may not be so infrequent as to be called rare, but they are at least uncommon. The rare or uncommon features include structures such as (i) multiple suppletion, (ii–iii) circumfixes and infixes (both found in Austronesian, for example), (iv) the coexistence of two or even three different patterns of case marking (as found, for example, in Georgian), (v) endoclitics (as in Udi), (vi) agreement morphemes repeated within a word (found in some Nakh-Daghestanian and Kiranti languages, for example), (vii) a case that marks non-focused nouns, but only in a subset of one gender (found in Dirayta), and (viii) tense marking on pronouns (found, for example, in Gurnu and certain other languages of Australia).

These rare and unusual structures pose a challenge for linguists. On the one hand, linguists must explain why they are rare or unusual; on the other hand, they must explain, given the rarity, why they exist at all. Some explanations that might be given are mentioned below.

- our innate endowment discourages this structure (perhaps as part of a more general feature)
- this structure does not function well
- this structure cannot be acquired easily by children
- this structure is not easily processed.

All of these statements may, in fact, be true. True or not, each one leaves us wondering how such a structure can exist at all.

I argue that these unusual structures are a product of the coincidences of history. They are rare, I suggest, because each requires a number of diachronic steps (changes) or conditions to develop. There is nothing necessarily rare or unusual about any of the conditions or steps, except the fact that

they are combined. Thus the rarity of these structures, I suggest, is a result of probability, the probability of exactly these changes and conditions coinciding in the order necessary. The fact that they exist at all is explained by the fact that they are not blocked by any mechanism.

2 On the explanation of the typological distribution of structures

As the simplest case, consider suppletion. Suppletion usually originates through a change linking two roots into a single paradigm. Suppletion involving two roots requires one change, while suppletion involving three roots requires two changes, and generally suppletion involving N roots requires $N-1$ changes. The approach outlined here predicts that suppletion involving two roots will be more common than suppletion involving three roots, and so on. It appears that this prediction is correct.

In a slightly more complex case, consider affixes. It is usually considered that prefixes and suffixes develop through two ordinary processes: first an independent word becomes a clitic, then a clitic becomes an affix. An infix, if one develops, generally develops out of a prefix or suffix (Yu 2007), thus requiring a third change. Circumfixes usually develop by linking a prefix with a suffix, thus requiring five changes (two to create the prefix plus two to create the suffix plus one to link them). Thus, the approach proposed here predicts that prefixes and suffixes will be more common than infixes, and that all three will be more common than circumfixes. Clearly, other things also need to be taken into consideration (especially the types of affixation already present in the language), and this method fails to predict that suffixes are more common than prefixes. Nevertheless the prediction made here is probably correct.[1]

One of the more complex examples I have discussed previously involves the Georgian case system (Harris 2008b); in this system, for a given verb, different cases may be required for subject and objects in different tense-aspect-mood categories. There are three such case subsystems, illustrated in (1a–c).

(1) a. *glex-i* *tesavs* *simind-s.*
peasant-NOM sows:I corn-DAT
'The peasant sows corn.'

b. *glex-ma* *datesa* *simind-i.*
peasant-ERG sowed:II corn-NOM
'The peasant sowed corn.'

c. *glex-s dautesavs simind-i.*
peasant-DAT has.sown:III corn-NOM
'The peasant has sown corn.' (adapted from Harris 1981: 1)

Although languages such as Hindi and Jacaltec have two case subsystems, as far as I am aware, the existence of three is unique to the Kartvelian family, to which Georgian belongs. I argue, following Harris (1985), that the current system developed historically from a relatively simple true ergative case marking system, through a reanalyzed antipassive, and through other specific changes, for all of which independent evidence exists. Each change involved is relatively common, but it is highly unusual for these changes to be combined.

Another example discussed in the same article is from endoclitics in Udi, a language of the Nakh-Daghestanian (North East Caucasian) family. In this language, clitics (meeting the criteria for distinguishing clitics from affixes) occur within a word between morphemes or within a morpheme (as in (2)), under certain morphosyntactic conditions.

(2) *jesir pasčaɣ-a bu-t'u-q'-sa ič ölkin-ä ta-ɣ-a-ne.*
captive king-DAT want$_1$-3SG-want$_2$-PRES self land-DAT thither-GO-SUBJVI-3SG
'The captive king wants to go to his (home)land.' (Dirr 1928: 67: 7)

In the first verb in (2), *buq'* 'want' is a single morpheme. As far as I am aware, intramorphemic clitics of this sort have not been reported for any other language,[2] and thus they may be considered very rare. I argue, following Harris (2002), that endoclitics in Udi originated through a series of individually common developments that included univerbation ("trapping" pronouns between two morphemes as they fused into a single word, cf. Watkins 1963) and analogy (Harris 2008b).

In Batsbi (Tsova-Tush) and some other Nakh-Daghestanian languages, many copies of the same agreement morpheme may occur in verbs (Harris 2009).

(3) *tiši^n c'a daħ **d**-ex-**d**-o-**d**-an-iš.*
old house(d/d).ABS PV CM-destroy-CM-PRES-CM-EVIDI-2PL.ERG
'Y'all are evidently tearing down the old house.'

It is shown that this, too, originated through changes that are individually common; probability predicts that such a large number will not often occur in a single language (Harris 2008a).

In a 1996 article, Tosco describes the very complex origins of a highly unusual subject case in Dirayta, an East Cushitic language. The case itself, the NFS case – non-focused subject – is unusual because (a) it involves only a subset of masculine nouns, not feminine, plural, or the other masculine nouns, (b) "those masculine nouns which are excluded from NFS marking are generally singulative in meaning", (c) "NFS case is expressed by marking masculine nouns with a feminine gender ending", and (d) NFS is synchronically the unmarked case (Tosco 1996: 28). Tosco lists the following seven changes as bringing this situation into existence:

1. Oromoid: Loss of the MASC definite marker *-*k*(V)
2. Konsoid: Extension of the FEM definite marker *-*t*(V) to MASC nouns
3. Dirayta: Reanalysis of the singulative -*itt* as definite
4. Konsoid: Loss of the definite value of *-*t*V, which becomes a frozen gender marker
5. Dirayta: Short vowel dropping in word-final position
6. Dirayta: Reanalysis of -*t* on MASC nouns as the NFS marker
7. Di[r]ayta: "Pruning" of -(V)*t* from MASC nouns not in NFS case role, and rise of a "new" ABS case. (Tosco 1996: 40)

This list does not even include the conditions which had to be present for this development to take place, though they are implied here. Although Tosco does not relate the complexity of this development to the rarity of the case, we may draw that connection. A case with the four characteristics listed above is rare because the co-occurrence of so many changes, each individually of a common type, is highly uncommon.

It is common for tense-aspect-mood (TAM) marking to be realized on verbs, but in some Australian languages it is realized also on nouns or pronouns, and this has been explored in some detail by Nordlinger and Sadler (2004). Blevins (2004) demonstrates that in one of these languages, Gurnu, at least five steps were required to get the tense from the verb onto the pronoun.

1. Earlier inflected verbs with regular phonological effects on following pronouns including initial C-loss under sandhi ...
2. Limitation of the general C-loss sandhi rule to this particular context ...
3. Failure of post-verbal pronouns to undergo complete grammaticization as bound morphemes ...
4. Reinterpretation of word boundary at the verb-pronoun boundary ...
5. Independent C-loss in pronouns giving rise to C-insertion for remaining V-initial bound pronouns ... (Blevins 2004: 310)

In this case, Blevins herself points out that the fact that so many distinct conditions and changes must occur in this order predicts that the phenomenon would be rare.

The examples given here illustrate the fact that in many instances common phenomena, such as prefixes or suffixes and simple suppletion, require few historical steps, while comparable but less frequent phenomena, such as infixes or circumfixes and multiple suppletion require more steps diachronically. Very infrequent phenomena such as those illustrated above from Georgian, Udi, Batsbi, Dirayta, and Gurnu require complex sequences of changes and conditions, though in most instances those changes and conditions are themselves common.

3 The Inadequacy of existing explanations

In order to explain the rarity of these and other structures, linguists have taken a variety of approaches, as indicated in the introduction to this article.

Greenberg has suggested an explanation of rare phenomena that may easily be confused with that I have proposed:

> In general one may expect that certain phenomena are widespread in language because the ways they can arise are frequent and their stability, once they occur, is high. A rare or non-existent phenomenon arises only by infrequently occurring changes and is unstable once it comes into existence (Greenberg 1978: 75).[3]

Perhaps Greenberg's explanation is right, but it now requires explanation of why certain changes occur infrequently, while others occur more often. That is, Greenberg's approach simply pushes the explanation back one level. My proposal differs from Greenberg's in that I am not suggesting that any change is necessarily infrequent, though that remains a possibility. I am proposing instead that any complex combination of changes or conditions will occur less often than any one of the changes or conditions that compose the combination, and that many rare phenomena can develop only through a complex combination of changes and conditions.

In his (2005) book, Newmeyer has discussed several proposed explanations for rare phenomena in language. He summarizes Chomsky's (1981) proposal as follows:

> [. . .] The idea was that typologically rare features should require a more complex formulation than typologically common ones, perhaps by requiring spe-

> cial marked parameter settings, by violating an implicational statement [. . .]" (Newmeyer 2005: 49, summarizing Chomsky 1981).

Many of the phenomena discussed briefly in Section 2 would indeed require more complex formulations of rules. Georgian case marking, because three different case patterns must be stated, requires a more complex set of case rules than do most languages (Harris 1981). Udi clitics similarly require complex rules, not only because they occur inside other morphemes, but because they occur in a variety of positions under different conditions (Harris 2002). Batsbi agreement markers, because they must be placed simultaneously in several different positions in a verb, require more complex rules. While the complexity of the requisite rules may seem to explain why most languages do not have three case systems, agreement clitics that are positioned variously, or repeated agreement within a single verb, it does not explain why these typologically rare phenomena do exist in these particular languages. If languages prefer simple systems, why do some languages have complex systems? Newmeyer himself (2005: 113–116) makes an important argument that "more complicated" is not always rarer, using stranding as his example. Stranding a preposition is simpler than moving it, yet preposition stranding is rare among the languages of the world.

The summary quoted above raises a partially independent potential explanation — specific patterns (in addition to or instead of complexity) may be discouraged by Universal Grammar, possibly through the mechanism of parameters or "implicational statements". Newmeyer (2005: 105) notes further that it is widely assumed that typologically pervasive patterns are "preferable" and therefore "favored" by Universal Grammar. But if Universal Grammar discourages certain systems, why would the languages I have named here maintain such a grammar, and indeed why would any language maintain an "undesirable" grammar?

Bowerman has suggested that structures (features, characteristics) that are rare in languages of the world are rare because they are difficult to acquire:

> One intriguing possibility is that the relative accessibility for children of alternative schemes for partitioning meaning in a given conceptual domain is correlated with the *frequency with which these schemes are instantiated in the languages of the world* [. . .] It is plausible that relative frequency is correlated with "ease" or "naturalness" for the human mind [. . .]" (Bowerman 1985: 1306; emphasis in the original).

Newmeyer (2005: 70) quotes Chomsky (1981: 9) to the same effect: "We would expect the order of appearance of structures in language acquisition

to reflect the structure of markedness in some respects", citing in addition Pinker (1984: 168–171). Hyams (1986), Newmeyer points out, holds the opposite view. Citing works on both rare and not rare late-acquired features and on early acquisition without regard to rarity, Newmeyer later (2005: 100) observes that there is little evidence to support the view quoted here.

Regarding Bowerman's hypothesis, Slobin (1997: 275) has observed that

> [. . .] we lack the necessary data and theory to evaluate it adequately. There are three kinds of problems: 1) The linguistic analysis that leads to the postulation of an accessibility or naturalness hierarchy is grounded only in the statistics of frequency of occurrence of grammaticized notions across languages. 2) There is no independent cognitive or psycholinguistic theory of what is "easy" or "natural". 3) It is a mystery that "difficult" or "unnatural" form-function relations are learned and used.

While the acquisition of many of the truly rare phenomena in language remain unstudied, one that has been studied is one part of the case marking in Georgian referred to in (iv). Imedadze and Tuite (1992: 104), in their survey of work on child acquisition of Georgian conclude "[. . .] that the presence of two distinct case-marking patterns used with different sets of verb forms does not present an especially difficult problem." It appears that there is no scientific support for any suggestion that other languages lack three systems of case marking because this would be difficult for children to acquire.

Hawkins and Cutler (1988) suggest, quite plausibly in my view, that the universal preference for suffixing over prefixing is explained by processing. The beginning of a word is easier to process than the end. They go on to suggest that the relative infrequency of infixing is likewise explained by processing (1988: 309). This may be correct, but as Hall (1988) has pointed out, processing alone does not explain how the numerical dominance of a more easily processed order is implemented. Further, if it is correct as I have suggested that creation of an infix requires one historical step more than creation of a prefix or suffix, it must be correct in any case that the infix would be less common than the prefix or suffix. It is not clear that the relative infrequency of infixes requires any additional explanation (though the preference for suffixes over prefixes clearly does).

One explanation that has been proposed for variation in case marking systems (e. g. ergative-absolutive vs. nominative-accusative, etc.) relates to the functional need to distinguish between subject and object (or agent and patient). [4] A similar argument might be constructed for the rarity of the multiple exponence in Batsbi illustrated in (3). It might be claimed that the function of

verb agreement is to indicate the subject and object (gender-class agreement in Batsbi is with the subject of an intransitive and the direct object of a transitive, regardless of their case marking). If this is the function of these markers, goes the argument, clearly that function is filled by a single marker, and additional markers are excessive and dysfunctional, and thus dispreferred. But if this argument were correct, why would Batsbi or any other language have a dysfunctional system? Why would multiple agreement such as that in (3) exist in any language?

All of the explanations discussed above predict that a rare feature would be quickly abolished, if it exists at all. If Universal Grammar indicates a preference for one structure over another, the dispreferred structure would surely disappear quickly. If children find a structure difficult to acquire, some will not acquire it, and it will vanish. If some structures are difficult to process, one would expect that they would be replaced by structures easier to process. And if some features are dysfunctional, speakers would surely rush to replace them with more practical structures. Yet, in some cases we know that very rare structures have existed for a very long time. For example, Georgian has been attested for more than a millennium and a half (since the fourth or fifth century C. E.), and through all of its history it has had feature (iv), three coexisting case systems, and feature (ii), abundant circumfixes. Udi dates from about the same period, and throughout its history it has had endoclitics (Schulze 2003). Because of their wide distribution in a family, we can infer that feature (iii), infixes, must have existed for a very long time in Austronesian (Blust 2003). Claims that a rare or less common phenomenon is difficult to process or acquire are not easily maintained in the face of structures that continue to be processed and acquired for hundreds of years. The same problem exists for explanations on the basis of innate endowment or functionality. These explanations cannot explain why a rare structure can continue for hundreds of years, if it is difficult to acquire or process or if it is discouraged by our innate endowment or by lack of functionality.

Any explanation based on rare structures being dispreferred makes the false prediction that these structures will not endure. In contrast, because the proposal made here is not based on any claim that the rare structures are in any sense dispreferred, it does not predict that they would quickly disappear. Among those discussed, the proposal made here is the only explanation that is consistent with the known long life time of some such features.

4 Conclusion

I have proposed an explanation based on probability, different from most of those in the typological and formal literature. These rare and unusual structures arise, I suggest, by means of a combination of the very ordinary changes that are found individually in language after language. What sets these rare structures apart is only the fact that establishing them requires more than one change or condition.[5] To some extent, the rarer a structure is, the more changes or conditions are required to build it. We may assume that this correlation would be perfect if every language were built "from scratch" with every generation. However, the correlation cannot be maintained very far, because the rare and unusual structure, once established, may continue to exist and be spread into many daughter languages.

The explanation proposed here has been criticized as being "obvious". That is, it is obviously true that a linguistic feature that requires many changes would be less frequent than a comparable feature that requires few changes. But this simple explanation has been overlooked in favor of more complex, sometimes less readily acceptable explanations. If it is obvious, this explanation should be part of the explanation of the distribution of rare and infrequent phenomena, but in fact this simple explanation is not ordinarily discussed in this context. A further problem is that any historical approach is often rejected out of hand as an explanation for any synchronic fact, without considering that in a case like this it is indeed "obvious".

None of the explanations discussed in Section 3 offers any way of approaching the real problem: If a structure is not part of our innate endowment, or is difficult to acquire or process, or does not function well, why is it that some languages do have the structure at issue? All of those explanations may be correct to an extent, but there remains for them the task of explaining how and under what circumstances a dispreferred structure gets past the barriers imposed by our innate endowment, functionality, acquisition, and processing.

Acknowledgments

The research reported here was supported by the National Science Foundation under Grant Number BCS–0215523. For comments on a related paper, I am grateful to my colleague Frank Anshen, as well as to participants at the conference "*Rara & Rarissima*", and to audiences where I have presented re-

lated papers, including those at the workshop "Explaining Linguistic Universals" held at Berkeley, at the Mediterranean Morphology Meeting 4, at the symposium "Linguistic Theory and Grammatical Change" in Bergen, Norway, at the University of Utrecht, and at Dartmouth College.

Abbreviations

In glosses of Georgian, Roman numerals represent the three tense-aspect-mood sets that govern different case systems. In glosses of Udi, subscripts indicate parts of a morpheme split by an intervening morpheme.
CM = class (gender) marker; DAT = dative; DM = default morpheme; ERG = ergative; EVID = evidential; NOM = nominative; PRES = present; SG = singular; SUBJV I = first subjunctive

Notes

1. This assertion is supported indirectly by the fact that most of the literature on relative frequency of affixation, such as Bybee et al. (1990), Hall (1988), and Dryer (2008), ignores infixes and circumfixes. The assertion is supported more directly by Greenberg, who mentions infixation, but refers to it as a "rare process" (1957: 92), and by Sapir, who observes that prefixing and suffixing are more common than infixing (1921: 72). Haspelmath (2001: 708) states that "infixes, circumfixes, and intercalated affixes appear only infrequently". And while there is a substantial literature on the theoretical treatment of infixes, including McCarthy and Prince (1993) and Yu (2007), circumfixes are rarely mentioned in the literature, apart from textbooks and descriptions of particular languages.
2. Intermorphemic clitics have been reported in Pashto (but see Kaisse 1981 and 1985) and in European Portuguese (Luís and Spencer 2005).
3. I am grateful to Jeffrey Good for calling this to my attention.
4. Song (2001: 157–167) provides a good summary of this position with respect to case marking.
5. See Whitman (2008) for a similar explanation.

References

Blevins, Juliette
2004 *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press.

Blust, R.
2003 Three notes on early Austronesian morphology. *Oceanic Linguistics* 42: 438–478.

Bowerman, Melissa
1985 What shapes children's grammars? In *The Crosslinguistic Study of Language Acquisition*, vol. 2, *Theoretical Issues*, Dan Isaac Slobin (ed.), 1257–1319. Hillsdale, NJ: Erlbaum.

Bybee, Joan L., William Pagliuca, and Revere D. Perkins
1990 On the asymmetries in the affixation of grammatical material. In *Studies in Typology and Diachrony: Papers Presented to Joseph H. Greenberg on his 75th Birthday*, William Croft, Keith Denning, and Suzanne Kemmer (eds.), 1–42. Amsterdam: John Benjamins.

Chomsky, Noam
1981 *Lectures on Government and Binding*. Dordrecht: Foris.

Dirr, Adolph
1928 Udische Texte. *Caucasica* 5: 60–72.

Dryer, Matthew S.
2008 Prefixing vs. suffixing in inflectional morphology. In *The World Atlas of Language Structures Online*, edited by Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, chapter 26. Munich: Max Planck Digital Library. Available online at http://wals.info/feature/39. Accessed on July 15, 2009.

Greenberg, Joseph H.
1957 Order of affixing: A study in general linguistics. In *Essays in Linguistics*, by J. H. Greenberg, 86–94. Chicago: University of Chicago Press.

Greenberg, Joseph H.
1978 Diachrony, synchrony, and language universals. *Universals of Human Language*, vol. 2, *Method and Theory*, Joseph H. Greenberg (ed.), 61–91. Stanford: Stanford University.

Hall, Christopher
1988 Integrating Diachronic and Processing Principles in Explaining the Suffixing Preference. In *Explaining Language Universals*, John A. Hawkins (ed.), 321–349. Oxford: Blackwell.

Harris, Alice C.
1981 *Georgian Syntax: A Study in Relational Grammar.* Cambridge: Cambridge University Press.

Harris, Alice C.
1985 *Diachronic Syntax: The Kartvelian Case.* (Syntax and Semantics, 18) New York: Academic Press.

Harris, Alice C.
2002 *Endoclitics and the Origins of Udi Morphosyntax*. Oxford: Oxford University Press.

Harris, Alice C.
2008a Explaining exuberant agreement. *Linguistic Theory and Grammatical Change: The Rosendal Papers*, Þórhallur Eyþórsson (ed.), 265–283. Amsterdam: Benjamins.

Harris, Alice C.
2008b On the explanation of typologically unusual structures. In *Linguistic Universals and Language Change*, Jeff Good (ed.), 54–76. Oxford: Oxford University Press.

Harris, Alice C.
2009 Exuberant exponence in Batsbi. *Natural Language and Linguistic Theory* 27: 267–303.

Haspelmath, Martin
2001 *Language Typology and Language Universals: An International Handbook*. Berlin: Walter de Gruyter.

Hawkins, John A., and Anne Cutler
1988 Psycholinguistic factors in morphological asymmetry. In *Explaining Language Universals*, John A. Hawkins (ed.), 280–317. Oxford: Blackwell.

Hyams, Nina
1986 *Language Acquisition and the Theory of Parameters*. Dordrecht: Reidel.

Imedadze, Natela, and Kevin Tuite
1992 The acquisition of Georgian. *The Crosslinguistic Study of Language Acquisition*, vol. 3, ed. by Dan Isaac Slobin, 39–109. Hillsdale, NJ: Erlbaum.

Kaisse, Ellen
1981 Separating phonology from syntax: A reanalysis of Pashto cliticization. *Journal of Linguistics* 17: 197–208.

Kaisse, Ellen
1985 *Connected Speech: The Interaction of Syntax and Phonology*. Orlando, FL: Academic Press.

Luís, Ana, and Andrew Spencer
2005 A paradigm function account of "mesoclisis" in European Portuguese. In *Yearbook of Morphology, 2004*, Geert Booij and Jap van Marle (eds.), 79–154. Dordrecht: Kluwer.

McCarthy, John J., and Alan S. Prince
1993 Generalized alignment. In *Yearbook of Morphology, 1993*, Geert Booij and Jaap van Marle (eds.), 79–153. Dordrecht: Kluwer.

Newmeyer, Frederick J.
2005 *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*. Oxford: Oxford University Press.

Nordlinger, Rachel, and Louisa Sadler
2004 Tense beyond the verb: Encoding clausal tense / aspect / mood on nominal dependents. *Natural Language and Linguistic Theory* 22: 597–641.

Pinker, Steven
1984 *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Sapir, Edward
1921 *Language*. New York: Harcourt, Brace and World.

Schulze, Wolfgang
2003 Caucasian Albanian (Aluan): The language of the "Caucasian Albanian" (Aluan) palimpsest from Mt. Sinai and of the "Caucasian Albanian" inscriptions. http://www.lrzmuenchen.de/%7Ewschulze/Cauc_alb.htm#intro.

Slobin, Dan I.
1997 The origins of grammaticizable notions: Beyond the individual mind. In *The Crosslinguistic Study of Language Acquisition*, vol. 5, *Expanding the Contexts*, Dan I. Slobin (ed.), 265–323. Mahway, NJ: Erlbaum.

Song, Jae Jung
2001 *Linguistic Typology: Morphology and Syntax*. London: Longman Linguistics Library.

Tosco, Mauro
1996 The strange case of the Dirayta subject case. *Journal of African Languages and Linguistics* 17: 27–45.

Watkins, Calvert
1963 Preliminaries to a historical and comparative analysis of the syntax of the Old Irish verb. *Celtica* 6: 1–49.

Whitman, John
2008 The classification of constitutent order generalizations and diachronic explanation. In *Language Universals and Language Change*, Jeff Good (ed.), 233–252. Oxford: Oxford University Press.

Yu, Alan C.L.
2007 *A Natural History of Infixation*. Oxford: Oxford University Press.

Right at the left edge: initial consonant mutations in the languages of the world

Pavel Iosad

1 Introduction

This paper presents a typological overview of initial consonant mutations (abbreviated: *ICM*) — pretheoretically, changes in the first consonant of a word which are not obviously caused by the phonetic / phonological context. They are relatively well-known from the modern Celtic languages such as Welsh (Ball and Müller 1992) and Irish (Ní Chiosáin 1991; Green 2006). The main goals of this paper are as follows:

1. Provide an overview of initial consonant mutation and initial consonant mutation-like phenomena attested cross-linguistically;
2. Examine whether initial consonant mutation can be thought of as a separate phenomenon or just a special case of (morphologically conditioned) phonological alternations;
3. Make some typological generalizations over initial consonant mutation (as long as it is possible on such a small set).

The paper is structured as follows. In Section 2, I propose an overview of initial consonant mutation cases attested cross-linguistically and try to define the concept of mutation more clearly. Section 3 presents a fuller inventory of languages exhibiting initial consonant mutations (or initial consonant mutation-like phenomena). Section 4 considers some issues in the phonological exponence of initial consonant mutation, with special reference to autosegmental approaches. Section 5 discusses factors which trigger mutation. Section 6 contains a brief discussion of initial consonant mutation typology. Section 7 concludes.

2 Overview of mutations

Initial consonant mutations are best known from the (Insular) Celtic languages Irish, Scottish, Manx, Welsh, Breton, and Cornish. Here, I will use

Welsh (see Morgan (1952); Thomas (1996) for very complete overviews) to illustrate some mutation types. The essence of the mutation process is that, in some morphosyntactic contexts, words appear with a consonant different from their citation form. This is shown in (1).

(1) a. ***t***ŷ
house
'house'

b. *dy* ***d***ŷ
2SG house
'your house'

c. *fy* ***nh***ŷ
1SG house
'my house'

d. *ei* ***th***ŷ
3SG.F house
'her house'

In (1), the word for 'house' shows up with an aspirated stop [t^{h}] in citation form (1a), with an unaspirated [d̥] after the 2SG possessive pronoun (1b), with a voiceless (aspirated) nasal [n̥^{h}] after the 1SG possessive pronoun (1c) and with a continuant [θ] after the 3SG feminine possessive clitic (1d). In all these cases, the change in the consonant is contingent on the identity of the preceding lexical item. In (2), on the contrary, it is both the preceding item (call it the *trigger*; in this case, the definite article) and the morphosyntactic properties of the word undergoing the alternation (the *target*) that drive the mutation: only feminine singular nouns demonstrate the change.

(2) a. *y* ***c***î
the dog (masc.)

b. *y* ***g****ath*
the cat (fem.)

c. *y* ***c***ŵ*n*
the dogs

d. *y* ***c****athod*
the cats

Finally, (3) exemplifies two cases where mutation is driven by rather general morphosyntactic conditions, rather than linear adjacency to a specific lex-

ical item. Specifically, the object NP *draig* 'dragon' undergoes mutation because of its position following the postverbal constituent, while the NP *dydd Llun* 'Monday' has its initial consonant changed because it plays the role of a temporal adjunct (Borsley, Tallerman and Willis 2007).

(3) *Lladd-odd Emrys* ***dd****raig* ***dd****ydd Llun*
kill-PST.3SG Emrys dragon Monday
'Emrys killed a dragon on Monday'

A fourth type of mutation is found in the isolate language Nivkh (Gilyak) spoken in the Russian Far East (Gruzdeva 1998; Mattissen 2003; Shiraishi 2006). Here, mutation is driven by the surface phonology, but blocked in certain syntactic configurations. Thus, (4) exemplifies a spirantization process: aspirated stops become continuants after (*inter alia*) another stop (irrespective of laryngeal specification). In (4a), the process applies when the verb is preceded by the object; but in (4b) it is blocked, since the verb is preceded by an adverb (verbal morphology is omitted in the source, which is Shiraishi 2006: 94).

(4) a. *ɲ-ɨmɨk tɨv-ux lep ɾ̥a-*
1SG-mother house-LOC bread bake-...
'My mother bakes bread in the house'

b. *j-ax niŋaq* ***q*** $^{\boldsymbol{h}}$ *o-ku-*
3SG-ACC a bit sleep-CAUS-...
'(She) let her sleep for a while'

I take these four types of phenomena to exemplify true initial consonant mutation. Before I give a fuller definition, I sketch some cases of what mutation is *not*.

First, I exclude cases where the first consonants of lexical words are caught up in across-the-board phonological processes whose domain happens to be larger than the (phonological) word. A well-known example is intervocalic spirantization in Tuscan Italian, or *gorgia toscana* (Giannelli and Cravens 1996). In this process, the voiceless stops /p t k/ are realized as continuants (normally [ɸ θ h], though there is a lot of variation). This process is active not only within words, but also in a clitic–host domain, as (5) shows:

(5) a. *casa*
house
'house'

b. *la [h]asa*
ART house
'the house'

I exclude such cases from direct consideration, since they do not target initial consonants specifically.[1]

I also refrain from considering most cases of consonant alternations involving the left edges of units smaller than the word; the ubiquitous post-nasal processes in Bantu and Austronesian languages (see Blust (2004) for a detailed overview of the latter) provide plenty of examples. Thus, in Kinyarwanda /r/ is realized as [d] following a nasal, as (6) shows.

(6) a. *uru-**r**imi*
CL6.SG-tongue
'tongue'

b. *in-**d**imi*
CL6.PL-tongue
'tongues'

While such "mutations" are initial in some sense, they are very often due to normal word-internal phonological processes, and thus, again, do not target first consonants specifically. Moreover, it seems that the majority of these cases do not involve any morphosyntactic information, which plays a conspicuous role in cases like Welsh and Nivkh.[2]

Based on the observed diversity of initial-consonant phenomena, in this paper I operate with the following definition:

> *Initial consonant mutation* refers to a change in the featural make-up of the initial consonant in a word, the context for which cannot be stated exclusively in terms of independently pronounceable phonetic or phonological entities.

This definition still leaves some leeway. In particular, what does it mean to be initial in a word? First, does it suffice to be initial only in some morphological forms? If yes, it is impossible to exclude Bantu languages like Tswana (Doke 1954), where a process normally associated with nasal-final prefixes is also triggered by a segmentally empty morpheme, as (7) shows.

(7) a. *go **b**ona*
INF see
'to see'

b. *go m-**p**ona*
INF CL1.OBJ-see
'to see him / her / it'

c. *∅-**p**ono*
CL9-seeing
'seeing'

Second, the concept of word is, of course, notoriously used with different meanings in various schools and with application to various domains.[3] I operate here with an intuitive notion of "word" as the actual instantiation of a lexical item, without committing to a particular stance. In the next section I list the languages which exhibit initial consonant mutation or initial consonant mutation-like processes.

3 The geography of initial consonant mutation

As mentioned above, the Insular Celtic languages present probably the best known case of initial consonant mutation (cf. any general overview, such as Ball and Fife 1993). It turns out, however, that initial consonant mutation or initial consonant mutation-like phenomena are attested more widely, at least in terms of linguistic areas.

3.1 Europe

Apart from the Celtic languages, initial consonant mutation-like phenomena are attested in *Romance*. In particular, the so-called "syntactic doubling" (*raddoppiamento fonosintattico*) in the dialects of Italy is a case in point (see Loporcaro (1997) for a full overview). While many cases of *raddoppiamento* are constrained by surface phonology (cf. D'Imperio and Rosenthall 1999), the gemination of the initial consonant is often contingent only on linear adjacency to a trigger. In the examples in (8) (from Standard Italian),[4] the first consonant of the noun is doubled after the clitic preposition *a*, but not after the article *la*: triggering the doubling is an idiosyncratic property of certain lexical items.

(8) a. *[k]asa*
house
'house'

b. *a [kː]asa*
at house
'at home'

c. *la [k]asa*
ART house
'the house'

Other dialects of Italy demonstrate more complex patterns: for example, in Neapolitan (Bullock 2001) initial [r] can double to [rː] or [dː], depending on its historical source (other historical voiced stops show similar behaviour):

(9) a. *ti [dː]ong-o*
2SG.OBL give-PRS.1SG
'I give (it) to you'

b. *tu me [r]aje*
2SG.DIR 1SG.OBL give.PRS.2SG
'You give (it) to me'

c. *se mett-e a [rː]ire-re*
REFL.OBL put-PRS.3SG to laugh-INF
'(S)he starts laughing'

Again, the choice of the doubled form is not determined solely by the phonological context: both the properties of the trigger and some abstract properties of the target influence the surface form.

Other Romance varieties also exhibit assimilation phenomena at the left edges of words; an example is Canary Islands Spanish (Oftedal 1985).

3.2 Africa

Due to the wide spread of prefixation in African languages, various assimilation phenomena at the right edges of prefixes or left edges of roots are also extremely common, and some of these fall under the purview of our definition of initial consonant mutation. This overview will by necessity be very brief and certainly not complete.

3.2.1 Atlantic

The systems of West Atlantic languages, first and foremost Fula (Pula(a)r, Fulfulde) are also often recognized as *comparanda* for the Celtic mutations

(cf. in particular Ternes 1990). There is an extensive literature on the mutation systems of Fula varieties (Arnott 1970; Paradis 1992; Breedveld 1995); the system of Seereer-Siin is treated by Fal (1980) and McLaughlin (1994, 2000); otherwise, Storch (1995) presents the most complete overview to date.

However, the majority of West Atlantic languages demonstrate only initial consonant mutation-like phenomena. While root-initial alternations are indeed driven by morphology (i. e. the triggering of alternations is an idiosyncratic property of prefixes independent of their surface phonology), alternating consonants are only seldom word-initial: for instance, in Bedik (Ferry 1968) only one of fifteen noun classes has a zero prefix, while in Seereer-Siin (McLaughlin 1994) this number is five (of sixteen classes).

Fula (Anderson 1976; Paradis 1992; Breedveld 1995), on the other hand, presents a more interesting case, since this language has no (surface) prefixes and thus the root-initial alternations are always by necessity word-initial. The next two examples show Fula mutation in the nominal and verbal systems respectively.

(10) a. ***p****ull-o*
Fula-CL1
'Fula person'

b. ***f****ul-ɓe*
Fula-CL2
'Fula people'

(11) a. *mi* ***w****ar-ii*
I come-PERF.ACT
'I have come'

b. ***mb****ar-ii-mi*
come-PERF.ACT-1SG
'I came'

In the nominal system, the noun class is sufficient as a predictor of the initial consonant: e. g. all nouns (as well as adjectives and determiners) belonging to class 1 (singular class for human beings) have a non-prenasalized stop as their first consonant if their lexical representation allows it (see below on the phonological exponence of mutation in Fula), while class 2 enforces continuants. In the verbal system, the initial consonant is determined by number and tense / aspect / modality features. In other words, Fula mutation can be described solely with reference to word-internal morphology —

unlike Celtic, where reference to word-external morphosyntactic contexts is unavoidable (see Section 5.2.1 for more extensive discussion).

3.2.2 (South-West) Mande

Most South-Western Mande languages (Mende, Kpelle, Loko etc.) demonstrate a process of initial mutation after a (historical) nasal (Kastenholz 1996; Vydrine 2004). There is some variation across the group: thus, in Kpelle the nasal is usually overt (12), while in Mende mutation is not triggered by any surface segment (13). The Mende examples are from Conteh, Cowper and Rice (1985).

(12) a. ***p***ólù
back
'back'

b. ḿ ***b***ólù
1SG back
'behind me'

(13) a. *ndòpóì ngúlɛ́í* ***gb***àndì-á
child oil hear-PERF
'The child heated the oil'

b. *ngúlɛ́í mìa ndòpóì* ***kp***àndì-á
oil FOC child heat-PERF
'The child heated *the oil*'

3.2.3 Bantu and other languages

Various prefix-root assimilations are extremely common in Africa, including the Bantu languages. Thus, in the Southern Bantu languages alone (zone S; Doke 1954), Nguni languages demonstrate alternations after prefix-final nasals, while all the three Sotho-Tswana languages, as well as Venda and Shona have alternations after surface-empty prefixes. For lack of space I do not treat them in detail: it is rather obvious, however, that such cases do not involve processes specifically targeting word-initial consonants.

3.3 Asia

In this paper I concentrate on two languages of this area: Nias (Brown 2001, 2005) and Nivkh (Gruzdeva 1998; Mattissen 2003; Shiraishi 2006) as well as briefly consider Burmese (Okell 1969).

3.3.1 Nivkh

This isolate language was briefly considered above in Section 2. As Shiraishi (2006) shows, in Nivkh certain phonological processes across word boundaries (involving changes in continuancy, aspiration, and voicing) are permitted only in certain syntactic configurations; in other words, reference to morphosyntax "outside" the target word is crucial.

3.3.2 Austronesian languages

We will mostly discount the various "nasal substitution" processes in Austronesian (see Blust (2004) for a very complete overview): they are essentially very similar to the the Bantu case discussed above.

One language, however, stands aside in this context. In Nias, mutation is both triggered by certain lexical items in an idiosyncratic manner (à la Celtic and Romance), as in (14), and used in certain morphosyntactic contexts, notably to mark absolutive case (15). All Nias examples are from Brown (2001), with an orthography modified to reflect the IPA more closely.

(14) a. ***f****oʔomo*
spouse
'spouse'

b. *i-be* *xɤ* ***v****oʔomo-nia nuxa s-oyo*
3SG.RLS-give PREP spouse-3SG cloth REL-be.red
'She gave red cloth to her husband'

(15) a. ***b****avi*
pig
'pig'

b. *la-bunu* ***ʙ****avi*
3PL.RLS-kill pig.ABS
'They killed a pig'

Brown (2001) interprets the latter type of mutation as induced by a case morpheme or clitic attaching to whole phrases (cf. the idea of clitics as the "morphology of phrases" in Anderson 2005). Nias mutation is probably the case most similar to the Celtic initial consonant mutation processes in terms of triggering.

3.3.3 Burmese

In Burmese (Okell 1969), causative verbs are formed by aspirating initial stops and devoicing initial sonorants:

(16) a. ***l**u?*
be.free
'be free'

b. ***l̥**u?*
CAUS:be.free
'to free'

Initial [j] also changes to [ʃ]; otherwise, it looks like a rather straightforward addition of a [+spread glottis] feature.

3.4 Australia

The non-Pama-Nyungan language Iwaidja (Pym and Larrimore 1979; Evans 1998) presents a case of root-initial alternations similar to those found in Atlantic, Bantu and Austronesian. In the majority of cases the alternations are caused by overt prefixes (though with no identifiable relation to their surface phonology), but in a very few cases the locus of alternation becomes word-initial. Table 1 on the facing page (from Evans (1998), with a modified orthography) shows the paradigm of the intransitive verb *ŋaḍbuni* 'fall'). The morphophonological alternation is between [k] and [g] / [ŋ]: the distribution between [g] and [ŋ] is irrelevant here.

However, as with the Atlantic, Bantu, and Austronesian cases referred to above, initial consonant mutation in Iwaidja can in no sense be said to target word-initial consonants specifically. It is entirely dependent on the identity of the preceding suffix, which Evans (1998) formalizes through an abstract *K* morphophoneme embedded in the suffixes' lexical representation.

Table 1. Mutation in an Iwaidja intransitive verb

Person	Singular	Plural
1	ŋa-**ŋ**aɖbuni	ad-**k**aɖbuni ŋad-**k**aɖbuni
2	aŋ-**ŋ**aɖbuni	gud-**g**aɖbuni
3	**k**aɖbuni	a-**ŋ**aɖbuni

3.5 Americas

There are at least four areas in the Americas where initial consonant mutation-like phenomena are attested: Numic languages (Uto-Aztecan), Mixe-Zoque, Mundurukú and Curuayá (belonging to the Tupi family), and various Macro-Jê languages.[5]

In the Numic languages, such as Comanche (Armagost 1989), Southern Paiute (Sapir 1930), and Chemehuevi (Press 1980), certain prefixes cause root-initial alternations not obviously related to their surface phonology, as the Comanche examples in (17) (Armagost 1989) show.

(17) a. ***p**uni*
'to see'

b. *waci**ʰ**p**uni*
'to spy on'

c. *na**β**uni*
'to see oneself'

However, these changes can only be triggered when a morpheme is present to the left of the target, and thus it seems that the alternation site is only word-initial precisely when no alternation happens.

The Tupi languages Mundurukú and Curuayá (Picanço 2005) present a case very similar to Nivkh: the alternation is phonologically driven, but blocked in certain morphosyntactic domains. Thus, in Mundurukú the segment [d] undergoes devoicing to [t] following noncontinuants (both nasal and oral), as (18) demonstrates in an inalienable possession construction.

(18) a. *dápsém **t**oj*
deer blood
'deer's blood'

b. *tawé* ***d****oj*
monkey blood
'monkey's blood'

However, this phonological process fails to happen in certain other morphosyntactic contexts: thus, for example mutation never happens after the alienable possessor prefix *-e-*.

In Zoque (Wonderly 1951; Akinlabi 1996) the 3SG possessor prefix is normally realized as palatalization of the first consonant. However, as Wonderly (1951) points out, the same effect is in evidence when a consonant is concatenated with a surface [j] (e. g. in compounding), and thus the "mutation" is in effect a surface-phonological process.

(19) a. i. ***f****aha*
belt
'belt'

ii. ***fʲ****aha*
3SG.POSS:belt
'his / her belt'

b. i. *poj* ***t͡s****ukum*
run go.out

ii. *po****t͡ʃ****ukum*
run:go.out
'run out'

A number of other South American languages (e. g. those belonging to the Macro-Jê family) use various methods of marking head-dependent adjacency (Rodrigues 2006). The exponence of this category varies widely, but in some languages it does include what looks like word-initial alternation (e. g. in Timbirá; Salanova 2004); for lack of space and data available to me, I do not concentrate on these cases further.

4 The phonology of initial consonant mutations

The phonological exponence of mutation has attracted much theoretical interest. Within classical generative phonology, there were several rule-based accounts, concentrating on the Celtic languages (cf. Ó Siadhail and Wigger 1975 for Irish, and Rogers 1972 for Scottish); much attention was also paid to

Celtic initial consonant mutation phenomena within Dependency Phonology (Ó Dochartaigh 1978; Ewen 1982). With the introduction of autosegmental approaches, many more initial consonant mutation phenomena were considered, usually within a wider context of morphologically conditioned alternations (cf. Lieber (1987), among others, for a wider survey including Welsh, Fula and Mende data; Swingle (1993) for Irish; Wolf (2007) for Breton). In this section I will present a brief review of some issues around the phonological exponence of mutation.

The general idea of the autosegmental approach is that initial consonant mutation is caused by a feature (or bundle of features) which are otherwise floating being attached to the left edge of a morpheme and inducing changes in the featural make-up of its segments. In the simplest case, this predicts that initial consonant mutation involves a consistent change in one or several feature values in all segments which bear a value for that feature. One such case is presented by Nivkh (Table 2).

Table 2. ICM in Nivkh

Type	Consonants									
Non-mutated	p^h	p	t^h	t	c^h	c	k^h	k	q^h	q
Mutated	f	v	r̥	r	s	z	x	ɣ	χ	ʁ

All Nivkh consonants with a [continuant] counterpart are involved in the initial consonant mutation process. Segments for which [cont] specification is redundant (namely all sonorants) are specified as [+cont] and thus cannot be affected by the prefixation of a [+cont] floating feature.

Even this relatively well-behaved system, however, needs a few remarks. The segment [r] does not behave like a sonorant, but rather as a fricative, in that it has a contrastive specification for [spread glottis] (for further arguments to this effect see Shiraishi 2006: 27).[6] Moreover, the coronal spirants [s] and [z] pattern with the palatal stops, rather than with the coronals.[7] The importance of facts such as these lies in demonstrating the importance of nontrivial featural specification (i. e. one based on the system of contrasts in the language rather than phonetic substance; cf. Dresher, Piggott and Rice 1994).[8] This problem is quite general, and not specifically related to initial consonant mutation, but we will see that this issue is a constant presence in other initial consonant mutation systems too.

Generally, initial consonant mutation and initial consonant mutation-like systems are "well-behaved" in this respect, in that the featural changes can to a large extent be captured by adding a feature. Consider the initial mutation system of Fula in Table 3 (Arnott 1970; Paradis 1992; Breedveld 1995; Koval' 2000).

Table 3. ICM in Fula

Grade	Consonants												
Continuant	w	b	r	d	j	ʔ/w/j	f	t	s	h	ɓ	ɗ	jʔ
Plosive	b	b	d	d	ɟ	g	p	t	c	k	ɓ	ɗ	jʔ
Prenasalized	mb	mb	nd	nd	ɲɟ	ŋg	p	t	c	k	ɓ	ɗ	jʔ

The names of the "grades" should be taken not as a literal description of the segments, but as labels referring to constellations of morphological contexts. Fula exemplifies a system where the segment appearing in one grade (mostly) allows predicting what segments appear in the others: this is why the "continuant" grade is normally considered to be underlying (e. g. by Breedveld 1995 and Paradis 1992). In an autosegmental account (cf. Lieber 1987), the "plosive" grade can be attained by prefixing a [−cont] feature and the "prenasalized" grade by whatever feature is responsible for this type of segment; the output will further need to be repaired to get rid of disallowed segments (e. g. prenasalized voiceless stops). However, a problem lurks with initial [w], [j], and [ʔ] in the continuant grade: each of these is ambiguous. It seems that some special marking is unaviodable: either morphological (i. e. different phonological changes in different morphological classes) or phonological (i. e. absolute neutralization; cf. Anderson 1976). Both of these are problematic.[9]

The final example comes from the Scottish Gaelic of Lewis (Borgstrøm 1940; Ladefoged et al. 1998). This variety presents a system that is largely typical of Scottish dialects in general. Table 4 on the next page presents it in a slightly simplified way (ignoring palatalization in obstruents and [m]); for a more general overview of phonological problems related to Celtic mutations, see Green (2006, 2007).

In this system, one problem is a chain shift: the fricative [f] deletes, but the stop [p^h] spirantizes to [f] without deleting (cf. Wolf 2007 for one autosegmental approach to chain shifts within parallel Optimality Theory).[10]

Table 4. Mutation in the Scottish Gaelic of Lewis (simplified)

Grade	Consonants													
Unmutated	p^h	t^h	k^h	p	t	k	f	s	m	$n^ɣ$	n^j	$l^ɣ$	l^j	$r^ɣ$
Mutated	f	h	x	v	ɣ	ɣ	∅	h	$v^{\sim}$	n	n	$l^ɣ$	l	r

Moreover, the deletion of [f] is a problem for an autosegmental approach in itself: how can *adding* a feature lead to *deleting* an entire segment? (See Bye and Svenonius (2009) for one possible answer.) Secondly, the mutation of obstruents (and [m], which patterns with obstruents in some other respects in this language) involves some sort of continuancy mutation: how can it be reconciled with changes involving velarization or palatalization of the sonorants? (This is less of a problem with approaches such as that of Ó Dochartaigh (1978) for similar phenomena in Irish, where the formal phonology directly leverages "more / less vocal tract opening".) Finally, this system also presents problems for straightforward phonetically-based features. Thus, if velarization and palatalization have uniform featural representations and mutation involves the prefixation of a single feature (bundle of features), it remains unclear why the relationship between mutated and non-mutated coronal nasals, laterals, and rhotics is so asymmetrical. Another similar problem is presented by the mutation of coronal obstruents: the addition of a continuancy feature to [d^j] might be expected to yield [$ð^j$]. That this does not happen could be taken as evidence for a highly ranked constraint against [$ð^j$]; however, this segment functions as the palatalized correspondent of [r] in this dialect. This is not a problem, of course, if surface [$ð^j$] is featurally specified as a sonorant.[11]

While very few problems that initial consonant mutation phenomena present to phonological theory are unique, it would seem that their importance might have been underappreciated, in light of the fact that initial consonant mutation phenomena are relatively rare and thus unknown to the non-specialist. At the same time initial consonant mutations present very clearly some of the issues that an autosegmental approach has to deal with, since no other solution within orthodox phonological theory seems to be available.

In the above discussion of autosegments, we have not touched upon the question of where they come from in the grammar. It is to triggering mutation that we turn in the next section.

5 Triggering initial consonant mutation

In this section I consider the nature of various contexts where initial consonant mutation is triggered cross-linguistically. I start with the simplest case, namely initial consonant mutation associated with specific lexical items, and then consider morphology and syntax as possible triggers for mutation.

5.1 Lexical triggers

Probably the most straightforward type of triggers is what I call "lexical" triggers. In this case initial consonant mutation is conditioned by the very presence of a specific lexical item (a word or non-null morpheme). Such triggers are attested in the majority of languages with initial consonant mutation or ICM-like phenomena: all Celtic languages, dialects of Italy,[12] some Mande and most Atlantic languages, the majority of Bantu and Austronesian languages, Iwaidja, Nias, and the Numic languages can be named here. Nivkh and Mundurukú / Curuayá (as well as the Jê languages, as far as I can tell) apparently stand out in this respect.

From the standpoint of theoretical phonology, such triggering presents very little problem. Assuming that the features are right (though as I have tried to show in Section 4 this is no easy task), we simply postulate the requisite feature bundle at the right edge of the trigger.[13] There are, however, several problems with this approach.

5.1.1 Locality

A necessary condition for "autosegmental" triggering is adjacency (at some tier): the autosegment ("morphophoneme") docks to the nearest available target (in the case of initial consonant mutation, to its right). Thus, we expect lexical triggers and targets to be adjacent. This, however, is not always the case. Thus, in Irish the numeral *dhá* 'two' and borrowed expletives (such as *fuckin'*) are "transparent" to mutations (Green 2007: 97–99).

(20) a. *ár* *[dʲ]each*
POSS.1PL house
'our house'

b. *dhá [h]each*
two house
'two houses'

c. *ár* *dhá [dʲ]each*
POSS.1PL two house
'our two houses'

In Breton, the word *holl* 'all' shows identical behaviour (Stump 1988): like Irish *dhá*, it causes a mutation when not preceded by a trigger, but when a possessive pronoun is present, it is the mutation required by the pronoun that surfaces. In Italian, parenthetical insertions can be transparent to *raddoppiamento* (Loporcaro 1996). It remains unclear how exactly such lexical triggering at a distance can be explained (Wolf 2007 sketches one such account, but does not flesh it out).

5.1.2 Restricted triggers

This subsumes cases where the mere presence of the trigger is not a sufficient condition for mutation. Ball and Müller (1992: 6–7) distinguish between *categorial* and *restricted* lexical triggers. In the former case, the same lexical item may trigger a mutation or fail to do so depending on its own category (thus, Irish *idir* causes mutation when it is a conjunction and does not cause mutation when it is the preposition 'between'); in the former, triggering is restricted by conditions on the target (thus, in Welsh the definite article only causes mutation on a following feminine singular word). Of course, accidental homophony can be proposed (i. e. two separate lexical items: *idir*[L] 'and' and *idir* 'between', where [L] is whatever triggers the mutation). For a full account of such restricted triggers, however, a discussion of morphology is unavoidable.

5.2 Morphological triggers

In this section I argue that an account which takes mutation-inducing autosegments or "morphophonemes" as exponents of morphological categories (along the lines of Wolf 2007) is not empirically and conceptually adequate for most cases of initial consonant mutation considered in this paper.

The majority of "morphological" initial consonant mutation triggering cases involve an explicit morpheme along with the segmental alternation, normally at the same edge as the alternation site (i. e. mutation is triggered by a prefix). This is the situation in Romance, most Bantu, Austronesian, and At-

lantic languages, Iwaidja, and many Mande languages. Under the classification proposed here, these are lexical triggers, since the autosegment is easily interpreted as being part of the prefix's lexical representation. In this section I will consider mutations which have been proposed to be the *sole* exponent of some grammeme, i. e. a subcase of *featural affixation* (Akinlabi 1996), itself a relatively rare phenomenon.

5.2.1 Fula

One such case is Fula. Recall that in Fula, alternations are word-initial (or root-initial: it is impossible to say which *prima facie*, since there are no prefixes in the language) but clearly related to grammemes normally expressed by suffixes. Most theoretical studies of Fula concentrate on the nominal system, where a class suffix is obligatorily present.[14]

(21) a. ***y**im-re*
song-CL
'song'

b. ***g**im-e*
song-CL
'songs'

In this case the standard analysis (Paradis 1992; Breedveld 1995) is to argue that the class morphemes are actually circumfixes: thus, for Breedveld (1995) the exponent of class in (21b) is ʔ-...-e. While this is not impossible, such pre-root elements are otherwise quite unknown in Fula: there is no *independent* evidence for such prefixes, and they are only postulated from theoretical premises (namely that something must be triggering the initial consonant mutation phonology).

Koval' (1997) provides further evidence against this approach. She notes that in some nouns the singular and plural are identical save for the mutation grade. A case in point is *yiit-e* 'fire', plural *giit-e*. Normally the class suffixes also undergo various alternations (Churma 1988; Breedveld 1995), and in this particular case two suffixes belonging to distinct classes are accidentally homonymous. Koval' (1997) notes that in some dialects the plural form is *giit-ee-li*, with two class suffixes (otherwise an anomaly[15]). According to Koval' (1997), such forms arise because initial consonant mutation is not sufficient to distinguish the singular and plural forms; in other words, sin-

Table 5. Fula verb paradigm (fragment)

Person	Full subject		Enclitic subject	
	SG	PL	SG	PL
1	mi **y**ar-ii	min **nj**ar-ii en **nj**ar-ii	**nj**ar-mi	— **nj**ar-ɗen
2	a **y**ar-ii	on **nj**arii	**nj**ar-ɗaa	**nj**ar-ɗon
3	o **y**ar-i	ɓe **nj**ar-ii	—	—

gular *yiit-e* and plural **giit-e* are non-distinct for the purposes of morphology. This generalization cannot be expressed in terms of any theory postulating mutation-inducing phonological elements as part of the morpheme: the morphemes for the two classes are then by necessity distinct.

A similar argument applies for the verbal system. The basic rule for mutation is as follows:

> In mutating verbs, the prenasalized grade of the initial appears (a) with a plural subject and (b) if the subject is an enclitic pronoun; else the initial appears as a continuant. [Koval' (2000): 170; my translation]

In general there is no subject-verb agreement in a Fula clause: the subject is only referred to once, either in a preverbal position (as a pronoun or full noun phrase) or cliticized to the verb (except the 3rd person and 1PL exclusive).[16] The "plural subject" rule applies regardless of subject type. Table 5 shows a fragment of a verbal paradigm (*yar-* 'drink').

If the prenasalization is due to some phonological material introduced in the morphology, then this material must be thought of as either forming a circumfix with the enclitic (e. g. 1SG *N-. . . -mi*) or representing an agreement morpheme. The former case is open to the same criticism as above, while the latter has the additional disadvantage of necessitating the introduction of a morphosyntactic process otherwise unattested in the language, namely number agreement: note that where agreement does happen (inside the determiner phrase), it proceeds by class, not by number.[17]

Summing up the discussion of Fula, I propose that initial consonant mutation-inducing phonological elements are problematic from two points of view. First, it raises phonological issues, as pointed out in Section 4, since a phonological account necessitates absolute neutralization. Second, viewing the putative autosegments as parts of morphemes is problematic in that it pre-

supposes morphosyntactic phenomena otherwise unattested in the language (prefixation, number agreement between subject and verb).[18]

One alternative is to interpret the mutation process as reflecting distributional requirements on the selection of root allomorphs in certain morphosyntactic contexts (Carstairs-McCarthy 1987; Stump 1995). This is much less problematic from a grammatical point of view; the downside, as emphasized by Wolf (2007), is a missed generalization: the phonological regularity of the relations between allomorphs (as we have seen, in Fula these relations are quite regular) is relegated to an accidental similarity (yet cf. Booij 2002). If such an account is accepted for Fula, we must exclude it from the "core" of initial consonant mutation languages, since the mutation then specifically targets the *root*-initial consonant instead of the word-initial one.

5.2.2 Welsh

Another language for which a morphological account has been proposed is Welsh. Kibre (1997) argues that the so-called soft mutation is an exponent of a [FEM SG] morpheme within the determiner phrase (in Welsh, adjectives are mutated following a feminine singular noun or article (importantly, *all* adjectives in this position are mutated, e. g. if there is more than one dependent on a single noun), and feminine singular nouns are mutated following the article). This is, however, not an empirically correct generalization: prenominal adjectives are *ceteris paribus* not mutated, but they still agree in gender with their head (example adapted from Morgan 1952: 12):

(22) **(l)lom aelwyd*
poor.FEM hearth
'poor hearth'

In (22), the adjective *llwm* 'poor' is not mutated (i. e. it is not *lwm*, though this form *per se* is possible in other contexts: *i lwm ddyn* 'to a poor person'), but still takes the feminine singular form *llom*. Thus the lack of mutation cannot be ascribed to a lack of agreement in prenominal position. This can be compared to the situation in Nias (Brown 2001), where mutation marks S arguments of monovalent verbs and P arguments of bivalent verbs, and can thus be construed as a mark of absolutive case. However, if the argument is fronted (normal word order in Nias is VOS), the mutation does not apply: probably further study is needed to determine whether this lack of mutation

is due to lack of case assignment in the preverbal position or some special feature of the mutation process. In Welsh, however, the evidence seems clear: any putative initial consonant mutation-inducing prefix cannot be a marker of [FEM SG] agreement.

Another morphological marker proposed by Kibre (1997) is a "non-topic" prefix which is attached to all constituents not occupying the immediately postverbal position. While this solution is formally impeccable, Kibre (1997) himself notes that this "non-topic" marker is somewhat unprecedented cross-linguistically, and, moreover, is not actually related to the information structure (i. e. actual topics can occupy the "non-topic" position and vice versa). He proposes that the marker is the product of grammaticalization of a morpheme which historically played a part in the semantics-syntax interface, but is now irrelevant in that respect. Thus, this lexical item is in fact no less *ad hoc* than the autosegmental "clitic" proposed by Lieber (1987) for Mende, which is simply inserted in a given syntactic configuration for no semantic reason.

Roberts (2005) proposes an analysis of Welsh which interprets some mutation as an exponent of Case; see below for brief discussion, and (Borsley, Tallerman and Willis 2007: chapter 7) for a detailed critique.

5.2.3 Mende

One language where a genuine morphological solution seems available is Mende. In this language the nasal present in other related languages has disappeared from the surface, but the mutation remains. Various accounts of Mende mutation have been proposed based on syntactic configurations having to do with c-command and branchingness (Rice and Cowper 1984; Conteh, Cowper and Rice 1985), case (Seidl 2001), agreement (Tateishi 1990), and prosodic boundaries (Cowper and Rice 1987; Tokizaki 2005). However, all these accounts presuppose a reversal of the historical direction of phonological changes: they view the so-called "strong" grade as underlying, and thus the elsewhere case. In fact the strong grade is historically postnasal, and thus conserved in very specific morphological environments (see Hyman (1973); Dwyer (1986) for the history of these environments, and Kastenholz (1996); Vydrine (2004) for the phonological history). For lack of space I do not consider this question closer here: Vydrine (2004) hints at an account where the "strong" grade is derived and appears in a well-defined morphological environment.[19]

5.3 Syntax

Finally, we consider some cases where syntax is claimed to be directly involved in mutations: initial consonant mutation happens whenever a certain configuration of syntactic constituents is present. We have briefly discussed one such case in Mende, reinterpreting it as morphological.

Another example comes from the "direct object mutation" in Welsh (Ball and Müller 1992; Borsley, Tallerman and Willis 2007). In Welsh constituents following the first postverbal phrase undergo soft mutation (including objects in VSO clauses, whence the name). However, this type of mutation applies also to subjects (and does not apply to objects of impersonal verbs, which do not raise to subject position); the evidence is carefully discussed in Borsley, Tallerman and Willis (2007: ch. 7). An alternative to the case-based account by Roberts (2005) (cf. also Zwicky 1984) is the "XP-trigger hypothesis", originally advanced by Borsley and Tallerman (1996):

> XP triggers soft mutation on the initial consonant of the right-adjacent constituent which it c-commands.

If this is indeed the correct generalization, then it is not entirely clear how mutation can be triggered autosegmentally: a meaning-less morpheme inserted into a given syntactic configuration would be quite unprecedented (apart from the "clitic" that Lieber (1987) introduces for Mende).

Syntax is also involved in Welsh mutation in a different way. Disruptions of normal word order often cause mutation (Tallerman 1999): for example, all nouns which follow adjectives are mutated, irrespective of gender and number. Again, it does not seem obvious what (sub)segmental morpheme can produce such an effect.

Nivkh and Mundurukú present a different type of syntax-induced mutation (or perhaps mutation blocking): in these languages, no autosegment is required as the phonology of mutation patterns is entirely surface-based. However, mutation only happens in certain syntactic domains, and thus syntactic information is indispensable to the initial consonant mutation process. This raises all sorts of questions with respect to modularity and interfaces, in particular whether syntactic information (such as constituent structure outside the word) is available in phonology (cf. Hayes (1990) for one positive answer).

6 Typology

As I have tried to demonstrate above, initial consonant mutation, despite being quite rare cross-linguistically, present a great deal of variation. I now return to the question posed at the beginning of the paper: can we view initial consonant mutation as a unified phenomenon, and what are the relevant typological generalizations?

I propose that the first of these questions receives a positive answer in view of the apparent importance of the beginning of the word. Possibly underpinned by its importance in psycholinguistic respects (e. g. word recognition, cf. Marslen-Wilson and Zwitserlood 1989), the left edge of a word is commonly assumed to be special in one phonological sense or another (Beckman 1999; Smith 2002). In this connection it is remarkable that initial consonant mutation always involves exactly the left edge of the word, and only very rarely does it impinge on non-initial segments. Thus, for example, vowel-initial words are, to my knowledge, never affected by initial consonant mutation processes. There are also, as far as I know, no examples of an "anti-Chaha" pattern, where mutation is realized on the leftmost available consonant, which is not necessarily word-initial (Lieber 1987). In Chaha, morphophonological labialization is realized on the rightmost non-coronal consonant (there are no labialized coronals in the language): thus *dænæg*^{w} 'he hit him', but *q*^{w}*ætær* 'he killed him', *sædæd* 'he chased him'. Lieber (1987) considers this gap accidental; I propose, however putatively, that it is not, in that it shows that whenever something disrupts the beginning of a word, the first segment must always be included.[20]

With respect to initial consonant mutation typology, I propose a major distinction between *endocentric* and *exocentric* initial consonant mutation and ICM-like processes (a similar distinction is proposed by Ternes 1990). Endocentric alternations are those which can be described solely with reference to the word they are associated with (its phonology, make-up in terms of morphemes, lexical class, and morphosyntactic features). Thus, all alternations caused by prefixes (whether segmentally or autosegmentally) are endocentric. Fula (in the interpretation proposed in 5.2.1) is thus also a case of endocentric mutation, since the morphemic make-up and morphosyntactic features are sufficient to determine the mutation grade. Exocentric mutation, on the other hand, cannot be described without reference to some factors outside the word, such as the syntactic context (Nivkh, Mundurukú, Welsh) or the position of the word with respect to some other item (Nias, Welsh, Irish).

Furthermore, I propose that exocentric mutations form the "core" of the initial consonant mutation "space". This is because endocentric mutations are, by definition, associated with morphosyntactic elements otherwise present in the language. This means that postulating relevant morphemes or morphological processes is quite straightforward, and thus the mutations are not special, given an adequate theory of morphophonology.

With exocentric mutations, on the other hand, such initial consonant mutation-triggering elements can be problematic from the point of view of the overall grammar, since their insertion cannot be likened to the insertion of "conventional" morphemes. Whether this necessitates introducing special morphological operations on whole words (cf. Hayes (1990); Green (2007) and Stewart (2004) specifically for Scottish Gaelic) is just one question that the rare phenomenon of initial consonant mutation poses to linguistic theory.

7 Conclusion

Lack of space has prevented discussion of other initial consonant mutation-related topics, such as the relationship between the properties of triggers and targets; trigger interaction; the interplay of initial consonant mutation and surface phonology (cf. Stump (1987) for Breton); the role of paradigmatic factors (cf. Shiraishi (2004) for Nivkh); the role of morphology (e. g. do mutations affect roots, stems or whole words, cf. Stewart 2004) etc. However, I hope to have demonstrated some of the importance of initial consonant mutations *as rara*.

First, if the interpretation of exocentric mutations as the "core" type is accepted, they are in fact quite rare: it is only the Celtic mutations (though they present several distinct types), Nias, Nivkh, and Mundurukú that undoubtedly present a case of endocentric initial consonant mutation (depending on the interpretation, the dialects of Italy and Fula may also be part of this group).

Second, while endocentric initial consonant mutation is a more or less straightforward subcase of morphologically induced phonological alternations and does not seem to raise theoretical questions significantly different from other phonological phenomena, exocentric initial consonant mutations are an altogether different proposition.

Quite apart from the phonological difficulties, they pose interesting questions to an additive theory of grammar which derives phonological alternations solely from the juxtaposition of (quasi-)phonological elements (cf.

Booij (2002) for a different view). In that sense, they do indeed form a special – albeit very small – class of phenomena. Only further study will show whether initial consonant mutations are truly a *sui generis* phenomenon, necessitating amendments to our view of grammar, or simply the outcome of an unusual confluence of otherwise unexceptional factors.

Abbreviations

1, 2, 3 = 1st, 2nd, 3rd person; ABS = absolutive case; ACC = accusative case; ACT = active voice; CAUS = causative; CL(1, 2...) = class (1, 2...); DIR = direct case; FEM = feminine gender; FOC = focus marker; ICM = initial consonant mutation; INF = infinitive; LOC = locative; OBJ = object; OBL = oblique case; PERF = perfect tense; PL = plural; POSS = possessive; PREP = preposition; PRS = present tense; PST = past tense; REFL = reflexive pronoun; REL = relative complementizer; RLS = realis mood; SG = singular; VOS = verb–object–subject

Acknowledgments

Some of the material presented in this paper was included in my presentation at the *Rara & Rarissima* conference. I thank the audience in Leipzig for their valuable input, in particular Larry Hyman and Peter Bakker. I would also like to thank Vladimir Plungian, who was my supervisor during the project that this paper grew out of. Thanks are also due to Peter Arkadiev, Patrik Bye, S. J. Hannahs, Sergei Knyazev, Bruce Morén-Duolljá, Elena Parina, Dmitri Sichinava, Jochen Trommer, and Valentin Vydrine. Special thanks are due to the editors of the volume. It is needless to mention that all errors and shortcomings are, of course, entirely mine.

Notes

1. That does not mean that such cases are totally uninteresting from an initial consonant mutation perspective: at least in the Celtic languages, mutations are ultimately the outcome of the grammaticalization of essentially the same phenomenon.
2. Though some cases of such interaction do exist: for example, in Indonesian "nasal substitution" takes place between a prefix and a root but not between two prefixes (Blust 2004: 123).
3. With respect to languages with mutation, Nivkh is a particularly instructive example: what Shiraishi (2006) interprets as syntactic domains, Mattissen (2003) at length argues to be "words".

4. Though Standard Italian as spoken in the north of Italy lacks this feature (De Mauro 1976).
5. Here I discount morpheme-initial phenomena that are never or only rarely word-initial, such as the Athabaskan *d*-effect.
6. Some other languages with initial consonant mutation-like phenomena where the (phonetically voiced) rhotic patterns with the voiced fricatives are Fula, Basari, Konyagi, Venda, and Chemehuevi, while in Tswana, Seereer, and Bedik they pattern with *voiceless* fricatives.
7. Other languages where coronal spirants pattern with palatals or postalveolars in initial consonant mutation-like phenomena are Fula (and other Atlantic languages), Javanese (and other Austronesian languages), and Nias.
8. Interestingly, the problem of featural specification of [r] with special reference to Nivkh was already raised by Trubetzkoy (1939).
9. Of course, it is entirely possible that closer (instrumental) analysis of Fula phonetics may reveal a difference between the realization of the ambiguous segments belonging to different mutation series, in which case the objection falls. I am not, however, aware of any such study.
10. The other Celtic languages: Irish (Christian Brothers 1999), Welsh (Ball and Müller 1992), Cornish (Lewis 1946) and Breton (Kervella 1946; Stephens 1993) present even more involved chain-shifting systems in their initial consonant mutation.
11. Other languages (without initial consonant mutation) where [ð] functions as a sonorant are Woods Cree (Ballard and Starks 2005) and Osage (Quintero 2004).
12. Interestingly, it seems that the existence of *raddoppiamento* triggered by surface phonological factors (namely stress) implies the presence of lexically triggered *raddoppiamento*, but not vice versa (cf. Loporcaro 1996).
13. In a less restrictive theory of (morpho)phonology, even the requirement for featural coherence is not necessary: the mutation can be triggered by fully abstract diacritic elements (sometimes called "morphophonemes"). The foremost exponent of this idea is Hamp (1951) with respect to Celtic; Pym and Larrimore (1979); Evans (1998) take a similar approach to Iwaidja.
14. Not all varieties of Fula demonstrate initial consonant mutation in verbs.
15. Except those dialects where a uniform plural suffix is being developed (Koval' 2000).
16. The distribution of preposed vs. enclitic subjects is dependent on tense-aspect-modality features and information structure.
17. A note on terminology: I follow an Africanist tradition in distinguishing *gender* as a pair of *classes*, each of which is interpreted as singular or plural within each particular gender. It is classes that are morphologically marked.
18. Fula is also problematic for realizational approaches where morphophonological phenomena arise from the requirement to make morphologically dissimilar forms phonologically dissimilar (cf. Kurisu 2001): since the relevant morphological categories are spelled out by explicit morphemes, lack of mutation fails to violate the realization requirement (Wolf (2007) also points this out for suffix-induced morphophonology in Nilotic languages).
19. This account was fleshed out in more detail in Iosad (2008).
20. I know of two possible counterexamples. One is Terena (Akinlabi 1996), where [nasal] spreads rightwards to both consonants and vowels until a stop is encountered; if the stop

is word-initial, the process is reminiscent of initial consonant mutation. However, such spreading clearly does not target word-initial consonants specifically and is thus not an initial consonant mutation by our definition. Another is from Jaxaaw Seereer-Siin, where continuancy can occasionally spread not just to the initial segment, but also to the second consonant of a root: *sox* 'to squeeze', *coq* 'a squeeze' (Fal 1980: 114). None of these cases is like Chaha, however, in that the "word-internal" alternation is always cued at the triggering (left) edge, whereas in Chaha the alternation site can be fully disassociated from the triggering edge.

References

Akinlabi, Akinbiyi
1996 Featural affixation. *Journal of Linguistics* 32: 239—289.

Anderson, Stephen R.
1976 On the description of consonant gradation in Fula. *Studies in African linguistics* 7.1: 93–136.

Anderson, Stephen R.
2005 *Aspects of the theory of clitics*. Oxford: Oxford University Press.

Armagost, James L.
1989 Comanche consonant mutation: initial association or feature spread? *Kansas Working Papers in Linguistics* 14.2: 1–19.

Arnott, David W.
1970 *The nominal and verbal systems of Fula*. Oxford: Clarendon.

Ball, Martin J., and James Fife (eds.)
1993 *The Celtic languages*. London / New York: Routledge.

Ball, Martin J., and Nicole Müller
1992 *Mutation in Welsh*. London / New York: Routledge.

Ballard, Elaine, and Donna Starks
2005 Liquids: laterals, rhotics and much more? In *Proceedings of the 2004 Conference of the Australian Linguistics Society*, Ilana Mushin (ed.). http://hdl.handle.net/2123/102

Beckman, Jill N.
1999 *Positional faithfulness*. Doctoral dissertation, University of Massachusetts, Amherst.

Blust, Robert R
2004 Austronesian nasal substitution: a survey. *Oceanic Linguistics* 43.1: 73–148.

Booij, Geert
2002 The balance between storage and computation in phonology. In *Storage and computation in the language faculty*, Sieb Nooteboom, Fred Weerman and Frank Wijnen (eds.), 115–138. Dordrecht: Kluwer.

Borgstrøm, Carl Hjalmar
1940 *The dialects of the Outer Hebrides*, vol. 1, A linguistic survey of the Gaelic dialects of Scotland. Oslo: Norwegian Universities Press. Norsk Tidsskrift for Sprogvidenskap, suppl. bind I.

Borsley, Robert D., and Maggie Tallerman
1996 Phrases and soft mutation in Welsh. *Journal of Celtic Linguistics* 5: 1–49.
Borsley, Robert D., Maggie Tallerman, and David Willis
2007 *The syntax of Welsh*. Oxford: Oxford University Press.
Breedveld, Johanna Odilia
1995 *Form and meaning in Fulfulde: A morphophonological study of Maasinankoore*. Leiden: Research School CNWS.
Brown, Lea
2001 *A grammar of Nias Selatan*. Doctoral dissertation, University of Sydney.
Brown, Lea
2005 Nias. In *The Austronesian languages of Asia and Madagascar*, K. Alexander Adelaar and Nikolaus P. Himmelmann (eds.), 562–587. London / New York: Routledge.
Bullock, Barbara
2001 Consonant gemination in Neapolitan. In *Phonological theory and the dialects of Italy*, Lori Repetti (ed.), 45–58. Amsterdam: Benjamins.
Bye, Patrik and Peter Svenonius
2009 Extended exponence and non-concatenative morphology. Manuscript, University of Tromsø.
Carstairs-McCarthy, Andrew
1987 *Allomorphy in inflexion*. London: Croom Helm.
Christian Brothers
1999 *New Irish grammar by the Christian Brothers*. Dublin: CJ Fallon.
Churma, Donald G.
1988 Consonant gradation in Fula suffixes: the ugly truth. *Studies in African linguistics* 19.1. 35–74.
Conteh, Patrick, Elizabeth Cowper, and Keren Rice
1985 The environment for consonant mutation in Mende. In *Current issues in African linguistics*, Gerrit Dimmendaal (ed.), vol. 3, 107–116. Dordrecht: Foris.
Cowper, Elizabeth and Keren Rice
1987 Are morphosyntactic rules necessary? *Phonology Yearbook* 4: 185–194.
De Mauro, Tullio
1976 *Storia linguistica dell'Italia unita*. Bari: Laterza.
D'Imperio, Mariapaola and Sam Rosenthall
1999 Phonetics and phonology of main stress in Italian. *Phonology* 16: 1–28.
Doke, Clement Martyn
1954 *The Southern Bantu languages*. London / New York: Oxford University Press.

Dresher, Bezalel Elan, Glyne Piggott, and Keren Rice
1994 Contrast in phonology: overview. *Toronto Working Papers in Linguistics* 14: iii–xvii. Mande. In *Current approaches to African linguistics*, Gerrit J. Dimmendaal (ed.), vol. 3, 149–170. Dordrecht: Foris.
Evans, Nicholas
1998 Iwaidja mutation and its origins. In *Case, typology and grammar*, Anna Siewerska and Jae Jung Song (eds.), 115—151. Amsterdam: Benjamins.

Ewen, Colin J.
1982 The phonological representation of the Welsh mutations. In *Language form and linguistic variation*, J. Anderson (ed.), 75–95. Amsterdam: Benjamins.

Fal, Arame
1980 *Les nominaux en Seereer-Siin: parler de Jaxaaw*. Dakar: Les nouvelles éditions africaines.

Ferry, Marie-Paule
1968 L'alternance consonantique en bedik. *Journal of West African Languages* 5.2: 91–96.

Giannelli, Luciano and Thomas D. Cravens
1996 Consonantal weakening. In *The dialects of Italy* Martin Maiden and Mair Parry (eds.), 32–40. London / New York: Routledge.

Green, Anthony Dubach
2006 The independence of phonology and morphology: The Celtic mutations. *Lingua* 116 (11): 1946—1985.

Green, Anthony Dubach
2007 *Phonology limited*. Potsdam: Universitätsverlag.

Gruzdeva, Ekaterina
1998 *Nivkh*. München: LINCOM Europa.

Hamp, Eric P
1951 Morphophonemes of the Keltic mutations. *Language* 27.3: 230–247.

Hayes, Bruce
1990 Precompiled phrasal phonology. In *The phonology-syntax connection*, Sharon Inkelas and Draga Zec (eds.), 85–108. Chicago: University of Chicago Press.

Hyman, Larry M.
1973 Notes on the history of Southwestern Mande. *Studies in African linguistics* 4: 183–196.

Iosad, Pavel
2008 Initial consonant mutation and information flow in Mende. Presentation at Old World Conference in Phonology 5, Université de Toulouse–Le Mirail.

Kastenholz, Raimund
1996 *Sprachgeschichte im West-Mande*. Köln: Köppe.

Kervella, Frañsez
1946 *Yezhadur bras ar brezhoneg* [A Comprehensive Grammar of Breton] La Baule: Skridou Breizh.

Kibre, Nicholas J.
1997 *A model of mutation in Welsh*. Bloomington: Indiana University Linguistics Club.

Koval', Antonina Ivanovna
1997 Imennyje kategorii v pular-ful'ful'de. [Nominal categories in Pular-Fulfulde] In *Osnovy afrikanskogo jazykoznanija: Imennyje kategorii* [Introduction to African linguistics: Nominal categories], V. A. Vinogradov (ed.), 92–220. Moscow: Aspekt Press.

Koval', Antonina Ivanovna
2000 Morfemika pular-ful'ful'de [Pular-Fulfulde morphemics]. In *Osnovy afrikanskogo jazykoznanija: Morfemika. Morfonologija* [Introduction to African linguistics. Morphemics. Morphonology], V. A. Vinogradov (ed.), 103–254. Moscow: Vostočnaja literatura.

Kurisu, Kazutaka
2001 *The phonology of morpheme realization*. Doctoral dissertation, University of California, Santa Cruz.

Ladefoged, Peter, Jenny Ladefoged, Alice Turk, Kevin Hind and St. John Skilton
1998 Phonetic structures of Scottish Gaelic. *Journal of the International Phonetic Association* 28: 1–41.

Lewis, Henry
1946 *Llawlyfr Cernyweg Canol* [A Handbook of Middle Cornish] Caerdydd: Gwasg Prifysgol Cymru.

Lieber, Rochelle
1987 *An integrated theory of autosegmental processes*. Albany: State University of New York Press.

Loporcaro, Michele
1996 Lengthening and *raddoppiamento*. In *The dialects of Italy*, Martin Maiden and Mair Parry (eds.), 41–51. London / New York: Routledge.

Loporcaro, Michele
1997 *L'origine del raddoppiamento fonosintattico: saggio di fonologia diacronica romanza*. Torino: Rosenberg & Sellier.

Marslen-Wilson, William and Pienie Zwitserlood
1989 Accessing spoken words: the importance of word onsets. *Journal of Experimental Psychology, Human Perception and Performance* 15.3: 576–585.

Mattissen, Johanna
2003 *Dependent-head synthesis in Nivkh: A contribution to a typology of polysynthesis*. Amsterdam: Benjamins.

McLaughlin, Fiona
1994 Consonant mutation in Seereer-Siin. *Studies in African linguistics* 23: 279–313.

McLaughlin, Fiona
2000 Consonant mutation and reduplication in Seereer-Siin. *Phonology* 17: 1–30.

Morgan, Thomas John
1952 *Y treigladau a'u cystrawen* [The mutations and their syntax] Caerdydd: Gwasg Prifysgol Cymru.

Ní Chiosáin, Máire
1991 *Topics in the phonology of Irish*. Doctoral dissertation, University of Massachusetts, Amherst.

Ó Dochartaigh, Cathair
1978 Lenition and dependency phonology. *Éigse* 17: 457–494.

Ó Siadhail, Mícheál and Arndt Wigger
1975 *Córas fuaimeanna na Gaeilge* [The sound pattern of Irish] Baile Átha Cliath: Institiúid Ard-Léinn Bhaile Átha Cliath.

Oftedal, Magne
1985 *Lenition in Celtic and in Insular Spanish*. Oslo: Universitetsforlaget.
Okell, John
1969 *A reference grammar of colloquial Burmese*. London: Oxford University Press.
Paradis, Carole
1992 *Lexical phonology and morphology: the nominal classes in Fula*. New York / London: Garland.
Picanço, Gessiane
2005 *Munduruká: Phonetics, phonology, synchrony, diachrony*. Doctoral dissertation, University of British Columbia.
Press, Margaret
1980 *Chemehuevi: a grammar and lexicon*. Berkeley / Los Angeles: University of California Press.
Pym, Noreen and Bonnie Larrimore
1979 *Papers on Iwaidja phonology and grammar*, A-2. Darwin: Summer Institute of Linguistics, Australian Aborigines Branch.
Quintero, Carolyn
2004 *Osage grammar*. Lincoln / London: University of Nebraska Press.
Rice, Keren and Elizabeth Cowper
1984 Consonant mutation and autosegmental morphology. In *Proceedings of CLS 20*, Joseph Drogo, Veena Mishra and David Testen (eds.), 309–320. Chicago: Chicago University Press.
Roberts, Ian
2005 *Principles and parameters in a VSO language: A case study in Welsh*. Oxford: Oxford University Press.
Rodrigues, Aryon D
2006 Macro-Jê. In *The Amazonian languages*, R. M. W. Dixon and Alexandra Aikhenvald (eds.), 165–206. Cambridge: Cambridge University Press.
Rogers, Henry
1972 The initial mutations in Modern Scots Gaelic. *Studia Celtica* 7: 63–85.
Salanova, Andrés Pablo
2004 Initial truncation in Mẽbengokre. Manuscript. Massachusetts Institute of Technology, Boston.
Sapir, Edward
1930 The Southern Paiute language (part I): Southern Paiute, a Shoshonean language. *Proceedings of the American Academy of Arts and Sciences* 65: 1–296.
Seidl, Amanda
2001 *Minimal indirect reference: A theory of the syntax-phonology interface*. London / New York: Routledge.
Shiraishi, Hidetoshi
2004 Base-identity and the noun-verb asymmetry in Nivkh. In *On the boundaries between phonetics and phonology*, Dicky Gilbers, Maartje Schreuder and Nienke Knevel (eds.), 159–182. Groningen: University of Groningen.
Shiraishi, Hidetoshi
2006 *Topics in Nivkh phonology*. Doctoral dissertation, University of Groningen.

Smith, Jennifer L.
2002 *Phonological augmentation in prominent positions*. Doctoral Dissertation, University of Massachusetts, Amherst.

Stephens, Janig
1993 Breton. In *The Celtic languages*, Martin J. Ball and James Fife (eds.), 349–409. London / New York: Routledge.

Stewart, Thomas W.
2004 *Mutation as morphology: Bases, stems and shapes in Scottish Gaelic*. Doctoral dissertation, Ohio State University.

Storch, Anne
1995 Die Anlautpermutation in den westatlantischen Sprachen. *Frankfurter afrikanistische Blätter* 2.

Stump, Gregory T.
1987 On incomplete mutations in Breton. In *A Festschrift for Ilse Lehiste*, Brian Joseph and Arnold Zwicky (eds.), 1–10. (= Ohio State University Working Papers in Linguistics; 35).

Stump, Gregory T.
1988 Nonlocal spirantization in Breton. *Journal of Linguistics* 24: 457–481.

Stump, Gregory T.
1995 Stem formation, stem indexing and stem choice. Manuscript. Lexington: University of Kentucky.

Swingle, Kari
1993 The Irish and other mutations. In *Proceedings of West Coast Conference on Formal Linguistics 11*, Jonathan Mead (ed.), 451–466. Stanford: Center for the Study of Language and Information.

Tallerman, Maggie
1999 Welsh soft mutation and marked word order. In *Functionalism and formalism in linguistics*, vol. 2, Michael Darnell, Edith Moravcsik, Frederick Newmeyer, Michael Noonan and Kathleen Wheatley (eds.), 277–294. Amsterdam: Benjamins.

Tateishi, Koichi
1990 A morphological analysis of Mende consonant mutation. *Current approaches to African linguistics* 7: 99–110.

Ternes, Elmar
1990 Initial mutations in Celtic and West African languages: synchrony and diachrony. *Afrika und Übersee* 73: 3–17.

Thomas, Peter Wynn
1996 *Gramadeg y Gymraeg* [A Welsh Grammar]. Caerdydd: Gwasg Prifysgol Cymru.

Tokizaki, Hisao
2005 Prosody and phrase structure without labels. *English Linguistics* 22.2: 380–405.

Trubetzkoy, Nikolai Sergejevič
1939 *Grundzüge der Phonologie*, vol. 7, Travaux du Cercle linguistique de Prague. Prague: Le cercle linguistique de Prague.

Vydrine, Valentin
2004 K rekonstrukcii fonologičeskogo tipa i imennoy morfologii pramande. [Towards a reconstruction of the phonological type and nominal morphology of Proto-Mande] Manuscript, Russian Academy of Sciences, Institute of Linguistic Studies, St. Petersburg.

Wolf, Matthew
2007 For an autosegmental theory of mutation. In *University of Massachusetts Occasional Papers in Linguistics 32: Papers in Optimality Theory III*, Michael O'Keefe, Ehren Reilly and Adam Werle (eds.), 315–404. Amherst: Graduate Linguistics Student Association.

Wonderly, William L.
1951 Zoque II: phonemes and morphophonemes. *International Journal of American Linguistics* 17.2: 105–123.

Zwicky, Arnold
1984 Welsh soft mutation and the case of object NPs. In *Proceedings of CLS 20*, Joseph Drogo, Veena Mishra and David Testen (eds.), 387–402. Chicago: Chicago University Press.

"Quirky" case: rare phenomena in case-marking and their implications for a theory of typological distributions

Andrej Malchukov

1 Introduction

The present paper provides an overview of rare phenomena in case-marking, in particular focusing on case marking patterns which are unusual in terms of distribution or function. It will also provide an explanation for rare (resp. common) patterns in terms of functional and diachronic factors in interaction. This paper will *not* address those patterns of case marking which are unusual in phonological realization (such as "tonal case", as found in a number of African languages). Some of these phenomena have been addressed in Dryer (2005), while more discussion of functionally unusual ("exotic") cases can be found in Malchukov (2009).

The paper is organized as follows. In Section 2, I address case patterns with unusual distribution, such as double case, distributed case, case attaching to a "wrong" host. In Section 3, I discuss those cases, which are functionally unusual: oblique cases with diffuse functions, "pragmatic" cases, as well as cases with unusual syntactic distribution ("quirky cases"). Section 4 discusses form-function mismatches in case marking, when a single case performs a double function (designative case in Tungusic), or when a function pertains to another domain (cf. modal and verbal cases in Kayardild). Section 5 addresses cross-linguistically rare patterns of case-marking, including marked nominatives, ergative case restricted to pronouns, double oblique marking of core arguments.

Section 7 recapitulates the discussion in the previous sections, showing how rare patterns arise, but also implicating which factors are responsible for their rare distribution. I will argue that rare patterns arise when a functional constraint conflicts with a grammaticalization path, or when the emergence of a particular pattern is dependent on several preconditions, or else when this pattern represents an incomplete grammaticalization cycle. The main results of the study are then briefly summarized in Section 8.

2 Distribution unusual

2.1 Double cases

Probably the best studied case of deviant distribution of case markers is the phenomenon of *double case* marking, or *Suffixaufnahme* (Plank (ed.) 1995). The most wide-spread pattern of the *double case* involves the genitive signaling the dependency within the noun phrase in combination with an external case signaling agreement with the head (see example (1) from Kayardild below). Given that the functions of the two cases are different,[1] double case marking is functionally well motivated and "logical" (Blake 2001: 103). Yet, most languages impose a restriction on double case marking as case markers tend to "paradigmaticize' as they get more grammaticalized (Lehmann 1995), and most languages take recourse to alternative strategies to avoid case doubling (Moravcsik 1995).

One possible explanation for the fact that the double marking is not as pervasive as expected is that it is redundant, as genitive marking alone (in combination with external marking on the head[2]) perfectly suffices to reconstruct syntactic roles of the elements of the noun phrase.

One of the most spectacular cases of multiple case marking comes from Australian languages such as Kayardild (Evans 1995). Consider the following example demonstrating case stacking in Kayardild:

(1) Kayardild (Evans 1995: 115)

Maku yalawu-yarra yakuri-na dangka-karra-nguni-na mijil-nguni-na
woman catch-PST fish-mABL man-GEN-INSTR-mABL net-INSTR-mABL
'The woman caught some fish with the man's net.'

In Kayardild, a noun can appear with up to four cases. Dench and Evans (1988) and Evans (1995) explain this pattern by a rule of "complete concord" which involves consistent percolation of case from head to dependent. This principle will account for the reason why the possessor takes, apart from the genitive, the instrumental case to agree with the possessor; the occurrence of the external "modal ablative" in this example, however, needs a separate explanation (see Section 4.2).

A pattern reminiscent of double marking[3] is *case layering*, as familiar from Indo-Aryan languages where case markers of postpositional origin attach to the oblique form of the noun; cf. Hindi: *bacce=ne* [child.OBL=ERG], *bacce=ko* [child.OBL=ACC]. While case layering is diachronically a result

of the renewal of case systems through the grammaticalization of postpositions, from a synchronic perspective such patterns are better analyzed as either postpositions in combination with the oblique case of a noun (this is the standard analysis of the Hindi forms above), or case proper attaching to the oblique stem (see Kulikov 2006 for further discussion of case layering in Indo-Aryan; and Blake 2001 and Spencer 2009 for a general discussion). As a final illustration, consider triple case marking in the Omotic language Maale (Amha 2001; König 2006), where the accusative functions as a stem formative, and the locative serves as input for formation of the ablative: *maa-o-idda-ppa* [house-ACC-LOC-ABL] 'from the house'.

Multiple case marking may also arise in languages with "templatic" morphology, if case markers are distributed across several slots in the template (see Spencer 1991 for discussion of "layered" vs. "templatic" morphology). Koasati may serve to illustrate the phenomenon of *distributed case* (Malchukov 2009). In this Muskogean language the order of syntactic and spatial cases is different: semantic (locative) cases precede the suffixal determiner ("article"), while syntactic cases (nominative and accusative) follow it (Kimball 1985: 345). This may yield double case-marking, as in the following example:

(2) Koasati (Kimball 1985: 348)

l·iyá:li-fa-kitt-on *(hí:ca-l)*
stepping-LOC-ART-ACC.FOC (see-1SG)
'I see its footprints.' (lit. 'place where it formerly stepped')

The origin of the distributed case pattern becomes evident once one takes into account other functions of the case markers in question. On the one hand, locative cases retain some derivational functions (note that in (2) the locative case acts as a sort of nominalizer). On the other hand, syntactic cases seem to have been recently reanalyzed from the discourse markers: they have a special focus form (as can also be seen in (2)), and are incompatible with other discourse markers (Kimball 1985). Thus the origin of syntactic cases from discourse markers and semantic cases from derivational markers would account for their linear position as well as for their combinability.

2.2 Displaced case in Iraqw

Iraqw provides an example of a deviant distribution of case, which – in violation of iconicity – attaches to the "wrong" noun (Mous 1993). In this

Cushitic language some case-markers are suffixes while others are enclitics which attach to the noun in preverbal position. Strikingly, the noun "is not necessarily the object of the case relation" (Mous 1993: 102). Compare the two following synonymous sentences:

(3) Iraqw (Mous 1993: 246)

a. *inós i hhar-tá hhawat=i hanmiis*
3SG S.3 stick-F.CON man=DIR give
'He is giving a stick to the man.'

b. *inós i hhawatú hhart=i hanmiis*
3SG S.3 man.CON stick=DIR give
'He is giving a stick to the man.'

In (3) the object (theme) is marked by the construct state marker (the usual case for objects when they follow pronominal inflection marker; Mous 1993: 242 and p. c.), while the recipient takes a directional enclitic. In (3b), however, the order of the two objects is reversed and the directional case-marker attaches to the theme instead (lit. 'gave man to the stick'). This kind of case displacement arises from the interaction of different factors, some of which are syntactic (word order flexibility) while others are morphological (cases are clitics) or morphophonological (clitics appear in a dedicated preverbal position and cliticize to the left). Note that none of these factors is exceptional *per se*; yet in combination they "conspire" to produce this unusual pattern.

3 Case function unusual

Above we have discussed those cases that are deviant in terms of morphological distribution. The cases to be considered in this section may be called functionally unusual in different ways. Note that these peculiarities are substantial and cannot be reduced to matters of terminology (on case terminology see Haspelmath 2009). More discussion of functionally unusual and "exotic" cases can be found in Blake (2001), Iggesen (2005), and Malchukov (2009).

3.1 "Old" cases with diffuse functions and syncretic cases

A case may be unusual if its function, rather than being too specific or restricted, is too general or diffuse to be captured in syntactic or semantic terms.

As expected, such cases are more typical of minimal case systems, where the general function of an "oblique" case is often difficult to establish (see Arkadiev 2009), yet they may also be found in languages with large or mid-sized case systems. As an example, consider the functions of the "oblique" case in Kayardild (Evans 1995: 149). Its general or basic meaning is unclear: it performs some dative-like functions (purpose, object of middle verbs), but otherwise its functions seem semantically disparate. The explanation for this disparity is historical (Evans 1995: 148–149). The "oblique" case is the old dative case (still preserved in the genetically related Yukulta), which has been replaced in its central functions (beneficiary, addressee) by emergent ("verbal") cases (see Section 4.2). Thus we are dealing here with a residual function of a case which does not conform to any recognizable semantic configuration. Another well-known source for semantically disparate functions is the (phonologically conditioned) case syncretism cf. Baerman (2009). For example, in Ancient Greek the same "dative" case is used in the instrumental function (when combined with inanimate nouns), and with the dative function (when combined with animate nouns). This polyfunctionality, hardly attested elsewhere, has been conditioned by a merger of two originally distinct cases in one form (Luraghi 2003: 51). Since dative-instrumental syncretism is conditioned by phonological rather than semantic factors its rareness is unsurprising. By contrast, those cases where polyfunctionality reflects a semantic similarity, are cross-linguistically wide-spread (cf., e. g., the frequent pattern of dative-allative polysemy). And conversely, on iconicity assumptions, as implemented in the semantic map approach (Haspelmath 2003), recurrent similarity of form reveals similarity of meaning.

3.2 "Pragmatic cases"

Another reason why it may be difficult to define the function of a case marker in syntactic or semantic terms is that its basic function belongs rather to the domain of pragmatics. As an example of a case which defies assimilation to a more familiar label, consider the "presentative" case in Samoan (Mosel and Hovdhaugen 1992), which is listed among 15 prepositional case-markers:

(4) Samoan (Mosel and Hovdhaugen 1992: 500, 143)

'O le ulugāli'ı ma l=a lā fānau
PRES ART couple and ART=POS 3DU children
'There was a couple and their children,…'

(5) *‘O le maile sā fasi e le teine*
PRES ART dog PAST hit ERG ART girl
‘The dog was hit by the girl.’

It is difficult to find a common denominator for the presentative case in semantic or syntactic terms. It is used to introduce a clause (see (4)), with nominal predicates, as well as fronted noun phrases (see (5)); (Mosel and Hovdhaugen 1992: 143, 772). In the latter use it is not restricted to any particular syntactic function: It can be an object (as in (5)), or other argument. As also suggested by the case label, the function of the presentative case is basically pragmatic: it introduces rhematic constituents, including contrastive / new topics. Yet, this marker is paradigmatic with other case-markers and therefore should be considered as a case-marker on language-particular grounds.

The case of Samoan is relatively straightforward, as the presentative marker shows few syntactic restrictions in distribution. In other languages, however, case-markers show a conflation of pragmatic and syntactic information. In Japanese the “nominative” marker *ga* is used with rhematic subjects; the predicate-focus case in Yukaghir is restricted to intransitive subjects and direct objects (Maslova 2003); the “subordinative case” suffix *-a* in Nama marks objects and (nominal) predicates (Hagman 1977: 114); the “background case” in Iraqw marks oblique constituents when topicalized / fronted (Mous 1993: 108); the name of the “topical non-subject” case in Tariana speaks for itself (Aikhenvald 2003).

Of course it is not unusual for syntactic cases to show a correlation with discourse categories, as a well-known correlation between nominative subjects and topical / given information can illustrate (Blake 2001: 133). Yet, with respect to “pragmatic cases” the situation is rather the reverse: here the pragmatic function seems to be basic, while the syntactic functions are derivative. For example, in Tukang Besi, the case-markers *te* and *na* cannot be uniquely associated with any syntactic function irrespective of word order, agreement, and information structure (Donohue 1999).

The case of Tukang Besi, as described by Donohue (1999; 2009), is particularly instructive, inasmuch as it shows a tight integration of pragmatic and syntactic information within one system, which leads to a very intricate system of case assignment. Thus the “core” case in *te* is interpreted as P if postverbal and does not control agreement (see (6)); if the verb is bipersonal

it is interpreted as A (see (6b)), unless it is fronted, in which case it is again interpreted as P (see (6c)).

(6) Tukang Besi (Donohue 2009)

a. *Te wowine no-'ita te kadadi.*
TE woman she:saw TE bird
'The woman watched birds.'

b. *No-'ita='e te wowine na kadadi.*
she:saw:them TE woman NA bird
'The woman watched the birds.'

c. *Te kadadi no-'ita='e te wowine.*
TE bird she:saw:them TE woman
'The woman watched the birds.'

Clearly, it is futile to try to interpret these constructions in terms of semantic roles; but it is also not straightforward to assign them a syntactic function, since such a characterization would have to make reference to word order (and agreement). The syntactic characterization offered by Donohue (2009: 776) is as follows:

te: if preverbal, *te* marks the subject;
if postverbal, *te* marks a core non-subject ('object').

na: necessarily postverbal, *na* marks the subject.

An alternative interpretation, which is in terms of discourse functions, is not totally straightforward either (Donohue 2009: 776):

te: if preverbal, *te* marks identificational focus;
if postverbal, *te* marks a non-given term.

na: necessarily postverbal, *na* marks a (non-focussed) given term.

It seems that part of the problem is that the reanalysis of pragmatic to syntactic markers is still under way. Yet, the pragmatic function is not totally clear either, which suggests intermediate stages of reanalysis. This seems to be true of other languages as well, insofar as the authors discuss the diachronic dimension. Thus, for Yukaghir it has been suggested that P-markers developed from Focus-markers but retain some residual discourse functions (Maslova 2009). The same might hold for the "subordinative case" in Nama which derives from a copula and seems to have had discourse (rhematic) function (Tom Güldemann, p. c.). Similarly, in languages with the marked nominative this case frequently originates from a definiteness / topicality marker (König 2006).

3.3 Idiosyncratic ("quirky") case: polysemy and pattern inheritance

The term "quirky case", which appears in the title, is commonly used to refer to idiosyncratic uses of cases that cannot be explained in synchronic terms. Note, however, that in many, perhaps most, cases diachronic explanation is readily available for such idiosyncratic uses. For example, in Hindi / Urdu the verb *la-* 'bring' exceptionally takes a nominative subject (in the past tense), aligning itself with intransitive rather than transitive verbs. An explanation to this alignment is historical: the verb *la-* derives from contraction of *le-* 'take' with the intransitive *a-* 'come' (Butt 2006: 125). In Ingush, the verb *ladieG-* 'listen' exceptionally takes the ergative-locative pattern, lacking the absolutive argument which is otherwise obligatory in this language. This is due to the fact that historically *ladieG-* originates from the compound *la* + *dieG* 'ear + put' (Nichols 1994: 119). A similar point can be made with respect to subject-experiencer constructions in European languages. As noted by Haspelmath (2001), European languages are typologically unusual in showing a clear preference for the subject-experiencer verbs (such as English *like*), rather than object experiencer verbs (such as German *gefallen* 'like; please'). Haspelmath attributes this predisposition to the fact that many of the subject experiencer verbs historically derive from the verbs denoting a physical action (e. g., *worry* is derived from 'strangle; seize by the throat'; Haspelmath 2001: 79).

In the cases discussed above, the explanation to the current pattern lies in the diachronic domain, in other cases, however, there are vestiges of this development observed on the synchronic level as well, in the form of case polysemy. Here then we are dealing with patterns of case unification rather than case inheritance (Malchukov 2005). While predisposition for subject experiencer verbs in European languages has been explained above by pattern inheritance, pattern unification may be responsible for the predisposition of certain experiencer verbs to the object experiencer construction. Thus, the fact that psych-verbs from Levin's (1993) 'amuse' class take an object experiencer may be due to the fact that many of them (including *amuse* itself) also allow an agentive interpretation (Grimshaw 1990: 23), while some others (e. g., *depress*, *strike*) can also refer to a physical (agentive) action (Levin 1993: 191). Another example from the same domain concerns predisposition of different types of perception predicates (verbs of active vs. inactive perception) for the inverse construction. It has been observed (Primus 1999; Malchukov 2005) that among perception verbs the inverse (DAT-NOM / ABS) pattern is char-

acteristic for 'see'-verbs, rather than 'look'-verbs. This is hardly surprising, as in the case of active perception verbs, the first argument is more agentive. Yet, in some languages, active perception verbs like 'look' also allow this pattern. On closer inspection, it turns out that these perception verbs are polysemous in these languages. Thus, in Lezgian, *akwa-* 'see' can also be used in the 'look' sense. Importantly, *akwa-* 'see' which takes the inverse DAT-ABS pattern, retains this pattern when used as 'look':

(7) Lezgian (Haspelmath 1993: 283)

Gila kwe-z [za wuč-da-t'a] aku!
now you-DAT [I.ERG do.what-FUT-COND] see.IMPFV
'Now look (lit. see) what I am going to do!'

Thus, a polysemous verb can share the same case pattern even if different meanings represent different verb types. This is evidently due to analogical pressures which may require '(case) pattern unification' on the part of a polysemous verb. On the other hand, 'pattern inheritance' that was implicated as being responsible for the rise of "quirky case" is due to the fact that in grammaticalization processes morphosyntactic changes (e. g., in subcategorization properties) may lag behind a functional reanalysis. (See Malchukov 2005 for more discussion of pattern inheritance and pattern unification in case-marking.)

4 Function-form mismatches

4.1 Designative case in Tungusic

An unusual case of a function-form mismatch is illustrated by the use of "designative" case in Tungusic languages.[4] In Tungusic languages, designative case appears on the direct object in combination with possessive markers, as illustrated by the following example from Even:

(8) Even (Malchukov 1995, and f. n.)

*Hin turki-**ga**-s emu-re-m*
your sledge-**DES**-2SG bring-NONFUT-1SG
'I brought the sledge for you.'

An unusual feature of this construction is that the possessor (cross-referenced by possessive suffixes on the object) is invariably interpreted as beneficiary.

This can be seen through comparing this construction to a parallel construction involving accusative case.

(9) Even (Malchukov 1995, and f. n.)

*Hin turki-**wu**-s emu-re-m*
your sledge-**ACC**-2SG bring-NONFUT-1SG
'I brought your sledge.'

The latter construction does not impose any particular interpretation on the possessor on the clausal level: in (9) it is likely interpreted as source (i. e. 'I brought the sledge from you') but also allows for other interpretations. In the designative construction, however, the possessor is always interpreted as a beneficiary. Therefore it can be argued that the designative case performs a double function, assigning two different functions to two different noun phrases: marking its host as object and simultaneously assigning beneficiary role to its possessor (Malchukov 1995).

The designative case in *-ga-* is also found in other Tungusic languages, such as Nanai (East Tungusic),[5] but its origin remains unclear (cf. Benzing 1955). One tempting explanation informed by typological data is to relate the designative case in *-ga-* to the homophonous verb *ga-* 'take': it is well known that 'take'-verbs constitute a common origin of both instrumental and object markers (Heine and Kuteva 2002: 288–289). On this account the designative construction could have originated from the fusion of a nonfinite form of *ga-* with a subordinate object. In particular, the designative construction like the one in (9) could have developed from the original construction (10) where the verb *ga-* takes the homophonous purposive converb ("supine") marker *-ga-* (as preserved in Nanai), with a subsequent simplification through haplology:

(10) Even

* *[Hin turki(w) ga-ga-s] emu-re-m*
your sledge(ACC) take-PURP.CONV-2SG bring-NONFUT-1SG
'I brought the sledge for you.'

This account can readily explain why the possessor (the erstwhile subordinate subject) is interpreted as a beneficiary: the source construction (10) would mean more literally 'I brought (it) for your sledge-taking'. Further, this scenario explains why the designative case obligatorily combines with possessive suffixes, as non-finite verbs take a possessive style agreement. If

this account stands, one factor contributing to this unusual pattern is that the designative case derives from a verb, while most other cases in Tungusic are of denominal origin (cf. Novikova 1960). Equally important is that the designative marker still shows vestiges of its verbal origin, which makes this case similar to the verbal case in Kayardild, discussed in the next section (see, in particular, Note 6 on page 160).

4.2 "Quirky" cases in Kayardild

Kayardild (Evans 1995) is celebrated for its unusual case-marking. Apart from the multiple case-marking already mentioned in Section 2.1, a particularly intriguing feature of Kayardild is the existence of "modal cases", which are used to express tense / aspect / mood features. Consider the following examples where the "modal proprietive" case is used to convey future meaning (11a), while the "modal ablative" case is used to convey the past meaning (11b).

(11) Kayardild (Evans 1995: 108)

a. *ngada warra-ja ŋarn-kiring-**ku***
I.NOM go-POT beach-ALL-**mPROP**
'I will go to the beach.'

b. *ngada warra-jarra ŋarn-kiring-**kina***
I.NOM go-PST beach-ALL-**mABL**
'I went to the beach.'

Note that modal cases (such as proprietive and ablative) can be used elsewhere in their normal "relational" function marking arguments of the verb. In the modal function however they appear in addition to (and externally to) relational cases. The diachronic scenario responsible for the rise of modal cases has been described by Dench and Evans (1988) and Evans (1995; cf. Blake 2001: 108). According to this scenario, originally these case markers marked non-finite forms in the complement and adverbial function. In accordance with the general rule of "total concord" operative in Kayardild (see Section 2.1) the cases percolated from the nonfinite verb heading the subordinate clause to its arguments (except for the subject argument). Subsequently, the erstwhile subordinate clauses became increasingly used as main clauses (a process called "insubordination" by Evans), while case forms on the verb have been fused with the verbal inflection. The net result is that in the course

of this development the cases survived on the dependents but not on the verbal head.

Another exceptional feature, which is restricted to Kayardild and a few other Tangkic languages, is the existence of "verbal cases". Verbal cases behave like normal cases syntactically (in particular, they also percolate to dependents), but additionally perform a verbalizing function. Consider (12) where the recipient is marked by the "verbal dative" case, which further takes the imperative inflection, just as the verb does:

(12) Kayardild (Evans 1995: 336)

wuu-ja wirrin-da ngijin-maru-th
give-IMP money-NOM 1SG-**vDAT**-IMP
'Give me the money!'

Again this feature is best understood in a diachronic perspective (Evans 1995: 182–183). Verbal cases have originated in noun-verb compounds, which would explain their verbal inflection (thus, the verbal dative in example (12) stems from *maru.tha* 'put').[6] Yet in itself verbal origin cannot explain the peculiarity of this pattern: after all verbs constitute a common source for case-markers cross-linguistically (Blake 2001). Evans (1995: 183) notes that two other factors may be responsible for the rise of verbal case, both related to availability of modal cases. On the one hand, there is need to recruit new cases, as the old cases of nominal origin increasingly acquire modal functions. On the other hand, retention of the verbal inflection on nouns may be due to the fact that Kayardild already had a means of expressing tense-aspect-mood features on nouns, by means of the "modal case". These factors are indeed relevant but, in my view, are not sufficient to explain the strange outcome: the fact that case-markers take verbal inflection. Another factor should not be overlooked in this connection: the role of case concord. Indeed if there were no rule of "complete concord" in Kayardild, one would probably never be tempted to identify these markers as cases in the first place, as the main argument for their case status is the adherence to the general rules of case percolation. From a diachronic perspective, another important factor is that verbal cases are "emergent" cases, where the full cycle of reanalysis of verbal forms into cases has not been completed. As is usual for grammaticalization processes, functional reanalysis is more advanced here (verbal cases behave like cases syntactically in terms of concord), while morphological reanalysis lags behind (the deverbal markers still retain verbal inflections).

Thus, both modal and verbal cases are borderline phenomena attested in incomplete cycles of reanalysis of case markers: only in the former case we are dealing with incomplete reanalysis of the erstwhile cases into the verbal inflection, while in the latter case we are dealing with incomplete reanalysis of the erstwhile verbs into case markers.

5 Rare alignment patterns

In this section I address cross-linguistically rare alignment patterns, such as marked nominatives, ergatives restricted to pronouns, and double oblique patterns. It will be argued that rareness of these patterns is accounted for by the fact that they are "dysfunctional", that is, violate one or several functional constraints in the domain of case marking. These constraints can be derived from basic functions of case marking, as established in the functional typological literature (see, e. g., Comrie 1978; Kibrik 1985):

(a) encoding semantic roles (under the indexing approach); and
(b) distinguishing between verbal arguments (under the discriminating approach), as well as complying with economy considerations.

Yet, as will be clear from the discussion below, these patterns are expected to arise under certain diachronic scenarios, even though these patterns will be cross-linguistically disfavored for functional reasons.

5.1 Marked nominatives

It is well known that there is a general tendency for a language to have one unmarked case which would be nominative in an accusative language, and absolutive in an ergative language (Dixon 1979; Tsunoda 1981). The usual explanation for this tendency is the interaction of distinguishability and markedness; in an intransitive sentence marking of a single argument is dispensable and in a transitive it can be restricted to one of the arguments (see, e. g., Comrie 1989; Kibrik 1985). Yet a number of African languages are known to be exceptions to this generalization (König 2006). Marked nominatives are also attested in some Muskogean languages like Koasati (Kimball 1985). In Koasati, nominative and accusative cases which have an overt marking contrast with the unmarked "autonomous" case used for indirect objects and possessors.

(13) Koasati (Kimball 1985: 331)

a. *ifá-k*
dog-NOM

b. *ifá-n*
dog-ACC

c. *ifá-Ø*
dog-AUT

The previous discussion of Koasati in Section 2.1 could already give a clue to an explanation of this unusual pattern. As noted above, nominative (and accusative) cases in Koasati have discourse-pragmatic functions, as is evident from the fact that they are incompatible with other discourse markers. The same explanation seems to hold for marked nominatives in African languages (Cushitic, Berber), where topic / definiteness markers are usual sources of marked nominative (König 2006). Thus, this rare pattern is actually a manifestation of a general grammaticalization path leading from topic to subject (Givón 1984; Shibatani 1991). While grammaticalization can explain why marked nominatives arise in some languages, economy considerations are responsible for the fact that such patterns are relatively rare. Another relevant consideration is that the original topic construction should involve dependent marking rather than head marking, or else this grammaticalization path would give rise to agreement rather than marked nominatives (Givón 1984).

5.2 Pronominal ergatives

Although ergative languages where both nouns and pronouns take ergative case are not uncommon, languages where only pronouns have ergative case are very rare. Indeed they constitute a counterexample to Silverstein's (1976) generalization concerning possible distribution of ergative and accusative patterns. The rationale behind this generalization relies on the notion of markedness, pronouns are most natural (least marked) in the function of transitive subjects, therefore they are expected to be least marked morphologically as well.

Yet, exceptions to this generalization have been reported for a number of languages, in particular, Iranian languages recently discussed by Filimonova (2005; cf. Payne 1980). Thus, in Parachi and Yazgulami ergative ("oblique") case marks the transitive subject (in past tenses), if the subject is pronom-

inal, while nominal subjects remain unmarked. As detailed by Filimonova (2005; cf. Plank 1985), this pattern has a straightforward diachronic explanation. This group of Iranian languages has been undergoing an ergative to accusative shift, and the ergative is retained on pronouns as they are generally more conservative in language change. Ultimately, it can be a frequency effect[7]: pronominal inflection is more conservative than nominal, since pronouns are more frequent than nouns (cf. Haspelmath 2006 on the role of frequency). Yet, although retention of the ergative case on pronouns is expected under this scenario, the pattern is rarely attested, and even if it emerges it seems to be unstable for well understood functional reasons. Indeed, in some dialects of Parachi the ergative form on pronouns has been replaced by the nominative one, while in Yazgulami ergative (oblique) marking spreads to intransitive subject as well, hence it no longer incurs a violation of Silverstein's generalization (Filimonova 2005).

A less known example of pronominal ergatives is provided by certain Mande languages like Guro, where the ergative is restricted to first and second person pronouns (Vydrine 2006). Consider the following example from Guro, which shows an alternation between ergative (basic series) and nominative case on a pronominal subject:

(14) Guro (Vydrine 2006: 53)

maa (∼ãã) bɛɛnɛɛ ɓaalaaá
1SG.ERG 1SG.NOM dog beat.PFV
'I beat a dog.'

Again the rise of this pattern is understandable from a historical perspective: in Mande languages the pronominal ergative derives historically from portmanteau pronouns (Vydrine 2006: 63, fn. 6). Thus the first person ergative pronoun *maa* in (14) historically derives from a contraction of the first person subject with the third person object pronoun (even though synchronically they cannot count as portmanteau pronouns any longer; Vydrine 2006: 54). Note also that in this case the ultimate explanation may be due to frequency. Vydrine (2006: 56) notes that ergative pronouns are used more readily and show fewer restrictions, the higher the status of subject on the animacy hierarchy is and the lower the status of the object. Since the most hierarchically natural scenario with subject pronouns outranking the object pronouns in prominence (that is, a "direct" $1,2 \rightarrow 3$ as opposed to the "inverse" $3 \rightarrow 1,2$ pattern) is expected to be most frequent, it is not surprising that this pattern grammaticalized earliest. Yet this functionally natural develop-

ment[8] yields a violation of a general markedness pattern, as established by Silverstein (1976) and Comrie (1978). This may account for the rarity of this pattern although an additional factor should not be overlooked: this scenario crucially depends on the possibility of a language to develop contracted pronouns in the first place.

6 Double oblique patterns

Double oblique patterns are cross-linguistically rare, but are attested in a number of Iranian languages (Payne 1980; Bossong 1985; Stilo 2004; Arkadjev 2005):

(15) Vafsi (Stilo 2004: 244)

luas-i *kærg-e=s* *bæ-værdæ.*
fox-OBL.SG chicken-OBL.SG=3SG PFV-take.PST
'The fox carried off (the) chicken.'

Why this pattern is rare is obvious: it does not comply with any of the functional motivations behind case-marking which are taken for granted in the literature. It does not distinguish between arguments, nor does it satisfy economy, as both arguments are marked. Also in this case the best approach to this puzzling pattern is diachronic. The double oblique pattern in these languages results from the meaning extensions of the originally polyfunctional dative-genitive case (Kerimova and Rastorgueva 1975; Arkadjev 2005). This polyfunctional case further developed into the marker of a (prominent) object, on the one hand, and to the ergative marker in past tenses, on the other hand. Note that both developments are not unusual. Ergative markers of genitive origin are attested in a number of languages (like Eskimo), and dative to accusative shift constitutes a well-known grammaticalization path (cf. Lehmann 1995). Thus, this pattern is due to a polysemy chain, with individual polysemy patterns well-attested elsewhere; schematically:

If ERG=GEN, GEN=DAT, DAT=ACC,
then ERG=ACC resulting in a A=P≠S pattern.

Thus, while the individual polysemy patterns between adjacent cases are well attested, languages which display the whole 'chain' are rare, as it would yield an ambiguous pattern with both subject and object bearing the same oblique case.

7 Discussion: accounting for typological distributions of case marking patterns

The discussion of rare phenomena in the domain of case-marking (as well as in other domains) raises two general questions: how these patterns developed and why they are rare. In the previous sections I provided an answer to the first question: as shown above, in most cases an unusual pattern can be straightforwardly explained in terms of common diachronic processes. Thus, I fully subscribe to the following statement by Givón,

> "in each instance, a *crazy* synchronic state of the grammar has arisen via diachronic changes that are highly *natural* and, presumably, motivated independently by various communicative factors." (Givón 1979: 235).

Here, I will address the second question building on the material introduced in the previous sections. As is clear from the discussion above, the two general explanations for the typological rareness of a certain case pattern are that this pattern is either 'dysfunctional' or requires a complex diachronic scenario.

The role of functional factors, which may favor or disfavor a diachronic development is particularly obvious in the rise of rare alignment patterns. As is clear from the preceding discussion, these developments arise from grammaticalization channels, well-attested elsewhere, yet the resulting pattern is 'dysfunctional', in that it violates one of the functional constraints:

- Grammaticalization of topics into subjects (see Section 5.1) is a well established grammaticalization path, yet reanalysis of topic markers into marked nominatives is rare, since it is hedged by economy constraints;
- Retention of the ergative case on pronouns (see Section 5.2) is expected under ergative to absolutive shift due to their higher frequency, but pronominal ergatives are rare since they are disfavored by markedness considerations;
- Grammaticalization of genitive-dative to ergative, on the one hand, and to accusative, on the other hand, follows well established grammaticalization channels, but both developments rarely occur simultaneously, since it would yield an ambiguous double oblique pattern (see Section 6).

Similar accounts implicating a counteracting functional factor can be provided for many other cases discussed above. Thus, case displacement (as observed in Iraqw; see Section 2.2) may be well motivated for phonological rea-

sons; yet such patterns are rare, as they violate iconicity (that requires that a noun should be marked for its own role features). For the same reason cases performing a double function (like Designative Case in Tungusic; see Section 4.1) are rare. And double case patterns, as discussed in Section 2.1, are rare since marking a genitive possessor for an external relation is redundant. Thus, in all of these cases we are dealing with a grammaticalization path that is inhibited due to a clash with a functional constraint, such as economy or ambiguity, reducing frequencies of the resultant patterns.

The presented analysis is couched in terms of competing motivations as is usual in typological literature (Croft 2003), but it is also compatible with (functional) optimality-theoretic (OT) approaches that derive the rise of language patterns from interaction of conflicting constraints (Prince and Smolensky 2004; Blutner et al. 2005; de Hoop and Malchukov 2008).[9] Admittedly, it is less conventional to characterize grammaticalization processes in terms of OT-style constraints. In that respect my approach is even closer to Haspelmath's (1999) account which recasts OT constraints as functional constraints arising in the process of diachronic adaptation.

Another reason why a certain pattern is less common is that some development requires a number of preconditions. Put differently, several factors "conspire" to produce a certain pattern. Conspiracies have been extensively discussed in typological literature (e. g., Croft 2003), but not in connection to the frequency of resultant patterns. Explicitly, the connection between rareness of a certain pattern and complexity of a diachronic scenario giving rise to this pattern has been recognized in the recent work by Harris (2006, 2008, this volume). For example, Harris shows that the rareness of the patterns of 'exuberant' (multiple) agreement as found in Tsova-Tush (Daghestanian) can be attributed to the coincidence of several factors. Our findings are consistent with this conclusion. Consider again the spectacular case of case displacement in Iraqw, which crucially depends on the following conditions (see Section 2.2):

– cases are clitics;
– clitics appear in a dedicated (preverbal) position;
– word order of arguments is pragmatically determined.

If one of these conditions were not met, the pattern would not arise. In fact, some case-markers in Iraqw are suffixes rather than clitics, and consequently they are exempt from case displacement. Note that none of these fac-

tors is exceptional *per se*, but *in combination* they yield this unusual pattern. Another instructive example, which relates the rise of a rare pattern to a complex diachronic scenario, is the rise of modal cases in Kayardild discussed in Section 4.2, which involves the following stages:

- case markers appear on subordinate forms in complement function (rise of 'complementizing case');
- the case percolates from the subordinate verb head to its arguments, in accordance to the rule of 'complete concord';
- use of erstwhile subordinates as main clauses ('insubordination');
- case forms on the verb are fused with the verb inflection and reanalyzed as subordinators, while case marking of arguments is maintained (rise of modal case).

Again these factors in isolation are not necessarily rare. In fact, use of case as complementizer is wide-spread cross-linguistically (Koptjevskaja-Tamm 1993; Malchukov 2004). The process of 'insubordination' is also familiar from European languages, where many finite forms are of participial origin. Similarly, the fusion of a case marker into the non-finite marker is frequent: as is well-known, this is a common origin for infinitive markers (Haspelmath 1988). The only truly rare "ingredient" in this process is the pervasiveness of case percolation which requires copying of case from the verbal heads to its arguments (the rule of 'complete concord'). Here indeed it can be concluded that "admixture" of this rare feature in a diachronic process started a "chain reaction", yielding other peculiar factors: as noted above, the emergence of both modal and verbal cases ultimately depends on the rule of complete concord. Yet, crucially, the rise of rare patterns need not involve exceptional factors; actually the features *per se* may be quite common. This has also been noted by Harris (2006, 2008, this volume), who concludes that it is not the rarity of features themselves but rather their coincidence that may lead to rise of a rare pattern.[10]

These two factors – conflicting factors and complex scenarios – can account for most of the cases discussed above. Note that these two explanations do not necessarily exclude each other. Thus, the pattern of case displacement in Iraqw presupposes several conditions, but it also violates Iconicity (or 'Relevance'), which requires a noun to be case-marked for its own features.[11] Arguably when several factors occur in combination, they have a cumulative effect of further reducing the probability of the pattern in question. Yet,

in some other cases another explanation seem to be relevant as well. Thus, rare case patterns frequently represent phenomena resulting from incomplete grammaticalization cycles. The rise of verbal and modal cases in Kayardild provides an instructive example in this respect. Similarly, "quirky" (idiosyncratic) uses (discussed in Section 3.3) are vestiges of specific grammaticalization processes, where morphosyntactic properties lag behind the semantic change. And 'pragmatic cases' (discussed in Section 3.2) represent a development of discourse markers into case markers that has not been completed. Further, at least some of the rare alignment patterns can be viewed as a transitional phenomenon accompanying an alignment shift (e. g., both pronominal ergatives discussed in Section 5.2 and double oblique patterns in Iranian discussed in Section 6 may be seen as intermediate stages of the ergative to accusative shift). Here, I will tentatively suggest that phenomena representing incomplete stages of grammaticalization are less stable diachronically as complex patterns representing a mixture of residual and emerging functions are more difficult to acquire. Thus, the use of "pragmatic cases" in Tukang Besi, determined by an intricate interplay of discourse and syntactic factors (see Section 3.2), may be not easy to acquire and is therefore likely to be unstable. It remains to be seen if this conclusion relating rareness of a particular phenomenon to incomplete grammaticalization / reanalysis can be generalized to other domains. It should be noted that an authoritative proposal to this effect has been recently made by Dahl (2004), who suggested that targets of grammaticalization ("mature structures", in terms of Dahl) are generally more stable diachronically than their sources.[12]

As noted above, these different explanations of typological distributions are not incompatible. From an evolutionary perspective, the frequency of a certain linguistic pattern X reflects the probability of languages developing this structure. On the one hand, this probability depends on whether certain structural preconditions for such a development are met (which in turn, depends on the number of preconditions, but also on the frequencies of the properties in question). On the other hand, the probability will also depend on whether this structure shows properties of a "good design" (defined in terms of satisfaction of functional constraints). This approach can also be extended to explain frequencies of certain grammaticalization paths. Thus, probabilities of certain grammaticalization paths will depend on availability of a source construction, on the one hand (where the source construction itself can be conceived as instantiation of different preconditions), and on communicative 'usefulness' of the grammaticalized concept (target), on the other

hand (cf. Zeevat 2006). In other words, I propose that frequencies of a certain linguistic structure can be derived from two variables: a) its 'generation costs'(determined – *à la* Harris – on the basis of the number of structural features involved); and b) its 'functional fitness'. In general, I am quite optimistic that a general theory of typological distributions (common vs. rare patterns) can be built on these premises, even though many details of this model still have to be worked out.[13]

8 Conclusions

In this paper, I presented an overview of rare phenomena in case-marking, as well as an account how these patterns developed. I further identified several general factors that determine why these factors are rare. First, a rare pattern may result from a *conflict* between a grammaticalization path and a *functional constraint* (see, e. g., the discussion of rare alignment patterns). Second, a pattern may be rare as it requires *co-occurrence of several different conditions*, quite often belonging to different domains (cf., e. g., the conditions under which case displacement in Iraqw arises). Third, functionally deviant cases may result from *incomplete grammaticalization cycles* (e. g., pragmatic cases, not fully reanalyzed, remnant or emergent cases).

It is argued that these different approaches to explaining typological distributions can be integrated in a single evolutionary model, which derives frequencies of a certain construction from its 'generation costs' (estimated in terms of the number of structural features involved), on the one hand, and its 'functional fitness' (estimated in terms of satisfaction of functional constraints), on the other hand. While the details of this model still have to be worked out, it is hoped that this paper makes another step in the direction of creating a general typologically-informed model of language change, which could account for the emergence and distribution of common and rare linguistic structures.

Abbreviations

ACC = accusative; ALL = allative; ART = article; AUT = 'autonomous' case; CON = construct state marker; COND = conditional; CONV = converb; DAT = dative; DES = designative (case); DIR = directional; ERG = ergative; FOC = focus;

F = feminine; GEN = genitive; IMP = imperative; IMPFV = imperfective; INSTR = instrumental; LOC = locative; M = masculine; mABL = modal ablative; mPROP = modal proprietive; NOM = nominative; NONFUT = nonfuture (tense); OBL = oblique; PFV = perfective; POS = possessive; POT = potential; PRES = present (tense); PST = past (tense); vDAT = verbal dative

Acknowledgements

I am grateful to Helen de Hoop, Martin Haspelmath, Nick Evans and Juliette Blevins for the useful discussion of the issues addressed in this paper. The usual disclaimers apply.

Notes

1. Indeed, some authors consider these two cases as representative of two distinct categories (cf. case proper vs. concordial case in Mel'čuk's 1998 framework).
2. Notably, when the head noun is absent, double case-marking is more widespread (Blake 2001: 103; Moravcsik 1995).
3. Still another phenomenon which is reminiscent of double case are 'compound cases', as known from Daghestanian languages (see, e. g., Comrie and Polinsky 1998; Kibrik 2003; Daniel and Ganenkov 2009). In this case also the dominant view is that compound cases involve two different categories: case proper in combination with "orientation" (in terms of Comrie and Polinsky) or "localization" markers (in terms of Kibrik).
4. See Creissels and Daniel (2006) for discussion of similar constructions in Nganasan (a Samoyedic language).
5. The relation of the designative case in *-ga-* with the "indefinite accusative" in *-ja*, as found in Evenki (Nedjalkov 1997) and some other Tungusic languages is not clear. These cases show some functional overlap, but also important differences (the latter is used in a partitive sense and does not necessarily take possessive agreement), which argues against common origin.
6. Interestingly on Evans' account (1995: 178–179), the verbal case has residual government properties inherited from the verb: thus, the verbal dative, as the verb 'put' from which it originates, governs both recipient and theme. This is reminiscent of the double role assignment on the part of the designative case of Even, which as was suggested above may also be due to the verbal origin of the designative case.
7. Filimonova does not implicate frequency in this connection, but notes that "pronouns being deictic words belong to the most archaic part of the lexicon" (Filimonova 2005: 98).
8. Currently, in Guro new series of contracted ergative pronouns with specialized aspectual and / or modal values are being developed (cf. 'contractive ergative imperfective', 'contractive ergative optative', etc; Vydrine 2006).
9. For example, de Hoop and Malchukov (2008) and Malchukov (2008) propose to account for the fact that Differential Object Marking is cross-linguistically more consistent than

Differential Subject Marking in terms of two interacting (functional) constraints. In the domain of Differential Object Marking both constraints converge on the same pattern, while in the domain of Differential Subject Marking they select (prefer) different patterns. Therefore Differential Object Marking is a cross-linguistically more frequent pattern, as compared to Differential Subject Marking, which shows more variation across languages.

10. A similar account has been proposed for the domain of phonology by Blevins (2004). Evolutionary Phonology developed by Blevins aims to account for why certain sound patterns are cross-linguistically frequent and some are rare (in contrast to generative phonology which is primarily concerned with why certain patterns exist and other do not). Blevins showed that sound patterns that are less common either result from less natural phonetic processes or involve more complex scenarios, when "a sound change is inverted, layered over with other changes or diluted with the effects of analogical change" (Blevins 2004: 78).
11. See Malchukov (2006) on the role of the iconic Relevance Principle for constraining correlation between transitivity parameters and transitivity alternations.
12. In computational approaches to language evolution stable states are viewed as linguistic 'attractors'. Such approaches have been more often applied for modeling phonetic changes (see, e. g. Kuhl 1991; Wedel 2004), but Zeevat (2006) develops a similar 'push-and-pull' model for semantic changes in grammaticalization processes.
13. This account bears some similarity to the "Reductionist" approach to (explaining) typological distributions, advocated by Newmeyer (this volume), who distinguishes between options provided by Universal Grammar, on the one hand, and distributional preferences guided by performance factors, on the other hand. My proposal is different insofar as it does not presuppose the notion of Universal Grammar, and, further, relates typological distributions both to (probabilities involved in) generation and evaluation.

References

Aikhenvald, Alexandra
2003 *A grammar of Tariana from northwest Amazonia*. Cambridge: Cambridge University Press.

Amha, Azeb
2001 *A grammar of Maale*. Dordrecht: Foris.

Arkadjev, Peter
2005 Dvuxpadežnye sistemy v indoiranskix jazykax: tipologičeskaja perspektiva. In *Indoiranskoe jazykoznanie i tipologija jazykovyx situacyj*, Bogol'ubov, M. N. (ed.), 74–92. St-Petersburg: Nauka. [Two-term case systems in Iranian languages in a typological perspective].

Arkadiev, Peter
2009 Poor (two-term) case systems: limits of neutralization. In *Handbook of Case*, Andrej Malchukov and Andrew Spencer (eds.), 686–700. Oxford: Oxford University Press.

Baerman, Matthew
2009 Case syncretism. In *Handbook of Case*, Andrej Malchukov and Andrew Spencer (eds.), 219–231. Oxford: Oxford University Press.

Blake, Barry J.
2001 *Case* (2nd edition). Cambridge: Cambridge University Press.

Blevins, Juliette
2004 *Evolutionary Phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Blutner, Reinhart, Petra Hendriks and Helen de Hoop (eds.)
2005 *Optimal communication*. Stanford, CA: Centre for the Study of Language and Information.

Bossong, Georg
1985 *Differentielle Objektmarkierung in den neuiranischen Sprachen*. Tübingen: Narr.

Butt, Miriam
2006 *Theories of Case*. Cambridge: Cambridge University Press.

Comrie, Bernard
1978 Ergativity. In *Syntactic Typology. Studies in the Phenomenology of Language*.Winfried P. Lehmann (ed.), 329–394. Austin / London: The University of Texas Press.

Comrie, Bernard
1989 *Language universals and linguistic typology*. Chicago: University of Chicago Press.

Comrie, Bernard and Maria Polinsky
1998 The great Daghestanian case hoax. In *Case, typology and grammar*, Anna Siewierska and Jae Jung Song (eds), 95–114. Amsterdam: John Benjamins.

Creissels, Denis and Michael Daniel
2006 Monotransitive 'give'-constructions. Exploring the periphery of ditransitives. Talk presented at the *Rara & Rarissima* conference, Leipzig.

Croft, William
2003 *Typology and Universals*. 2nd edition. Cambridge: Cambridge University Press.

Dahl, Östen
2004 *The growth and maintenance of linguistic complexity*. Amsterdam: Benjamins.

Daniel, Michael and Dmitry Ganenkov
2009 Rich case systems: limits of elaboration in Daghestanian. In *Handbook of Case*, Andrej Malchukov and Andrew Spencer (eds.), 668–686. Oxford: Oxford University Press.

Dench, Alan and Nicholas Evans
1988 Multiple case-marking in Australian languages. *Australian Journal of Linguistics* 8: 1–47.

Dixon, R. M. W.
1979 Ergativity. *Language* 55: 59–138.

Donohue, Mark
1999 *A Grammar of Tukang Besi*. Berlin / New York: Mouton de Gruyter.

Donohue, Mark
2009 Case in Austronesian languages: distinguishing case functions in Tukang Besi. In *Handbook of Case*, Andrej Malchukov and Andrew Spencer (eds.), 770–779. Oxford: Oxford University Press.

Dryer, Matthew
2005 Position of case affixes. In *The World Atlas of Language Structures*, Matthew Dryer, Martin Haspelmath, David Gil and Bernard Comrie (eds.), 210–211. Oxford: Oxford University Press.

Evans, Nicholas D.
1995 *A Grammar of Kayardild*. Berlin: Mouton de Gruyter.

Filimonova, Elena
2005 The noun phrase hierarchy and relational marking: problems and counterevidence. *Linguistic Typology* 9: 77–113.

Fortescue, Michael
1984 *West Greenlandic*. London: Croom Helm.

Givón, Talmy
1979 *On understanding grammar*. New York: Academic Press.

Givón, Talmy
1984 *Syntax: A Functional-Typological Introduction*, Vol. I. Amsterdam: Benjamins.

Hagman, Roy S.
1977 *Nama Hottentot grammar*. Bloomington: Indiana University Press.

Harris, Alice C.
2006 *Explaining exuberant agreement*. Talk presented at the "Rara & Rarissima" conference, Leipzig.

Harris, Alice C.
2008 On the explanation of typologically unusual structures. In *Language universals and language change*, Jeff Good (ed.), 54–78. Oxford: Oxford 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.

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

Haspelmath, Martin
1993 *A grammar of Lezgian*. Berlin: Mouton de Gruyter.

Haspelmath, Martin
2001 Non-canonical marking of core arguments in European languages. In *Non-canonical marking of subjects and objects*, A. Aikhenvald, R. W. M. Dixon and M. Onishi (eds.), 53–85. (Typological studies in languages, 46). Amsterdam: Benjamins.

Haspelmath, Martin
1999 Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft* 18 (2): 180–205.

Haspelmath, Martin
2003 The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In *The new psychology of language*, vol. 2, Michael Tomasello (ed.), 211–243. New York: Erlbaum Associates.

Haspelmath, Martin
2006 Against markedness (and what to replace it with). *Journal of Linguistics* 41 (2): 1–46.

Haspelmath, Martin
2009 Case terminology. In *Handbook of Case*, Andrej Malchukov and Andrew Spencer (eds.), 505–518. Oxford: Oxford University Press.

Heine, Bernd and Tania Kuteva
2002 *World lexicon of grammaticalization*. Cambridge: Cambridge University Press.

de Hoop, Helen and Andrej L. Malchukov
2008 Case-marking strategies. *Linguistic Inquiry* 39: 565–587.

Iggesen, Oliver
2005 *Case-asymmetry*. München: Lincom Europa.

Kenesei, István, Robert M. Vago and Anna Fenyvesi
1998 *Hungarian*. London: Routledge.

Kerimova, A. A. and Rastorgueva, V. S.
1975 Kategorija padeža [The category of case]. In *Opyt istoriko-tipologičeskogo issledovanija iranskix jazykov*. [*A Diachronic-Typological Study of the Iranian Languages*.] Vol. 2., V. S. Rastorgueva (ed.), 117–199. Moscow: Nauka.

Kibrik, Aleksandr E.
1985 Toward a typology of ergativity. In *Grammar inside and outside the clause*, Johanna Nichols and Anthony C. Woodbury (eds), 268–324. Cambridge: Cambridge University Press.

Kibrik, Aleksandr E.
2003 Nominal inflection galore: Daghestanian, with side glances at Europe and the world. In *Noun Phrase Structure in the Languages of Europe*, Frans Plank (ed.), 37–112 . Berlin / New York: Mouton de Gruyter

Kimball, Geoffrey. D.
1985 *A descriptive grammar of Koasati*. Doctoral dissertation, Tulane University, New Orleans.

König, C.
2006 Marked nominative in Africa, *Studies in Language* 30 (4): 705–782.

König, C.
2009 Marked nominatives. In *Handbook of Case*, Andrej Malchukov and Andrew Spencer (eds.), 535–549. Oxford: Oxford University Press.

Koptjevskaja-Tamm, Maria
1993 *Nominalizations*. London: Routledge

Kuhl, Patricia
1991 Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50: 93–107.

Kulikov, Leonid, Andrej L. Malchukov and Peter de Swart (eds)
2006 *Case, Valency and Transitivity*. Amsterdam / Philadelphia: Benjamins.
Lehmann, Christian
1995 *Thoughts on Grammaticalization*. München: Lincom.
Levin, Beth
1993 *English Verb Classes and Alternations*. Chicago: University of Chicago Press.
Luraghi, Silvia
2003 *On the Meaning of Prepositions and Cases. A Study of the Expression of Semantic Roles in Ancient Greek.* Amsterdam / Philadelphia: Benjamins.
Malchukov, Andrej
1995 *Even*. München: Lincom.
Malchukov, Andrej
2004 *Nominalization / verbalization: constraining a typology of transcategorial operations*. München: Lincom.
Malchukov, Andrej L.
2005 Case pattern splits, verb types, and construction competition. In *Competition and Variation in Natural Languages: the Case for Case*, Mengistu Amberber and Helen de Hoop (eds.), 73–117. Oxford: Elsevier.
Malchukov, Andrej L.
2006 Transitivity Parameters and Transitivity Alternations: Constraining Co-variation. In *Case, Valency and Transitivity* (Studies in Language Companion Series; 77 (2)), Leonid I. Kulikov, Andrej L. Malchukov, and Peter de Swart (eds), 329–358. Amsterdam / Philadelphia: Benjamins.
Malchukov, Andrej L.
2008 Animacy and asymmetries in differential case marking. *Lingua* 118: 203–221.

Malchukov, Andrej
2009 Rare and exotic cases. In *Handbook of Case*, Andrej Malchukov and Andrew Spencer (eds.), 635–651. Oxford: Oxford University Press.
Mallinson, Graham and Barry Blake
1981 *Language typology: cross-cultural studies in syntax*. Amsterdam: North Holland.
Maslova, Elena
2003 *A grammar of Kolyma Yukaghir*. Berlin / New York: Mouton de Gruyter.
Maslova, Elena
2009 Case in Yukaghir languages. In *Handbook of Case*, Andrej Malchukov and Andrew Spencer (eds.), 789–797. Oxford: Oxford University Press.
Mel'čuk, Igor A.
1998 *Kurs obščej morfologii*, vol. 2. Moskva: Jazyki russkoj kul'tury.
Merlan, F.
1982 *Mangarayi*. (Lingua Descriptive studies; 4). Amsterdam: North Holland.
Moravcsik, Edith A.
1995 Summing up Suffixaufnahme. In *Double case. Agreement by Suffixaufnahme*, Frans Plank (ed.), 451–484. New York / Oxford: Oxford University Press.
Mosel, Ulrike and Even Hovdhaugen
1992 *Samoan reference grammar*. Oslo: Scandinavian University Press.

Mous, Maarten
1993 *A grammar of Iraqw*. (Cushitic Language Studies; 9). Hamburg: Buske.

Nedjalkov, Igor
1997 *Evenki*. London: Routledge.

Newmeyer, Frederick J.
this volume Accounting for rare typological features in formal syntax: three strategies and some general remarks. In *Rethinking Universals: How rarities affect linguistic theory*. Jan Wohlgemuth and Michael Cysouw (eds.), 195–221. (= Empirical Approaches to Language Typology; 45). Berlin / New York: Mouton de Gruyter.

Nichols, Johanna
1994 Ingush. In *North East Caucasian languages*, Vol. 2, Rieks Smeets (ed.), 79–147. New York: Caravan Books.

Novikova, K. A.
1960 *Očerki dialektov evenskogo jazyka. Ol'skij govor*. I. Leningrad: Nauka [*A sketch of Even dialect systems. The Ola dialect.* Part I].

Novikova, K. A.
1980 *Očerki dialektov evenskogo jazyka. Ol'skij govor*. II. Leningrad: Nauka [*A sketch of Even dialect systems. The Ola dialect.* Part II].

Payne, J. R.
1980 The decay of ergativity in Pamir languages. *Lingua* 51: 147–186.

Plank, Frans(ed.)
1995 *Double case. Agreement by Suffixaufnahme*. New York / Oxford: Oxford University Press.

Plank, Frans (ed.)
2003 *Noun Phrase Structure in the Languages of Europe*. Berlin: Mouton de Gruyter.

Primus, Beatrice
1999 *Cases and thematic roles*. Tübingen: Niemeyer

Prince, Alan and Paul Smolensky
2004 *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford, Blackwell.

Silverstein, Michael
1976 Hierarchy of features and ergativity. In *Grammatical categories in Australian languages*, R. M. W. Dixon (ed.), 112–171. Canberra: Australian Institute of Aboriginal Studies.

Shibatani, Masayoshi
1977 Grammatical relations and surface cases. *Language* 53: 789–809.

Shibatani, Masayoshi
1991 Grammaticization of topic into subject. In *Approaches to grammaticalization*, Vol. II., Bernd Heine and Elisabeth Closs Traugott (eds.), 93–134. Amsterdam: Benjamins.

Song, Jae Jong
2001 *Linguistic typology: morphology and syntax*. London: Longman.

Spencer, Andrew
1991 *Morphological theory: an introduction to word structure in generative grammar*. Oxford: Blackwell.

Stilo, Don
2004 *Vafsi Folk Tales*. Wiesbaden: Reichert.

Tsunoda, Tasaku
1981 Split case-marking in verb types and tense / aspect / mood. *Linguistics* 19: 389–438.

Vydrine, Valentin
2006 Emergence of morphological cases in South Mande: from the amorphous type to inflectional? In *Case, Valency and Transitivity* (Studies in Language Companion Series; 77 (2)), Leonid I. Kulikov, Andrej L. Malchukov, and Peter de Swart (eds.), 49–65. Amsterdam / Philadelphia: Benjamins.

Wedel, Andrew
2004 *Self-organization and categorical behavior in phonology: simulating evolution of patterns in lexical contrast*. Doctoral dissertation, University of Arizona.

Zeevat, Henk
2006 Grammaticalisation and evolution. In *Proceedings of the 6th International Conference on the Evolution of Language*, 372–378.

Negatives without negators

Matti Miestamo

1 Introduction

Das grammatische Raritätenkabinett[1] assembled by Frans Plank features the following *rarum* concerning the expression of clausal negation (number 33):

> "negation expressed negatively, by omission of material present in affirmative clause".

Well-known instances of this so-called Zero Negative construction are found in some Dravidian languages in which negation can be signaled by the mere absence of tense marking without an overt marker of negation. In this paper, I take a look at this rare type of negative construction from a typological perspective.

Section 2 briefly introduces the Dravidian Zero Negative construction. Section 3 discusses the phenomenon in a cross-linguistic perspective, suggesting possible typological parallels to the absence of tense and negation markers. Section 4 addresses the possibility of ellipsis of negators in languages in which these are usually present in the negative construction. Section 5 first discusses whether and how the absence of tense and negation markers is functionally motivated and examines then the implications of negatives without negators to the markedness of negation vis-à-vis affirmation. Section 6 offers a brief conclusion.

2 The Dravidian zero negative

The *Dravidian zero negative* (Master 1946), reported from (earlier stages) of many South and Central Dravidian languages, is a negative construction in which negation is expressed without an overt marker of negation, by the mere absence of tense marking present in the corresponding affirmatives. The affirmative verb forms can be rendered schematically as ROOT-TENSE-PERSON and the negative ones as ROOT-PERSON. In Old Kannada, for example, we find the affirmative-negative correspondences in (1).

(1) Old Kannada (India; Dravidian: Southern Dravidian) (Pilot-Raichoor (2010: 268–269)[2]

a. *noːḍ-uv-eṃ*	b. *noːḍ-id-eṃ*	c. *noːḍ-eṃ*
see-FUT-1SG	see-PST-1SG	see-1SG
'I will see'	'I saw'	'I do / did / will not see'

As can be seen in (1), there is no overt marker of negation, and the negative (1c) differs from the corresponding affirmatives (1a) and (1b) only by the presence vs. absence of tense marking. The tense distinction made in the affirmative is lost in the negative.

I will not engage in a detailed description of Zero Negatives in different Dravidian languages (see Masters (1946); Subrahmanyam (1971); Pederson 1993; Pilot-Raichoor (1997, 2010) for more detailed data and analysis); the brief characterization given here should suffice as a basis for the typological discussion that I now turn to.

3 Typological parallels

More generally, the Dravidian Zero Negatives may be seen as instances of a construction type in which no overt marker of negation appears and the negative differs from the corresponding affirmative by the absence of marking of a category (or categories) present in the affirmative (be it tense or any other category). This construction type can be called the *Subtracting Zero Negative Construction*.

In my typological study of *standard negation* – the basic strategies languages use for negating declarative verbal main clauses (Miestamo 2005) – not a single language in an areally and genealogically representative variety sample of 297 languages shows this type of construction. There are three Dravidian languages in the sample, Brahui, (Modern) Kannada and Malayalam, but none of these languages have a productive Subtracting Zero Negative Construction according to my sources.

Outside of the Dravidian family, I am not aware of any languages showing this type of construction, either. This section discusses some typological parallels, i. e. structures that have some properties in common with the Dravidian Zero Negatives (or more generally, with the Subtracting Zero Negative Construction). There are two basic aspects that need to be discussed: that the marking of tense or some other categories is absent (Section 3.1), and that there is no overt marker of negation (Section 3.2).

3.1 Absence of tense marking

The absence of tense marking in the negative is better understood in the context of the distinction between *symmetric* and *asymmetric negation* (Miestamo 2005). Negatives can be divided into two basic types, *symmetric* and *asymmetric*, according to whether and how the structure of the negative differs from the structure of the affirmative in addition to the presence of negative markers. The symmetry-asymmetry distinction can be observed in constructions and paradigms.

In a *symmetric negative construction*, the negative does not differ structurally from the corresponding affirmative in any other way than by the presence of the negative marker(s), e. g., in German (2), where the presence of *nicht* is the only structural difference between negatives and affirmatives.

(2) German (Germany; Indo-European: Germanic) (constructed example)

a. *singen* 'to sing', 1SG

	AFFIRMATIVE	NEGATIVE
PRES	*ich singe*	*ich singe* ***nicht***
PST	*ich sang*	*ich sang* ***nicht***
PERF	*ich habe gesungen*	*ich habe* ***nicht*** *gesungen*
PLUPERF	*ich hatte gesungen*	*ich hatte* ***nicht*** *gesungen*

b. *singen* 'to sing', 2SG

	AFFIRMATIVE	NEGATIVE
PRES	*du singst*	*du singst* ***nicht***
PST	⋮	⋮

In an *asymmetric negative construction* the structure of the negative differs from the structure of the corresponding affirmative in other ways (as well), i. e. not (merely) by the presence of the negative marker(s), e. g., in Finnish cf. (3), where the negative marker is the negative auxiliary *e-* carrying personal inflections, and the lexical verb appears in the non-finite Connegative form.

(3) Finnish (Finland; Uralic: Finnic) (constructed example)

a. *laula-n*
sing-1SG
'I sing.'

b. ***e-n*** *laula*
NEG-1SG sing.CNG
'I do not sing.'

In a *symmetric paradigm* the correspondences between the paradigms used in affirmatives and negatives are one-to-one, e. g., in German (2), where

all affirmative declaratives, regardless of the choice of categories like tense or person, can be negated and no grammatical distinctions are lost.

In *asymmetric paradigms* the correspondences between the paradigms used in affirmatives and negatives are not one-to-one; in almost all cases of paradigmatic asymmetry some grammatical distinctions made in the affirmative are lost (neutralized) in the negative, e. g., in Maung (4), where the distinction between realis and irrealis made in the affirmative is lost in the negative obligatorily marked as irrealis.

(4) Maung (Australia; Australian: Iwaidjan) (Capell and Hinch 1970: 67)

a. *ŋi-udba*
1SG>3-put
'I put.'

b. *ni-udba-ji*
1SG>3-put-IRR.NPST
'I can put.'

c. ***marig*** *ni-udba-ji*
NEG 1SG>3-put-IRR.NPST
'I do not put.', 'I cannot put.'

Constructional asymmetry is found in 83 out of a sample of 179 languages (46 %) and paradigmatic asymmetry in 53 languages (30 %).[3]

The Dravidian Zero Negatives show both constructional and paradigmatic asymmetry.[4] The construction is asymmetric in that there are other structural differences between the affirmatives (1a) and (1b) and the negative (1c) than the presence of a negative marker, viz. the absence of tense marking. The paradigm is asymmetric in that the tense distinctions made in the affirmative (1a) and (1b) are lost in the negative (1c). Note that as there is no overt negator, the asymmetry (the absence of tense marking) serves as the indicator of the negative semantics of the clause.

It is not at all rare in the world's languages that the marking of grammatical categories in negatives is different from their marking in affirmatives. The categories that are affected the most often by asymmetry are tense-aspect-mood (TAM) categories. Constructional asymmetry affecting the marking of tense-aspect-mood is found in 46 out of 179 languages (26 %), and in 44 out of 179 languages (25 %) there is paradigmatic asymmetry whereby some tense-aspect-mood distinction(s) available in the affirmative are excluded in the negative (these numbers only include instances in which negation affects the marking of tense-aspect-mood categories directly).[5] It is not always straightforwardly clear whether a particular category in a language should be analysed as tense, aspect or mood, but it seems safe to say that tense distinctions are involved in well over a third of the paradigmatic cases.

I will now discuss some examples of how tense-aspect-mood categories can be affected. As was already seen in the Maung example (4) above, irrealis mood may be obligatory under negation in some languages and the distinction between realis and irrealis may then be lost. A different type of neutralization happens in Páez (5), where the affirmative can make a distinction between Progressive and Habitual (5a) and (5b), but the negative suffix {*-me:*} cannot occur with the Habitual and only the Progressive form is available in the negative (5c); the paradigm is thus asymmetric. The construction is symmetric in that the only difference between the negative and the (Progressive) affirmative verb form is the presence of the negative marker (the absence of the final vowel in (5c) is due to an automatic phonological process and thus does not count as asymmetry in the relevant sense).

(5) Páez (Colombia; Páezan) (Jung 1989: 102–104)[6]

a. *âts,h-a' ts,hab-na u'x-we-ts-thu*
now-TOP village-to go-IMPF-PROG-DECL.1SG
'I'm going to the village right now.'

b. *skwela-na-t, u'x-we-'*
school-to-FACT.3PL go-IMPF-HAB
'They go to school.'

c. *u'x-we-ts-**me:**-th*
go-IMPF-PROG-**NEG**-DECL.1SG
'I don't go.', 'I'm not going.'

In most of the languages in which some tense-aspect-mood distinctions made in the affirmative are lost in the negative, there still is some tense-aspect-mood marking in negatives, i. e. only some categories are excluded, as in Páez.

However, there are also a few cases in which the negative does not contain any tense-aspect-mood marking at all even though the corresponding affirmative does; in the 179-language sample there are three languages in which such a construction can be clearly identified. In one of the standard negation constructions available in Rama (example 6), the negative element *yaana* appears pre-verbally or clause-initially and the verb has no tense marking, i. e. the tense suffixes found in affirmatives (6a) and (6b) are absent in negatives (6c). In Imonda (example 7), postposed *hoi* is one of the means of expressing standard negation and with this negative marker tense marking does not appear on the verb. In Ogbronuagum (8), the Factitive, Future, and

Progressive categories are negated by a construction in which the negative marker is fused with the subject proclitic and no tense-aspect-mood marker is present.

(6) Rama (Nicaragua; Chibchan: Rama) (Grinevald Craig, no date: 106, 145, 190)

a. *nah sung-i*
1 see-PRES
'I see (it)'

b. *i-siik-u*
3-come-PST
'He came.'

c. ***yaana*** *urnga ma-ni-tang*
NEG food 2-1-give
'I am not going to give you food.'

(7) Imonda (Papua-New Guinea; Border) (Seiler 1985: 157, 172)

a. *muit ale-la-f*
Muit stay-EMPH-PRES
'Muit is here / there / in.'

b. *toad ale* ***hoi***
boys stay **NEG**
'There were no boys.'

(8) Ogbronuagum (Nigeria; Niger-Congo: Cross-River) (Kari 2000: 34–35)

a. *o-tó-lé akídɪ*
1PL-FUT-eat beans
'We shall eat beans.'

b. ***ɔjí-mɔ́ɔlʊ*** *ínə*
1PL.NEG-catch fish
'We shall not catch fish.'

What these cases have in common with the Subtracting Zero Negative Construction is that tense-aspect-mood marking found in the affirmative is absent in the negative and the tense-aspect-mood distinctions in question are not available.[7] The constructions are asymmetric as there is no non-negative form differing from the negative by the mere presence / absence of the negative marker, and the paradigms are also asymmetric in that some tense-aspect-mood distinctions made in the affirmative are unavailable in the negative.

In the above examples, just as in Dravidian Zero Negatives, the verbal forms are finite in the sense that they are neither dependent on a higher verbal element nor verbal forms used as dependent forms in other contexts. There is a subtype of asymmetric negation in which the lexical verb loses its finiteness in the negative, and among these constructions one can find some more cases of absence of tense-aspect-mood marking in the negative. In Sentani (9), the negative marker is a vowel prefix, all verbal marking (tense-aspect-mood, person-number) is lost on the verb, and the verb appears in the Non-

Temporal form, which is a non-finite form used in some dependent clauses as well.

(9) Sentani (Indonesia: West Papua; Sentani) (Hartzler 1994: 52–53)

a. *neyæ u-eu-ne*
he say-3SG.SUBJ.R-3SG.OBJ
'He is saying to him.'

b. ***o-****boro-i*
NEG-hear-NTMP
'I / you / he / she / we / you / they didn't / do(es)n't / won't hear.'

In Sentani the absence of tense-aspect-mood marking is not due to negation in the same sense as in the cases discussed earlier: the negative construction uses a non-finite form of the lexical verb, and because this non-finite form happens to lack tense-aspect-mood marking in this language, such marking is also absent in negatives. In negative constructions in which the lexical verb loses its finiteness, tense-aspect-mood marking is only rarely lost; it happens in only two more languages in the 179-language sample: Inanwatan and Rama (in a different construction from the one discussed above).

A few more cases of absence of tense marking may come about through the ellipsis of auxiliaries introduced to carry verbal inflections when the lexical verb loses its finiteness. In Apalaí, the negative marker *-pyra* appears on the lexical verb that becomes non-finite and the copula is introduced to carry tense and subject marking (10).

(10) Apalaí (Brazil; Cariban) (Koehn and Koehn 1986: 64)

a. *isapokara [Ø]-ene-no*
jakuruaru.lizard [1>3]-see-IMPST
'I saw a jakuruaru lizard.'

b. *isapokara on-ene-****pyra*** *a-ken*
jakuruaru.lizard 3-see-**NEG** 1-be.IMPST
'I did not see a jakuruaru lizard.'

The copula is, however, optional if the grammatical categories expressed by it are obvious from the context. Absence of tense-aspect-mood marking in negatives through the optional loss of copula in this type of negative construction can be found in seven of the 179 languages: Apalaí, Araona, Lower Grand Valley Dani, Quileute, Suena, Waorani and Yareba.

In this section I have shown that the loss of tense-aspect-mood distinctions in negatives is by no means unique to Dravidian languages (although the

complete absence of tense-aspect-mood marking present in the corresponding affirmatives is rather uncommon). I will now turn to the much rarer aspect of the Dravidian negatives — the absence of an overt marker of negation.

3.2 Negatives with no overt markers of negation

As discussed above, Subtracting Zero Negative Constructions have not been found outside of the Dravidian family. The *Raritätenkabinett* suggests that such a construction might be found in two non-Dravidian languages, viz. Achumawi and Malakmalak, and a similar claim has also been made concerning Karitiâna. In this section I will examine all three cases, and discuss to which extent they fit the definition of the Subtracting Zero Negative Construction. Finally, I will take a look at negation in Igbo, which clearly does not show a Subtracting Zero Negative Construction, but comes closest to having no overt negative marker in standard negation in my 297-language variety sample.

3.2.1 Malakmalak

The *Raritätenkabinett* mentions Malakmalak as a possible case of Zero Negation. The source for this information is Forest (1993), but if we take a closer look at Forest (1993: 61–63) or the original source (Birk 1976), there is no reason to see a Subtracting Zero Negative Construction in this language. Negation can be expressed using a negative construction in which the negative particle (adverb) *akana* occurs preverbally (11).

(11) Malakmalak (Australia; Australian: Northern Daly) (Birk 1976: 124)

akana *pilp yi-nma-ŋayi-wa*
NEG slap 3SG.M.SUBJ-FUT-3SG.F.OBJ-FUT
'He will not slap her.'

Often, however, negative meaning is expressed with constructions whose primary function is not (pure) negation. The Adversative (marked by the suffixes *-tan* and *-wur*) is one of the means for expressing these indirect negatives. According to Birk (1976: 92), the Adversative "renders of no account the action described by the verb root to which it is suffixed". In (12a) we can see *-tan* in a sort of frustrative ("in vain") function and in (12b) it is translated as a negative.

(12) Malakmalak (Australia; Australian: Northern Daly) (Birk 1976: 92–93)

a. *kay-**tan** a-ya-nö*
call-**AVR** 1SG.SUBJ-PNCT-3SG.M.OBJ
'I called him but there was no reply.'

b. *tikka-**tan** yö-nuŋka-kka*
come.back-**AVR** 3SG.M.SUBJ-FUT-FOC
'He will not be coming back.'

Another inherently non-negative suffix, *-manŋa*, can be used as a negator in a restricted number of contexts.[8] What is relevant for our concerns is that even if Malakmalak often expresses meanings that would be expressed with direct negatives in most other languages with less direct ways of negation, these categories are overtly marked, and clearly do not constitute parallels to the Dravidian Zero Negatives. The avoidance of explicit marking of negation is of a very different kind — functional (semantic / pragmatic) rather than formal as in the Subtracting Zero Negative Construction. Wintu (USA; Penutian: Wintuan) shows a similar kind of avoidance of the use of direct negatives (see Lee 1946).

A slightly different case may be found in Ungarinjin (Australia; Australian: Wororan; Rumsey 1982). In this language, negatives are obligatorily marked as irrealis, i. e. the negative particle occurs together with the irrealis form of the verb. But the irrealis form can occasionally be used with a negative meaning without a preceding negative particle. Is this then a negative without a negator? According to Rumsey (1982: 91), the irrealis form of the verb comes from an earlier negative form, i. e. a negative form has been reanalysed as irrealis. In the rare cases in which the irrealis form still expresses negation alone, it may perhaps be analysed as a negative marker as was the case at an earlier stage of the language. In any case, the category is overtly marked and thus clearly different from the Subtracting Zero Negative Construction.

3.2.2 Achumawi

In Achumawi (13), standard negation is expressed with a construction in which the lexical verb is nominalized, and the marker *tsé-* is attached to the existential copula which appears as the finite element of the clause.

(13) Achumawi (USA; Hokan: Palaihnihan) (de Angulo and Freeland 1931: 97, 112)

a. *s-ă'm-á*
1SG-eat-FV
'I eat.'

b. ***tsé**-s-ùw-í d-ámm-ì*
NEG-1SG-be-FV NMLZ-eat-FV
'I do not eat.'

According to Forest (1993: 53) there is no negative element in the construction — the element *tsé-* is taken to be a special participial form of the verb 'be'. However, nowhere in de Angulo and Freeland (1931; the source that Forest also uses) is *tsé-* treated as a form of the verb 'be'. The following quote shows clearly that they treat it as a negative element (de Angulo and Freeland 1931: 112):

> The "auxiliary" [form of the verb be] is used especially to form the negative, in connection with the element *tsé-* (to be classed as an "adjective"?). The combination is then followed by the Denominative form of the verb to be expressed negatively, thus; *tsé-sùwí d-ámm-ì* "I don't eat" (not-I-am eating)

Although they are not certain about the categorial status of *tsé-*, they clearly treat it as a negative element, glossing it as "not" and calling it "the ordinary negative" (de Angulo and Freeland 1931: 116) and "negative particle" (de Angulo and Freeland 1931: 117); they do not give any indication that it would be a form of the verb 'be'.[9] There can be no other conclusion than that *tsé-* is a negative marker.

However, if we look beyond standard negation in Achumawi, we can find constructions in which it seems to be the case that no overt marker of negation is present. The element *nám* '(not) yet' has positive meaning when followed by a verb with Indicative pronominal prefixes and negative meaning when Subordinate pronominal prefixes are found on the following verb (14).

(14) Achumawi (USA; Hokan: Palaihnihan) (de Angulo and Freeland 1931: 87)

a. *nám y-ă'm-á*
yet 3SG-eat-FV
'He is still eating.'

b. *nám t-ă'm-á*
yet 3SG.SBRD-eat-FV
'He is not eating yet.'

This is the only example of this construction found in the source, and there is no further information given. In a passage dealing with the Subordinate category (de Angulo and Freeland 1931: 89–90), we learn further that its functions are as follows: it is used in interrogatives, in complement clauses and to express the idea 'to have performed an action once before'. It is diffi-

cult to interpret *nám* as involving a negative element when it occurs in front of a Subordinate verb but not involving one in connection with an Indicative verb. On the basis of the information available in de Angulo and Freeland (1931), we may consider this to be a negative construction without negators. However, it is not an instance of Subtracting Zero Negative Construction, since it is not the case that the construction differs from the corresponding affirmative by the absence of marking of a category present in the affirmative — the difference is in the use of a different set of subject prefixes (Subordinate instead of Indicative). Forest (1993: 17, citing Hilders and Lawrance 1957) mentions a similar construction from the Eastern Sudanic (Nilo-Saharan) language Teso.

A third construction to be discussed from Achumawi is the negation of the Periphrastic Future of Eventuality (15).

(15) Achumawi (USA; Hokan: Palaihnihan) (de Angulo and Freeland 1931: 112)

a. *d-ùʔmáˑd-ìˑgú-s-ùw-á*
NMLZ-sleep-PURP-1SG-be-FV
'I will be sleeping.'

b. *d-ùʔmáˑd-ìˑgú-**d-ùts-ì**-s-ùw-á*
NMLZ-sleep-PURP-**NMLZ-be-FV**-1SG-be-FV
'I will not be sleeping.'

In this construction, negation is expressed by the addition of *dùtsì*, a nominalized form of the verb 'be', without the presence of any overtly negative element. According to de Angulo and Freeland (1931: 112, 116–117), this is due to the fact that the regular *-áˑmè* (negative after Denominatives) is not compatible with the Purposive *-(ìˑ)gú*. A possible analysis would of course be to see *dùtsì* as the marker of negation in this particular construction, and the connection with the verb 'be' would then be only diachronic. Be it as it may, this construction is not an instance of Subtracting Zero Negative Construction, since morphemes are added rather than deleted when the affirmative is turned into a negative.

3.2.3 Karitiâna

According to Landin (1984: 1, 11–12), Karitiâna marks affirmatives overtly while negatives are unmarked, i. e. negatives differ from affirmatives by the

absence of the Affirmative markers and no negative marker is present. Furthermore, tense-aspect suffixes are also absent in negatives. This is illustrated by the pair of examples in (16).

(16) Karitiâna (Brazil; Tupian: Arikem) (Landin 1984: 11)[10]

a. *y*	*ta-oty-j*	*ỹn*	b. *y*	*oty*	*ỹn*
1SG	AFF-bathe-TNS	1SG	1SG	bathe	1SG
'I will bathe.'			'I will not bathe.'		

Some further differences between affirmatives and negatives are also observed. In negatives containing a transitive verb the third person direct object pronoun *i-* is inserted. In the case of intransitive stems with initial stress the epenthetic element *ry-* appears, and on consonant-final stems the epenthetic element *-y* is found. Thus in the case of vowel-final intransitives with non-initial stress, negatives seem to differ from the corresponding affirmative by the absence of the Affirmative marker (and tense marking).

Storto's analysis of Karitiâna negatives differs from Landin's and the essential differences can be briefly summarized as follows (Storto 1999 and p. c.). There is an overt negator in the negative construction, the postverbal element *pa^{d}ni*, but this element can also be omitted (17).[11]

(17) Karitiâna (Brazil; Tupian: Arikem) (Storto 1999: 68–69)

a. *i-soʔoːt-o*	***(pa^{d}nĩ)***	b. *i-a-okɨ*	***(pa^{d}nĩ)***
3SG-see-EPN	**NEG**	3SG-PASS-kill	**NEG**
'(S)he didn't see.'		'(S)he wasn't killed.'	

Since the prefixal markers that Landin calls Affirmative are absent not only in the negative, but also in some non-declarative contexts such as imperatives, interrogatives and quotes, they are analysed as Declarative mood markers, marking "a statement that the speaker believes to be true." (Storto 1999: 163, n. 56). Their absence does not signal negation.

Everett's (2006: 328–332) account is, again, different. He notes (Everett 2006: 329) that the negator *pa^{d}ni* "can (and typically does) follow the verb in negative clauses", cf. (18d). Apart from this optional negator, intransitive and transitive verbs behave differently. In the case of vowel-final intransitives with non-inital stress, negation is unmarked (if *pa^{d}ni* is not used) (18a). Consonant-final intransitives receive the negative suffix *-ɨ* / *-ĩ* (18b) and (18c) and initially stressed intransitives may be prefixed with the negative marker *ɾi-* (18c); the elements treated as epenthetic elements by Landin and Storto are thus taken to be morphological negative markers. In transitive negatives the

prefix *i-* appears (18d). This element was analysed as a third person marker by both Landin and Storto, but Everett treats it as an irrealis marker, since it is also used in interrogatives and imperatives and is not restricted to third person contexts (see also Everett 2006: 253–256).

(18) Karitiâna (Brazil; Tupian: Arikem) (Everett 2006: 255, 328–329)[12]

a. *i-oti (ĩn)*
1SG.ABS-bathe (1SG)
'I did not bathe.'

b. *i-terekterej-**ĩ** (ĩn)*
1SG.ABS-dance-**NEG** (1SG)
'I did not dance.'

c. *i-**ɾi**-mbik-**i** ĩn*
1SG.ABS-**NEG**-sit-**NEG** 1SG
'I did not sit.'

d. *ĩn **i**-soki **padni** eppa*
1SG **IRR**-break **NEG** oar
'I did not break the oar.'

e. ***i**-atoɾ-i ĩn bɨ̃pãn*
IRR-take-FUT 1SG arrow.OBL
'I will not take the arrow.'

f. *ĩn i-opiso-t*
1.ERG INTR-hear-NFUT
'I heard.'

g. *ĩn na-opĩ̃ː-t (ĩn)*
1SG NSAP-cut-NFUT 1SG
'I cut it.', 'I cut something.'

h. *ĩn **i**-opĩ̃ː-t (ĩn)*
1SG **IRR**-cut-NFUT 1SG
'I did not cut it.', 'I did not cut.'

The markers analysed as Affirmative by Landin and Declarative by Storto are also treated differently by Everett (2006: 284–290, 409–424): he agrees with the earlier analyses that they are not used in negatives, imperatives and interrogatives, but does not consider them to be markers of mood. Instead, they are analysed as markers of voice, marking Speech-Act-Participant and Non-Speech-Act-Participant voices. Examples of affirmatives with and without the use of the Non-Speech-Act-Participant prefix can be seen in (18f) and (18g), cf. (18h). The examples in (18) also show that tense is not systematically absent in negatives, but tense marking varies according to the clausal construction used in a more complex way.

The Karitiâna data have been analysed in different ways by different authors. All these analyses point to the direction that there is an environment (vowel-final intransitives with non-inital stress) in which negation may be completely unmarked in case the postverbal negator is ellipted. In other contexts there is always an overt marker, although in the case of transitives, following Everett's analysis, the marker is not a dedicated negative marker but a more general irrealis marker (cf. the discussion of Ungarinjin on page 177). Contrasted with affirmatives marked with the (Non)-Speech-Act-Participant voice, example pairs can indeed be found in which the affirmative contains

a marker and the negative differs from it by the absence of this marker, e. g., (16a) vs. (16b). These look very much like instances of Subtracting Zero Negative Construction. However, if the system is considered as a whole, Karitiâna cannot be said to use a Subtracting Zero Negative Construction to express negation.

3.2.4 Igbo

In Igbo, the basic tense-aspect-mood and polarity distinctions are made by the presence vs. absence of the vowel prefix and by varying tone. The (indicative) system shows a basic distinction between imperfective and perfective aspect. The imperfective has the prefix in the negative but not in the affirmative (19a), (19b), whereas in the perfective the opposite pattern is found, the prefix occurring in the affirmative but not in the negative (19c).[13] Affirmatives and negatives also differ in their characteristic tone patterns.

(19) Igbo (Nigeria; Niger-Congo: Igboid) (Green and Igwe 1963: 75, 119, 140)

a. *anyì cì anʊ́*
we carry bit.of.meat
'We are (were) carrying bits of meat.'

b. *anyɪ a-cɪ́ akhʊ́*
we PFX-carry palm.nut
'We are (were) not carrying palm nuts.', 'We did not carry palm nuts.'

c. *hã̂ gà-rà ɪ-cɪ́ akhʊ́ ɔ̀fɔ á-cɪ-ta èkwè cì-tà-ghì*
they go-PST PFX-carry palm.nut Ofo PFX-carry-PST Ekwe carry-PST-EMPH
'They went to fetch some palm nuts. Ofo fetched some and Ekwe did not.'

With inseparable subject pronouns, the prefix is never used, and in such cases some distinctions are made by tone only cf. (20a) and (20b).

(20) Igbo (Nigeria; Niger-Congo: Igboid) (Green and Igwe 1963: 75, 119)

a. *ɔ cì anʊ́*
(s)he carry bit.of.meat
'(S)he is / was carrying bits of meat.'

b. ɔ cɨ́ akhʊ́
(s)he carry palm.nut
'(S)he is / was not carrying palm nuts.', '(S)he did not carry palm nuts.'

The functional load carried by tone is high. Sometimes even the tonal patterns are identical, thus e. g. (20b) is ambiguous between a negative and an exclamative affirmative. Such ambiguities may arise in different parts of the verbal system, and various (non-inflectional) affixes can be used to disambiguate. Negatives often have the emphatic suffix *-ghɪ*, but this is not obligatory, and it can occur in the affirmative too (cf. Green and Igwe 1963: 60).[14]

Following the analysis of the basic aspectual and polarity distinctions outlined above, we may conclude that there is no overt segmental negator in Igbo — the distinction between affirmation and negation is expressed by the interaction of tone and presence vs. absence of prefix. Negatives have their own characteristic tone patterns and in this general sense we may say that these tone patterns are the overt markers of negation.

In practice, however, the same tone patterns may occur in different functions in different contexts (cf. the ambiguity of (20b)), and thus they are not dedicated markers of negation. The specific combination of proclitics and tones that occur in the negative marks the clauses as negative, but the same segmental elements and tone patterns have other functions in the verbal system as well. Discussing all the possible combinations in detail is beyond the scope of this paper, but the above discussion shows clearly that identifying a specific marker as an overt marker of negation is not straightforward in Igbo.

As I have argued in Miestamo (2005: 272), a similar situation obtains in Degema (Nigeria; Niger-Congo: Edoid). In terms of the symmetry-asymmetry distinction, we can say that, just as in the Dravidian Zero Negatives, the asymmetry alone renders the sentences negative in these languages.

4 Ellipsis of negators

Expression of negation without negators may come about through the ellipsis of negative markers. As already seen in example (3) on page 171, Finnish expresses negation with an asymmetric construction in which the negative marker is the negative auxiliary *e-* carrying personal inflections, and the lexical verb appears in the Connegative form (which is formally an uninflected stem and does not contain any negative marking).

Several authors, e. g., Savijärvi (1977, 1981), have reported that in their dialectal material, the negative auxiliary is sometimes ellipted (21).

(21) Häme dialect of Finnish (Finland; Uralic: Finnic) (Savijärvi 1977: 28)

su-lla muu-ta-ka tyä-tä o
2SG-ADE other-PART-NPI work-PART be.CNG
'(But) you have no other work either.'

The negative meaning is clear in (21), but no negative marker is present. The Connegative form of the verb and the negative polarity clitic *-kA* are enough to convey negative force here. It should also be noted that even if we changed these elements into elements occurring in declarative affirmatives, the word order would not be possible for a positive statement, at least not in the context in which this clause has been uttered. There are thus enough clues that this is indeed a negative, although no overt negator is present.

Ellipsis of the negative auxiliary has also been reported from other Uralic languages, e. g., Estonian dialects (see Honti 1997: 165–166 and references therein; Klaus 2009).

Similar negatives without the negative auxiliary are found in contemporary spoken Finnish, and as Kotilainen (2007) observes, they are also increasingly attested in written language, being used frequently in internet chats, cf. (22).

(22) Colloquial Finnish (Finland; Uralic: Finnic) (Kotilainen 2007: 7)

<cacha> Tycho, mä töissä oo!! :O

<cacha> Tycho mä tö-i-ssä oo :O
NAME NAME 1SG work-PL-INE be.CNG :O
'<cacha> Tycho, I'm not at work!! :O'

In (22) the negative meaning is again clear, although no overt marker of negation is present. The Connegative form of the verb is enough to express the negative meaning here, and again, word order is different from any felicitous positive expression in the same context. Kotilainen (2007) points out that these constructions occur as a strong reaction to what has been said before, cf. the exclamation marks and the shouting smiley in (22). According to him, this is not a case of ellipsis of the negative auxiliary in contemporary Finnish, but has been conventionalized as a special affective negative construction.

Summarizing, negatives without overt negation can be found in Finnish either due to ellipsis of the negator or in a special affective construction, and

the other structural differences between negatives and affirmatives then convey negative meaning. However, in contrast to the Dravidian Zero Negatives, these are not instances of standard negation, i. e. they do not belong to the basic means Finnish employs for the simple / neutral negation of declarative verbal main clauses. Furthermore, they do not fall under the definition of the Subtracting Zero Negative Construction, since negation is not expressed by subtraction of material present in the affirmative.

Ellipsis of negative marking has also been reported from the Australian language Nunggubuyu. There is extensive structural asymmetry between affirmatives and negatives in this language: affirmatives and negatives differ in realis-irrealis marking, most negatives being obligatorily irrealis-marked, and there are some other differences in the marking of verbal categories as well; furthermore, nouns and demonstratives in the scope of negation also show different behaviour from the affirmative, see Heath (1984: 526) for a summary and Miestamo (2005: 106–108, 328–329, 429–430) for discussion. According to Heath (1984: 531), the negative marker is sometimes ellipted, and then the asymmetries serve to mark negation alone.

This section has shown that in some languages in which negation is asymmetric, the negative marker may sometimes fall out and the asymmetries can then distinguish these negative clauses from affirmatives without the presence of an overt marker of negation. I have not encountered this in the grammars of any other language of my sample. This may be because the phenomenon is indeed very rare. I would, however, tend to think that it might be somewhat more widespread, but as its textual frequency is low in the languages in which it occurs, descriptive grammarians usually either ignore it or leave it out as a marginal phenomenon.

5 Discussion

5.1 Functional motivations

In this section I will address the functional motivations behind the Dravidian Zero Negatives. I will start from the motivations for the loss of tense distinctions and discuss the absence of tense marking further below. The following principles are proposed as general motivations for the cross-linguistically recurrent types of negative structures in (Miestamo 2005: Ch. 5): Symmetric negatives are language-internally analogous to the affirmative, copying its linguistic structure; they are ultimately motivated by pressure for system cohe-

sion. Asymmetric negatives copy different aspects of the functional asymmetry between affirmation and negation and are thus language-externally analogous to these functional-level asymmetries (for the notions of language-internal and language-external analogy, see also Itkonen 2005). The specific functional-level asymmetry proposed as explanation for the existence of paradigmatic asymmetry in which grammatical distinctions are lost in negatives, i. e. the type also represented by the Dravidian Zero Negatives, emerges from the discourse context of negation.[15]

Negatives are typically used as denials of propositions that are in some sense present in the context, so when negatives are uttered, their temporal and other properties are usually familiar to the speaker and hearer and need not be further specified. In some languages this has grammaticalized as obligatory neutralization of (some) grammatical distinctions in the negative. Another functional level asymmetry between affirmatives and negatives that can contribute to the loss of grammatical distinctions in the negative is that it is often more difficult and less relevant to attribute temporal and other specifications to non-realized, e. g., negated, events than to realized ones.

Concerning the absence of negative marking in Dravidian, many authors (e. g., Pederson 1993) have argued that the Zero Negatives have come about through phonological erosion of the negative marker. According to Pilot-Raichoor (1997, 2010), the phonological erosion account is not plausible, since all hypotheses following that line in the literature contain some unresolved problems and controversies. She argues that the construction has a semantic motivation linked to the specific meanings of the Dravidian tense morphemes, and that favourable conditions for its structural development were created by the shift from analytic to synthetic type that the languages underwent in that historical period. Not being a Dravidianist, I will not engage in a detailed discussion of the history of the Dravidian construction, but keep my focus on the typological perspective.

There is one way in which typological investigations could lend support to Pilot-Raichoor's account: demonstrating that the type of tense-aspect-mood semantics found in Dravidian correlates with the presence of a Subtracting Zero Negative Construction. A typological investigation charting the cross-linguistic frequency and distribution of such tense-aspect-mood meanings is beyond the scope of the present study. But even if this information were available, no such correlation could be demonstrated since Subtracting Zero Negative Constructions are not found in any non-Dravidian language. In this light, one must agree with Pilot-Raichoor's conclusion that the uniqueness of the

Dravidian Zero Negative construction is due to the rarity of the simultaneous occurrence of all the factors conditioning its development — a situation underlying most (if not all) *rara* and *rarissima*.

We have seen that the absence of negative markers is very rare. Tense and aspect distinctions are commonly lost in negatives, and in some languages, the tense-aspect-mood markers present in the affirmative are completely absent in the negative. Yet, in all these cases, except in the Dravidian Zero Negatives, there is overt marking of negation. This conforms to the observation, which has been common knowledge at least since Jespersen (1917), that negatives very often need reinforcement and tend to be expressed by more, rather than less, phonetic material. The best known example of the effects of this need is the development of negation, e. g., in English and French, commonly referred to as *Jespersen's Cycle*, whereby emphatic elements are needed to make negative meaning more explicit and these elements are then reanalysed as negative markers, which may later allow for the phonetic reduction and loss of earlier negative elements. And indeed, the need for reinforcement is doing its work in Igbo as well (cf. Section 3.2.3): the emphatic suffix seems to be becoming an overt negative marker used with the constructions containing no dedicated segmental markers of negation. The ultimate motivation for the need for reinforcement comes from the discourse context of negatives: as the prototypical use of negation is denial of semantic contents that are implicitly or explicitly present in the context, they often constitute somewhat abrupt speech acts, and therefore need emphasis on the negativity (see Miestamo 2005: 197–199, 209–210 for more discussion). The Dravidian Zero Negatives go drastically against this well-motivated tendency.

Coming back to the cases of ellipsis in Section 4, since constructions first arise in performance and only later conventionalize as grammatical constructions, the ellipsis of negators may indeed provide a source for negatives without negators. However, as negation needs clear and often emphatic expression, the elliptic cases cannot easily develop into grammatical constructions as such, and a new negative element is likely to emerge to make sure that negative force is conveyed. Concerning the newly conventionalized affective negative construction without overt negator in Finnish (see example (22) on page 184), Kotilainen (2007) suggests that the pronominal negative polarity item *mitään* 'anything' might be on its way to becoming a negative marker in this construction (cf. the Estonian negative element *mitte* which is etymologically the same element).

5.2 Negatives without negators and typological markedness

The foundations of the typological conception of markedness were laid by Greenberg (1966). Following Croft (2003), the criteria for typological markedness can be reduced to overt coding, behavioural potential and frequency.[16] According to the overt coding criterion, the marked category is expressed by at least as many morphemes as the unmarked one. Behavioural potential is divided into two criteria: on the one hand, the paradigmatic potential of the unmarked category is at least as high as that of the marked one, i. e. at least as many grammatical distinctions can be made in connection with the unmarked category as with the marked one, and on the other, the distributional potential of the unmarked category is at least as high as that of the marked one, i. e. the unmarked category may itself be embedded in at least as many contexts as the marked one. Finally, according to the frequency criterion, the unmarked category occurs at least as frequently as the marked one.

The frequency criterion is not of interest here, since there is certainly no reason to expect that a standard negation construction without an overt marker of negation would behave differently from one with overt negators with respect to the textual frequency of negatives vs. affirmatives — affirmatives are more frequent than negatives no matter how negation is expressed. The behavioural criteria are more interesting in this context, since the Dravidian Zero Negatives do indeed show negatives as more marked than affirmatives: more grammatical distinctions are made in the affirmative than in the negative (just as in the numerous cases in which grammatical distinctions are unavailable in the negative discussed in Section 3.1). In this context, the most interesting criterion is overt coding, as some have claimed that the Dravidian Zero Negatives constitute a counterexample to the markedness of negation vis-à-vis affirmation in this sense.

It is true that there is more phonological and morphological material in the Old Kannada affirmative verb forms (1a) and (1b) on page 170 than in the negative (1c), and if we look at the whole verb form, we may say that the affirmative is coded by a longer form; but this is not what the overt coding criterion is about. There is no overt marking of affirmation either – the tense markers express tense-aspect-mood, not affirmation – and since the overt coding criterion says that the marked category is expressed by at least as many morphemes as the unmarked one, a situation in which neither category is overtly marked is not in conflict with the markedness of negation. Polarity – either affirmative or negative – is not overtly coded in any morpheme in the verb,

and what we are dealing with here is a global constructional meaning carried by the whole verb form; this is also true of the other cases of negation without negators discussed in this paper.

In conclusion, the Dravidian Zero Negatives do not constitute counter-evidence to the markedness of negation on the overt coding criterion either. There are no languages, Dravidian or other, in which (non-emphatic) affirmation receives overt marking while negation is unmarked (see Miestamo 2007 for more discussion on the typological markedness of negation).

6 Conclusion

In this paper, I have discussed, from a typological perspective, the Dravidian Zero Negative construction, or more generally the Subtracting Zero Negative Construction — a construction in which no overt marker of negation appears and the construction differs from the corresponding affirmative by the absence of marking of categories present in the affirmative. I have shown that it is indeed unique to Dravidian. That tense-aspect-mood distinctions are lost is not in itself rare, but expressing negation without any overt marker of negation is, and the combination of these two features makes the Dravidian construction unique. I have shown that, very rarely, negation can be expressed without overt negators in some other languages as well, but these constructions are clearly different from the Dravidian Zero Negatives. I have also discussed some suggested cases of zero negation that turn out not to be such. I have looked at the ellipsis of negators, and finally, I have examined the functional motivations behind negatives without negators and discussed these negatives from the point of view of typological markedness.

As an overall conclusion to this paper, I would like to reformulate the answers to two questions: First, why are negatives without negators rare? My view is that negatives make strong speech acts and need strong expression; it is not communicatively efficient to express negation only covertly. Why, then, are negatives without negators possible in the first place? My answer to this question is that there is functional asymmetry between affirmation and negation, and this grammaticalizes as structural asymmetry in many languages; this structural asymmetry provides other cues to identify negation in addition to the presence of overt negative markers, and thus enables negation without negators in some rare cases.

Abbreviations

1 = first person; 2 = second person; 3 = third person; ABS = absolutive; ADE = adessive; AFF = affirmative; AVR = adversative; CNG = connegative; DECL = declarative; EMPH = emphatic; EPN = epenthetic; ERG = ergative; F = feminine; FACT = factative; FOC = focus; FUT = future; FV = final vowel; HAB = habitual; IMPST = immediate past; IMPF = imperfecive; INE = inessive; INTR = intransitive; IRR = irrealis; M = masculine; NAME = proper name; NEG = negative; NFUT = non-future; NMLZ = nominalization; NPI = negative polarity item; NPST = non-past; NSAP = non-speech-act-participant; NTMP = non-temporal; OBJ = object; OBL = oblique; PART = partitive; PASS = passive; PERF = perfect; PFX = prefix; PL = plural; PLUPERF = pluperfect; PNCT = punctual; PRES = present; PROG = progressive; PST = past; PURP = purposive; R = realis; SBRD = subordinate; SG = singular; SUBJ = subject; TNS = tense; TOP = topic

Acknowledgments

I wish to thank my audience at the *Rara & Rarissima* conference and the editors of this volume for their useful comments on the manuscript. The research reported here has been conducted at the universities of Helsinki and Antwerp, and financed by these universities and the Academy of Finland; the support of these institutions is gratefully acknowledged.

Notes

1. See http://typo.uni-konstanz.de/rara/intro/
2. For each language the main geographical location and genealogical affiliation (family: genus) is given following the classification by Dryer (2005) in *The World Atlas of Language Structures*.
3. The numbers given in this section are based on a 179-language subsample of the 297-language variety sample. The subsample is areally and genealogically balanced so that each continent-size linguistic area (macroarea) is represented in proportion to its internal genealogical diversity (see Miestamo 2005: 35–36 for details).
4. Although the subtypes of asymmetric negation proposed in Miestamo (2005) are not discussed in this paper, it may be interesting for some readers to note that the constructional and paradigmatic asymmetry in the Dravidian Zero Negatives belongs to type A / Cat / TAM.
5. To be more precise, these numbers only concern the subtypes of asymmetric negation labelled A / NonReal and A / Cat / TAM in Miestamo (2005), in which the asymmetry affects tense-aspect-mood categories directly: 10 of the 44 cases of neutralization are of

subtype A / NonReal and 34 are of subtype A / Cat / TAM (see Miestamo 2005: 179–180); 8 languages show constructional asymmetry of subtype A / NonReal and 39 show constructional asymmetry of subtype A / Cat / TAM, one language, viz. Warndarang, showing both types (see Miestamo 2005: 173, 175, 360). In the other subtypes there are some cases in which the marking of tense-aspect-mood categories is affected indirectly, e. g., the Sentani construction to be discussed in example (9) on page 175.

6. The comma marks palatalization in Jung's orthography.
7. Note, however, that in both Imonda and Rama there are other negative constructions in connection with which tense-aspect-mood marking does appear, and in Ogbronuagum, as well, overt tense-aspect-mood marking appears in the perfect negative construction.
8. This is actually a combination of two suffixes both of which have a primary function characterizable as elative: *-manʸ* 'departing from', *-ŋa* 'away from the speaker' (see Birk 1976: 87, 91–95, 105, 114).
9. The element *tsé-* occurs in the same position prefixed to the copula as many adjective stems (de Angulo and Freeland 1931: 85); formally it might then be classified as an adjective. Phonetically it does resemble some forms of the verb 'be' but not even a diachronic link to the verb be is suggested in the source. The participle of 'be' is *tìdzí*. Even if there were a diachronic link, it would still clearly be a negative element in the synchronic analysis of the construction.
10. The page numbers refer to the pagination in the version downloaded from the internet.
11. Landin (1984: 15) also mentions this element but according to him, it only occurs in negative sentences used as responses to polar questions.
12. In the original source, the gloss of (18c) lacks the negative elements, but a comparison with the text and with the glosses of the other examples makes it is obvious that this is a typographic error.
13. The alternative perfective translations of the negated imperfectives in (19b) and (20b) are due to the fact that the negative perfective cannot occur in an isolated utterance or initiating discourse, and the aspectual distinction is neutralized in these contexts.
14. In fact, Emenanjo (1987: 172) treats it as a negative suffix but says it "may be optionally deleted at the surface structure level".
15. Those cases of neutralization in which the negative is obligatorily marked as irrealis or some other non-realized category are of course more directly motivated by the fact that negation itself belongs to the non-realized (cf. Miestamo 2005: 208–209).
16. Croft speaks about the "structural criterion", but I find the term "overt coding criterion" clearer. Similarly, Croft uses the term "inflectional potential" for one of the behavioural potential criteria, but I prefer "paradigmatic potential" which is more neutral about the coding means by which the paradigmatic choices are expressed.

References

Birk, David. B. W.
1976 *The Malakmalak Language, Daly River (Western Arnhem Land).* (= Pacific Linguistics; B-45). Canberra: Australian National University.

Capell, Arthur, and H. E. Hinch
1970 *Maung Grammar*. (= Janua Linguarum, Series Practica; 98). The Hague: Mouton.

Croft, William
2003 *Typology and Universals.* 2nd edition. Cambridge: Cambridge University Press.

de Angulo, Jaime, and L. S. Freeland
1931 The Achumawi language. *International Journal of American Linguistics* 6: 77–120.

Dryer, Matthew S.
2005 Genealogical language list. In *The World Atlas of Language Structures*, Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), 584–644. Oxford: Oxford University Press.

Emenanjo, E. Nolue
1987 *Elements of Modern Igbo Grammar: A Descriptive Approach.* Ibadan: University Press.

Everett, Caleb
2006 *Gestural, perceptual, and conceptual patterns in Karitiana*. Doctoral dissertation, Rice University.

Forest, Robert
1993 *Négations: Essai de syntaxe et de typologie linguistique.* (= Collection Linguistique; LXXVII). Paris: Klincksieck.

Green, M. M., and G. E. Igwe
1963 *A Descriptive Grammar of Igbo*. London: Oxford University Press.

Greenberg, Joseph
1966 *Language Universals, with Special Reference to Feature Hierarchies.* The Hague: Mouton.

Grinevald Craig, Colette
no date *A grammar of Rama*. Report to National Science Foundation BNS 8511156.

Hartzler, Margaret
1994 Sentani. In *Typological Studies in Negation*, Peter Kahrel and René van den Berg (eds.), 51–64. (= Typological Studies in Language; 29). Amsterdam: Benjamins.

Heath, Jeffrey
1984 *Functional Grammar of Nunggubuyu*. Canberra: Australian Institute of Aboriginal Studies.

Hilders, John H., and Jeremy C. D. Lawrance
1957 *An Introduction to the Ateso Language.* Kampala: Eagle Press.

Honti, László
1997 Die Negation im Uralischen II. *Linguistica Uralica* 1997 (3): 161–176.

Itkonen, Esa
2005 *Analogy as Structure and Process: Approaches in Linguistics, Cognitive Psychology and Philosophy of Science.* (= Human Cognitive Processing 14). Amsterdam: Benjamins.

Jung, Ingrid
1989 *Grammatik des Paez, ein Abriss*. Doctoral dissertation, University of Osnabrück.

Kari, Ethelbert E.
2000 *Ogbronuagum (The Bụkuma Language)*. (= Languages of the World / Materials; 329). München: Lincom Europa.

Klaus, Anneliis
2009 *Eitus eesti murretes* [Negation in Estonian dialects]. MA thesis, Tartu University.

Koehn, Edward, and Sally Koehn
1986 Apalai. In *Handbook of Amazonian Languages*, vol. 1, Desmond C. Derbyshire, and Geoffrey K. Pullum (eds.), 33–127. Berlin / New York: Mouton de Gruyter.

Kotilainen, Lari
2007 *Kiellon lumo. Kieltoverbitön kieltokonstruktio ja sen kiteytyminen.* [The fascination of negation. The negative construction without negative verb and its crystallization] (= Suomi; 193). Helsinki: Suomalaisen Kirjallisuuden Seura.

Landin, David J.
1984 An outline of the syntactic structure of Karitiâna sentences. In *Estudos sobre línguas tupí do Brasil*, Robert A. Dooley (ed.), 219–254. (= Série Lingüística; 11). Brasilia: Summer Institute of Linguistics.
http://www.sil.org/americas/brasil/publcns/ling/KTStruct.pdf

Lee, D. Demetracopoulou
1946 Stylistic use of the negative in Wintu. *International Journal of American Linguistics* 12.2: 79–81.

Master, Alfred
1946 The Zero Negative in Dravidian. *Transactions of the Philological Society* 1946: 137–155.

Miestamo, Matti
2005 *Standard Negation: The Negation of Declarative Verbal Main Clauses in a Typological Perspective*. (= Empirical Approaches to Language Typology; 31). Berlin / New York: Mouton de Gruyter.

Miestamo, Matti
2007 Symmetric and asymmetric encoding of functional domains, with remarks on typological markedness. In *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations*, Matti Miestamo, and Bernhard Wälchli (eds.), 293–314. (= Trends in Linguistics, Studies and Monographs; 189). Berlin / New York: Mouton de Gruyter.

Pederson, Eric
1993 Zero Negation in South Dravidian. In *CLS 27, Papers from the 27th Regional Meeting of the Chicago Linguistic Society 1991, Part Two: The Parasession on Negation*, Lise M. Dobrin, Lynn Nichols, and Rosa M. Rodriguez (eds.), 233–245. Chicago: Chicago Linguistic Society.

Pilot-Raichoor, Christiane
1997 Le zéro négatif dans les conjugaisons dravidiennes. *Faits de langues* 10: 77–102.

Pilot-Raichoor, Christiane
2010 The Dravidian zero negative: Conceptualisation and diachronic context of its morphogenesis. In *Rara & Rarissima: Documenting the fringes of linguistic diversity*, Jan Wohlgemuth and Michael Cysouw (eds.), 267–304. (= Empirical Approaches to Language Typology; 46). Berlin / New York: Mouton de Gruyter.

Rumsey, Alan
1982 *An Intra-Sentence Grammar of Ungarinjin, North-Western Australia*. (Pacific Linguistics; B-86). Canberra: Australian National University.

Savijärvi, Ilkka
1977 *Redundanssi ja kieltoverbin ellipsi suomen kielen negaatiojärjestelmässä.* [Redundancy and the ellipsis of the negative verb in the Finnish negation system] (= Jyväskylän yliopiston suomen kielen ja viestinnän laitoksen julkaisuja; 14). Jyväskylä: Jyväskylän yliopisto.

Savijärvi, Ilkka
1981 Redundanz und Ellipse des Negationsverbs im Negationssystem der finnischen Dialekte. In *Heutige Wege der finnischen Dialektologie*, Heikki Leskinen (ed.), 157–174. (= Studia Fennica; 24). Helsinki: Suomalaisen kirjallisuuden seura.

Seiler, Walter
1985 *Imonda, A Papuan Language*. (= Pacific Linguistics; B-93). Canberra: Australian National University.

Storto, Luciana
1999 *Aspects of a Karitiana grammar*. Doctoral dissertation, Massachusetts Institute of Technology.

Subrahmanyam, P. S.
1971 *Dravidian Verb Morphology (A Comparative Study).* (Annamalai University, Department of Linguistics. Publication 24) Annamalainagar: Annamalai University.

Accounting for rare typological features in formal syntax: three strategies and some general remarks

Frederick J. Newmeyer

1 Introduction

My goal in this paper is to explore the treatment of rare typological features in formal syntax, in its mainstream Principles-and-Parameters (P & P) manifestation. I present the three principal strategies for handling them, which I illustrate with case studies, followed by an evaluation of the strategies.

Rarities present a particular challenge for P & P, given the central idea of this approach that seeming complexity and idiosyncrasy are purely epiphenomenal. As Chomsky has put it:

> So you don't have complex rules for complex constructions because there aren't any rules and there aren't any constructions. [...] Rather, there are extremely general principles [...] and then there are options that have to be fixed, parametric choices [...] Within this framework of fixed principles and options to be selected, the rules and constructions disappear, they become artifacts. (Chomsky 2002: 94)

Eliminating rules and constructions immediately raises important questions: where precisely does typology fit into this picture, and in particular, where do typological oddities fit in? If language-particular constructions and rules have disappeared, it is not at all obvious how fixed principles and options are going to derive some feature that is present only in a few languages. One might say that it is not the job of Universal Grammar (UG) to concern itself with such matters, but, in fact, Chomsky has been quite explicit that typological generalizations (and presumably the odd exception to them) do fall under the purview of Universal Grammar:

> There has also been very productive study of generalizations that are more directly observable: generalizations about the word orders we actually see, for example. The work of Greenberg has been particularly instructive and influential in this regard. *These universals are probably descriptive generalizations that should be derived from principles of UG.* (Chomsky 1988: 33; emphasis added)

The juxtaposition of those two quotes simply leaves unanswered the question of what the theory has to say about rare features or rare correlations of features. So consider the fact that only one percent of languages manifest the correlation SVO & Rel N (Dryer 1991). What should the theory do about such a case? One possibility would be for it to pay no attention to such languages, as is perhaps suggested by the first quote. After all, it seems unlikely that general principles combined with parametric choices could end up deriving that one percent of languages that manifest the correlation. On the other hand, one might attempt to derive the one percent figure from theoretical machinery already deemed necessary, as suggested by the second quote. The theoretical challenge is that if all that exists are parameters and options, then what, if anything, tells us why some of these options are frequently chosen and why some are virtually never chosen? That is, what theoretical device explains why there are rare features and how they should be accounted for?

In fact, there is no general consensus on these questions. Rare features have been handled in P & P syntax in three quite different ways, one of which, at least, seems at root incompatible with the first Chomsky quote. The three strategies are as follows:

(A) Reductionism: The rare feature is derivable from the interaction of processes known to be motivated in the grammars of the world's languages.

(B) UG Stipulation: Universal Grammar is structured so that the rare feature is predicted to be rare.

(C) Language-Particular Stipulation: Rare features are outside the domain of UG principles and parameters *per se*, and are attributed to language-particular rules.

To put the issue in "folksy" terms, two of the three options "punish" grammars for having rare features. With UG Stipulation (B), it is Universal Grammar itself that imposes a penalty on grammars containing typological rarities, by marking them as deviant from what Universal Grammar would regard as ideal. With Language-Particular Stipulation (C), the punishment is *self*-inflicted. Grammars say in effect: " I know that I'm allowing some rare feature that doesn't mesh too well with ideas about a simple and elegant UG. But I'm willing to pay the price, namely, this ugly extra rule that I carry around with me."

In this paper I argue for Reductionism (A), the non-punishment option, and try to explain how it can be made a reality without giving up too much in the way of the essential notions of current syntactic theory.

The paper is organized as follows. In Section 2, I exemplify the three strategies for handling rarities mentioned above. Section 3 compares and evaluates the three strategies, while Section 4 probes the adequacy of the idea that there is a correlation between typological rarity and grammatical complexity. Section 5 is a brief conclusion.

2 Three strategies to handle rare typological features in formal syntax

Let us now examine a few examples of each of the three options for handling rare features, followed by a brief discussion of the handling of rarities within Optimality Theory.

2.1 Reductionism

In this strategy, the rare feature is derivable from the interaction of processes known to be widespread in the world's languages. Three examples follow.

2.1.1 V2 languages

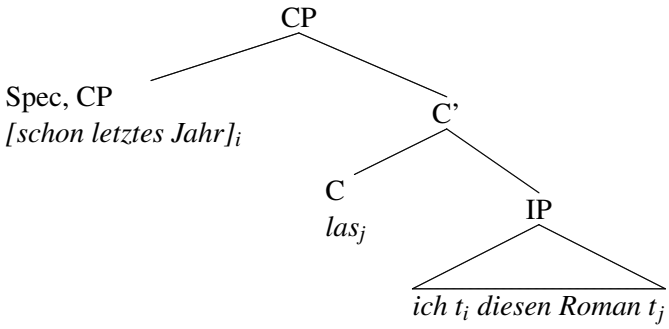
In V2 languages, the finite verb must occupy second position in the main clause. V2 is attested only in Germanic, in languages in contact with Germanic (Sorbian, Old French and some other early Romance languages, Rhaeto-Romance, Breton, and Middle Welsh), and in the Indo-Aryan language Kashmiri (Raritätenkabinett 2003: entry #79). Consider an example from German in (1a)–(1f). Note that in all six variants of the same proposition, the finite verb is in second position:

(1) (German)
   a. *[Ich] [las] [schon letztes Jahr diesen Roman]*
   b. *[Ich] [habe] [schon letztes Jahr diesen Roman gelesen]*
   c. *[Diesen Roman] [las] [ich schon letztes Jahr]*
   d. *[Diesen Roman] [habe] [ich schon letztes Jahr gelesen]*
   e. *[Schon letztes Jahr] [las] [ich diesen Roman]*

f. *[Schon letztes Jahr] [habe] [ich diesen Roman gelesen]*

One popular account of V2, initiated by den Besten, derives it from the movement of the inflected verb into C and the movement of the initial XP into CP:

(2) Analysis of (1e) (den Besten 1983; Platzack 1987)



Each movement operation has been proposed in the analysis of many diverse phenomena in a wide variety of languages. In other words, the den Besten and Platzack account provides no explanation of why V2 is rare.

2.1.2 Multiple wh-fronting

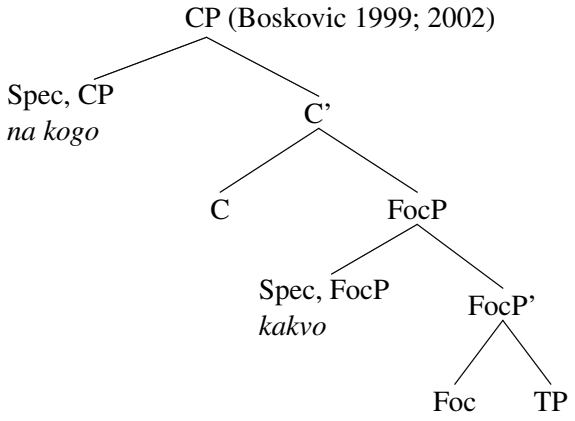
In multiple *wh*-fronting languages, more than one *wh*-phrase may appear displaced in fronted position. No cross-linguistic data is available, but this phenomenon is believed to be very rare. The following is an example from Bulgarian:

(3) Bulgarian (Boskovic 2002)

Na kogo kakvo dade Ivan
to who what gave Ivan
'What did Ivan give to who?'

Boskovic argues that multiple *wh*-fronting is not a unified phenomenon, but for Bulgarian, the leftmost *wh*-phrase moves to [Spec, CP] (that is, ordinary *wh*-movement), while the other frontings are instances of focus-movement:

(4)



Both *wh*-fronting and focus-movements are widespread cross-linguistically. However, the particular conjunction of the two leads to the rare situation found in Bulgarian.

2.1.3 Optional wh-fronting languages

In some languages, a *wh*-phrase may appear *in situ* or fronted, with no discernible discourse differences. No cross-linguistic data is available, but this phenomenon is said to be very rare. The following is an example from Babine-Witsuwit'en (Athabaskan; Canada; Denham 2000):

(5) a. *Lillian ndu yunkët*
Lillian what 3s.bought.3s
'What did Lillian buy?'

b. *Ndu Lillian yunkët*
what Lillian 3s.bought.3s
'What did Lillian buy?'

This phenomenon poses an interesting analytical problem for the Minimalist Program, in that optionality is excluded in the syntactic derivation *per se*. In the analysis of Denham (2000), Babine-Witsuwit'en has optional selection of C from the lexicon. If C is chosen, then movement is triggered; if C is not chosen, then the *wh*-phrase remains *in situ*. In other words, optional selection of C parallels other optional selection, such as that of Topic or Focus. But, crucially, Denham provides no explanation of why the optional selection of C is so rare cross-linguistically.

2.2 UG stipulation

In this strategy, Universal Grammar is structured so that the rare feature is predicted to be rare. Three examples follow.

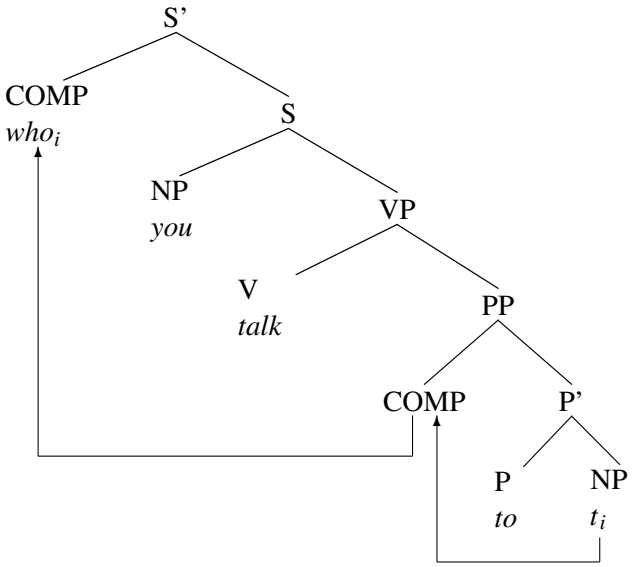
2.2.1 Preposition-stranding

Some languages permit the extraction of NP out of PP, leaving the Preposition in its base-generated position. This phenomenon is attested only in Germanic (but not in German and only marginally in Dutch), and marginally in French. In English we find:

(6) a. *Who did you speak to?*
b. *Your little boy should be spoken to.*

In one of the first analyses of stranding, Universal Grammar is said to provide PP with a COMP position as a marked option. In languages that choose that option, [COMP, PP] is an "escape hatch" for extraction, thereby licensing stranded P (van Riemsdijk 1978):

(7)



In other words, Universal Grammar is structured so that stranding is predicted to be rare.

2.2.2 The order A–N–Dem–Num

This particular order of elements is found only in Koiari (Trans-New Guinea; Papua New Guinea) and Bai (Tibeto-Burman; China) (Cinque (2005), citing Dutton (1996) for the former and Dryer (2008) for the latter). Cinque builds into Universal Grammar a fixed Merge (or underlying) order of the four categories, along with parameters of movement which are assigned varying markedness values. The derivation of A–N–Dem–Num involves the Merge order in (8a) followed by the two parametrically marked movements in (8b) and (8c):

(8) a. Merge order: [... $[_{WP}$ Dem ... $[_{XP}$ Num ... $[_{YP}$ A $[_{NP}$ N]]]]]
b. Movement of NP plus pied-piping of the *picture of who* type of the lowest modifier (highly marked), followed by
c. Raising of [A N] without pied-piping around both Num and Dem (highly marked)

In other words, the order is rare because Universal Grammar simply happens to be structured in a way that stipulates the rarity.

2.2.3 Rare correlations of syntactic properties

There has always been the sentiment that Universal Grammar prefers "consistency". That is, for any particular language, Universal Grammar specifies that complements either precede or follow their heads, specifiers either precede or follow their heads, and so on. For that reason, theories of Universal Grammar have generally been devised so that their most unmarked settings generate typological consistency. But few if any languages are completely consistent. As observed by Smith (1981: 39),

> "there is virtually no typological implicational statement that does not have exceptions".

So in different ways in different periods in the history of generative grammar, there have been attempts to construct a theory of Universal Grammar so that consistency emerges as the simplest unmarked option and that inconsistency and the putative rarity that goes along with it requires some deviation from what Universal Grammar might "prefer".

The idea in generative syntax goes at least as far back as 1965. Bach (1965) noted that the correlations in (9) can be derived from Greenberg (1963):

(9) a. SVO & N Rel ordering
b. SOV & Rel N ordering

Bach proposed that Universal Grammar specify (9) as the unmarked correlation. The grammars of the minority of languages like Chinese that violate (9a) and (9b) would require special complicating statements overriding the marking conventions.

To take a related example, Travis (1989) noted eight possible orders of V, DO, a complement PP, and an adjunct PP:

(10) a. PP_2 PP_1 NP V
b. PP_2 PP_1 V NP
c. PP_2 NP V PP_1
d. PP_2 V NP PP_1
e. PP_1 NP V PP_2
f. PP_1 V NP PP_2
g. NP V PP_1 PP_2 (Kpelle)
h. V NP PP_1 PP_2

Kpelle (Niger-Congo; Guinea) has order (10g), which was claimed by Travis to be very rare. She proposed three separate parameters to derive all of (10a–10h):

(11) a. Headedness
b. Directionality of theta-role assignment
c. Directionality of Case assignment

Travis went on to propose markedness relations among these three parameters so that the rare ordering in Kpelle would be highly marked.

2.3 Language-particular stipulation

In this strategy, rare features are outside the domain of Universal Grammar principles and parameters *per se* and are attributed to language-particular rules. Three examples follow.

2.3.1 Preposition stranding (again)

Probably the most popular analysis of stranding attributes it to a language-particular rule of reanalysis. Specifically, a preposition and an adjacent verb are reanalyzed, creating a complex verb that can properly govern the trace of movement, thereby licensing stranding of the preposition:

(12) a. *You talked* $[_{PP}$ *to who*$]$ > *You* $[_{V}$ *talked to*$]$ *who* > *Who*$_i$ *did you* $[_{V}$ *talk to*$]$ e_i*?*

b. *e was spoken* $[_{PP}$ *to Mary*$]$ > *e was* $[_{V}$ *spoken to*$]$ *Mary* > *Mary*$_i$ *was* $[_{V}$ *spoken to*$]$ e_i

In other words, the few languages that manifest stranding "pay" for their rarity by adding this extra-parametric language-particular rule.

2.3.2 OVS order

This order is very rare, manifested by only 5 languages out of 402 (1.24 %) in the sample of Tomlin (1986). Consider the treatment of Hixkaryana (Carib; Brazil) in Baker (2001), based on an earlier proposal in Kayne (1994). This language for the most part has OVS word order:

(13) Hixkaryana (Derbyshire 1985)

kanawa yano toto
canoe took person
'the man took the canoe'

One's first thought might be that what is needed is a parameter allowing for OVS order. But in fact Baker (2001: 166) rejects the idea that a special word order parameter is involved here. Rather, he argues that Hixkaryana is (parametrically) SOV and allows the fronting of VP by a movement rule:

(14) S[OV] $\rightarrow$ [OV]S

In other words, in this account word order is determined both by a parameter and a language-specific rule.

2.3.3 Long-distance agreement

Long-distance agreement is agreement between a verb and an argument in a clause embedded under that verb. This phenomenon is apparently quite rare,

as agreement is typically a purely local phenomenon. Below is an example from Tsez (North Caucasian; Russian Federation):

(15) Tsez (Polinsky and Potsdam 2001)

enir [užā magalu bāc̄ruɬi] b-iyxo
mother [boy **bread.3.ABS** ate] **3**-know
'The mother **knows** [the boy ate **the bread**]

Note that the verb of the higher clause agrees with the object in the lower clause. Polinsky and Potsdam revise the theory of agreement to relax the strictly local domain for this process. But crucially, they posit that Tsez has a language-particular feature, namely that the embedded clause has an initial Topic Phrase and there is movement of the object to this position, where it undergoes agreement with the higher verb.

2.4 Rare features in Optimality Theory

This discussion would not be complete without a look at the handling of rare features within Optimality Theory (OT). It is basically a version of (A), the reductionist approach, with an element of (C), language-particular stipulation, thrown in. In Optimality Theory, the constraints are posited to be universal. What is language-particular is their ranking.

Consider the Optimality Theory analysis of subject choice in Aissen (1999). Different languages have different criteria on voice and subjecthood, as sketched in (16):

(16) Subject choice in five language types:

a. Fox (Algic; United States) and Nocte (Sino-Tibetan; India): All sentences are active, with grammatical function entirely determined by semantic role

b. English: A patient can be a subject if it is thematically prominent (e. g. the passive)

c. Lushootseed (Salishan; United States): Like English, but passive clauses with first and second person agents are excluded

d. Lummi (Salishan; United States and Canada): Like Lushootseed, but active is disallowed when the subject is third person and the object is first or second.

e. Squamish (Salishan; Canada): Active clauses with third person subjects are excluded if the object is second person

Example (17a–17e) sketches Aissen's (1999) analysis.

(17) The constraint rankings in for the five language types:

a. Fox and Nocte: ***Su / Pat*** » … {*GR / Pers, *GR / χ}

b. English: *Su / x » ***Su / Pat*** » *GR / Pers

c. Lushootseed: *Obl / Local » ***Su / Pat*** » *GR / Pers

d. Lummi: *Obl / Local » *Oj / Local » *Su / x » ***Su / Pat*** » *GR / Pers

e. Squamish: *Obl / Local » *Oj / 2 » *Su / x » ***Su / Pat*** » *Oj / 1 » *GR / Pers

As we progress from Fox to Squamish, we see increasing sensitivity of voice to person, that is, the increasing domination of the constraint ruling out subject patients by constraints ruling out associations of grammatical functions with person. Now, not all of the five types are equally common cross-linguistically. I believe the Squamish type to be quite rare. But nothing inherent to Optimality Theory tells us that, so the rarity of the Squamish type is not accounted for internally to the grammar.

3 A comparison and evaluation of the three strategies for capturing rare features

Section 3.1 argues that Strategy (B), UG Stipulation, is inherently undesirable, and Section 3.2 argues for a performance-based account of typological rarities.

3.1 Against UG stipulation to handle rare features

Let us now move to a critical comparison and evaluation of these three approaches. In my view, Strategy (B), UG Stipulation, is a priori the least desirable of the three, in that it complicates the theory of Universal Grammar, without at the same time leading to new insight about grammar. Saying that preposition stranding is rare because the theory says that a COMP node for PP is rare is not much of an explanation, since it simply substitutes one mystery for another. That is, why should a COMP node for PP be a rarity?

Furthermore, a claim that says something like "Universal Grammar disfavors COMP's for PP" runs into two immediate problems. First, the theory would be no more complex if the situation were the precise opposite, that is, if a COMP node for PP were the norm and a COMP node for full sentences were a rarity. Second, consider what it means to say that some feature is part of Universal Grammar. By definition, it means that that feature is innate. Now, we are just talking about intuitive plausibility here, but it seems utterly implausible to me that biological evolution would have shaped the human genome to favor complementizer nodes for some phrasal categories more than for others.

So let us put UG Stipulation aside for now and consider Reductionism (Strategy A) and Language-Particular Stipulation (Strategy C). In my view, each has its pluses and minuses. Reductionism has the advantage of being most in accord with the spirit of Principles-and-Parameters (P & P), in that it involves just principles and options, without language-particular rules and constructions. Pursuing this option a little further, however, it also has seeming defect (and I stress "seeming") of failing to account for the rarity of the feature. Why, for example, should the conjunction of two common processes yield a rare feature? If we can point to many cases where an inflected verb moves into C and many cases where an initial phrase moves into [Spec, CP], then why should *both* movements, which is what is needed to derive a V2 language, be uncommon?[1]

Strategy (C), Language-Particular Stipulation, is a move away from a "pure" Principles-and-Parameters theory toward a construction-based approach. After all, what could "language-particular stipulation" mean except to say that rules and / or constructions have crept back into the theory? And that is exactly what is meant by the claim, for example, that English and other stranding languages have a special reanalysis rule, that Hixkaryana and other OVS languages have a special VP fronting rule, and so on. Ideally, we would not want a theory to allow for *both* parameters *and* rules. Yet I find the idea that grammars "pay" for rare features by needing extra rules to be intuitively appealing.

3.2 A performance-based account of typological rarity

What should we conclude, then? I suggest that of the three strategies, Reductionism should be appealed to whenever possible, precisely because of its very nature. If we can derive V2 languages by means of the interaction of

processes already motivated in the grammar, then so much the better. This strategy fails to explain *why* the feature is rare, but that would be a bad thing only if it were the job of Universal Grammar to account for why a particular grammatical feature happens to be common or rare. As I have argued at length elsewhere, that is not the job of a UG-based theory. My position is that most typological generalizations fall out from a theory of performance, rather than from a theory of Universal Grammar (see Hawkins 2004; Newmeyer 2005).

Let me turn to a rare phenomenon to illustrate the advantages of a performance-based account. It turns out that there is a robust hierarchy pertaining to the material that can intervene between P and N:

(18) Prepositional Noun Modifier Hierarchy (PrNMH; Hawkins 1983)

If a language is prepositional, then if RelN then GenN, if GenN then AdjN, and if AdjN then DemN.

In a nutshell, if a language allows long things to intervene between a preposition and its object, then it allows short things. This hierarchy predicts the possibility of prepositional phrases with the structures depicted in (19) (along with an exemplifying language):

(19) a. $_{PP}$[P $_{NP}$[____N...] (Arabic, Thai)

b. $_{PP}$[P $_{NP}$[____N...]; $_{PP}$[P $_{NP}$[Dem N...] (Masai, Spanish)

c. $_{PP}$[P $_{NP}$[____N...]; $_{PP}$[P $_{NP}$[Dem N...]; $_{PP}$[P $_{NP}$[Adj N...] (Greek, Maya)

d. $_{PP}$[P $_{NP}$[____N...]; $_{PP}$[P $_{NP}$[Dem N...]; $_{PP}$[P $_{NP}$[Adj N...]; $_{PP}$[P $_{NP}$[PossP N...] (Maung, English)

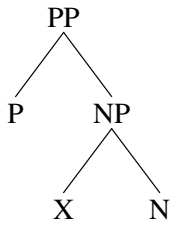
e. $_{PP}$[P $_{NP}$[____N...]; $_{PP}$[P $_{NP}$[Dem N...]; $_{PP}$[P $_{NP}$[Adj N...]; $_{PP}$[P $_{NP}$[PossP N...]; $_{PP}$[P $_{NP}$[Rel N...] (Amharic)

But virtually no language allows, say, a relative clause to intercede between a preposition and its noun complement, but not an adjective. The parsing-based explanation of the hierarchy is straightforward and invokes the processing principle in (20):

(20) Minimize Domains (Hawkins 2004): The hearer (and therefore the parsing mechanism) prefers orderings of elements that lead to the most rapid recognition possible of the structure of the sentence.

To illustrate, note that the greater the distance between the P and the N in a structure like (21), the longer it takes to recognize all the constituents of the PP:

(21)



Given Hawkins's idea that grammars try to reduce the recognition time, the hierarchy follows. Since relative clauses tend to be longer than possessive phrases, which tend to be longer than adjectives, which tend to be longer than demonstratives, which are always longer than "silence", the hierarchy is predicted on parsing grounds. How could one capture this generalization by means of any grammar-internal mechanism, short of building the entire hierarchy into Universal Grammar?

In fact, Hawkins notes that there are a few rare exceptions to this hierarchy.[2] In Karen (Sino-Tibetan; Myanmar), genitives are the only daughters of NP to precede N and in Sango (Creole; Central African Republic), Adj-N cooccurs with N-Dem. Do these cases refute the hierarchy and its performance-based explanation? The answer is "no." One's first thought might be that exceptions pose as great a challenge for parsing principles as for Universal Grammar principles. After all, in both cases, some theory-based generalization has been violated. But the difference is that one expects performance principles to admit exceptions. Rather than being like the either-or (or yes-no) switch settings inherent to Universal Grammar parameters, they are part-and-parcel of a theory of language use. And nobody, as far as I know, believes that an algebraic theory suffices to explain facts about language use. Rather, usage-based generalizations are generalizations about populations (whether of speakers or of languages). To give an analogy, the generalization that cigarette smoking causes lung cancer is not threatened by the fact that there exist individuals who smoke five packs of cigarettes per day over their lifetimes and do not develop lung cancer. The rare languages that violate this hierarchy are parallel, in crucial respects, to these individuals.

There is another, very different, reason to prefer the Reductionism strategy. Of the three, it is the most conservatively cautious. Surely not all rare features are rare by virtue of creating parsing difficulties. Some are rare by

historical accident. Reductionism is the only option of the three that does not take a grammar-internal explanation of the rarity as the default.

4 The Rarity-Complexity Hypothesis

Let us turn now to the strategy of Language-Particular Stipulation. Clearly, it is a theoretically weaker and less interesting approach than Reductionism, in that there is no presumption that the rules involved generalize beyond the particular language for which they are posited. Still, there is a way that Language-Particular Stipulation could be made more appealing. Suppose the hypothesis in (22) were theoretically motivated:

(22) The Rarity-Complexity Hypothesis (RCH): The rarer the grammatical phenomenon, the more complex its analysis.

The cases that we looked at in Section 2.3 seem to give some support to this hypothesis. Preposition stranding, OVS order, and long-distance agreement are all rare and more complex to formulate than their more typologically common counterparts. Section 4.1 provides further support for the hypothesis, while Section 4.2 argues that it is unmotivated as a general principle.

4.1 Support for the Rarity-Complexity Hypothesis

In every historical period, claims have been made that typological rarity is "registered" internally to the grammar. That is, cross-linguistically rare features are said to require more complex descriptive devices or supplementary statements not needed for common ones. For example, Emonds (1980) took on the question of why VSO languages are much rarer than SVO languages. His answer was that VSO order is literally derived from SVO order, so VSO languages contain an extra rule in their derivation:

> VSO languages are in fact rare compared even with the SVO type alone. Deriving them via a local rule predicts this [. . .] That is, certain types of rules make a language "more complicated" and hence, rarer. (Emonds 1980: 44)

Consider Baker's Government-and-Binding-style account of why certain typological features are more common than others. Essentially, the more "choices" a language learner needs to make, the rarer the language type is claimed to be. As far as VO versus OV is concerned:

> Since the difference between English-style and Japanese-style word order is attributable to a single parameter [...] we expect roughly equal numbers of English-type and Japanese-type languages. (Baker 2001: 134)

Why are VSO languages so much rarer than SVO languages, then? Because two more parameters enter into the characterization of the former than of the latter:

> Within the head-initial languages, however, it requires two further decisions to get a verb-initial, Welsh-type language: Subjects must be added early and tense auxiliaries must host verbs. If either of these decisions is made in the opposite way, then subject-verb-object order will still emerge. If the decisions were made by coin flips, we would predict that about 25 percent of the head-initial languages would be of the Welsh type and 75 percent of the English type. This too is approximately correct [...] (Baker 2001: 134)

There are post-Government-and-Binding approaches where it is hypothesized that, all other things being equal, the more movements required, the typologically rarer the language type. So take Kayne's asymmetric program (Kayne 1994). There exist rare "typologically inconsistent" languages such as Vata (Niger-Congo; Côte d'Ivoire) which have final complementizers, but in which *wh*-elements move to the left. Kayne gives a rough sketch of how such languages have more complicated grammars than those with the expected correlation. In essence, for most SOV languages, IP moves leftward into [Spec,CP], thereby blocking *Wh*-Movement into that position. Rare languages like Vata, which have final complementizers, but *Wh*-Movement to the left, have an extra rule.

Cinque (1996) also attempts to explain typological generalizations, and exceptions to them, in Kayne's framework. For example, Kayne predicts that no language will have N-Dem and Num-N. But some languages, including Berber, Welsh, Hebrew, and Zapotec, do manifest this correlation. According to Cinque, such languages "pay the price" for their deviance by requiring an extra movement of demonstratives.

4.2 The Rarity-Complexity Hypothesis is incorrect

Is the Rarity-Complexity Hypothesis always motivated? That is, is there in general a robust correlation between the number of operations applying in the derivation of a particular construction type and its cross-linguistic rarity?

I do not think so. Stylistic variants and speech act types are a natural point of departure. The grammars of all languages provide the speaker with the possibility of alternative means of expressing the same propositional content (consider passives, clefts, pseudo-clefts, and so on). Likewise, the grammars of all languages provide means for questioning and negating constituents and for conveying imperatives, hortatives, and other speech acts. With rare exceptions, it has been assumed that for any given language, the least complex derivation is that of the simple active declarative sentence. Stylistic variants and so on have typically been formed by means of a rule or projection not needed for the simple active declarative sentence. In other words, the derivations of such sentence types are usually assumed to be more complicated than the derivations of simple active declarative sentences. But are such sentence types rarer? It depends on how one defines "rarity", but in general there is little reason to think so. Every language allows stylistic variants, has a means of forming questions, of negating constituents, and so on. So it seems unlikely that the Rarity-Complexity Hypothesis holds for such sentence types.

The Rarity-Complexity Hypothesis would also seem to suggest, all other things being equal, that an analysis not involving displacement should be more common cross-linguistically than one involving displacement, at least if we confine ourselves to overt displacement. In other words, questions formed by overt movement should be rarer than in situ questions. Actually, that seems right. Dryer (1991) has observed that a solid majority of the world's languages lack a *wh*-fronting-type rule. But by the same token, a non-movement option should always be more common cross-linguistically than a movement option, regardless of the particular phenomenon. That seems dubious. A lot depends on one's particular theoretical approach, of course, but I have no reason to think that passives, say, that involve only a change in case-marking are more common cross-linguistically than those that involve a change of case-marking along with displacement. Along the same lines, displaced focuses are certainly more common than in situ focuses.

Let us turn now to some more specific phenomena that seem to call into question the Rarity-Complexity Hypothesis. Given space constraints, I focus on one: clausal negation. Miestamo (2005) classifies negation strategies as in Roman I and II below (it must be kept in mind that many languages employ more than one strategy):

[I] Symmetric negation, in which a negative marker occurs without any other modification of the clause.

(23) Indonesian

a. *Mereka menolong kami*
they help us.EXCL
'They helped us'

b. *Mereka tidak menolong kami*
they NEG help us.EXCL
'They didn't help us'

[II] Asymmetric negation, in which the placement of the negative marker involves one of the following types of clausal modification:

i. A loss of finiteness of the lexical verb:

(24) Hixkaryana

a. *kɨ-amryekɨ-no*
1SUBJ-hunt-IMPST
'I went hunting'

b. *amryekɨ-hɨra w-ah-ko*
hunt-NEG 1SUBJ-be-IMPST
'I did not go hunting'

ii. The appearance of an irrealis marker of some sort:

(25) Maung (Iwajdjan; Australia)

a. *ŋi-udba*
1SG.3-put
'I put'

b. *marig ŋi-udba-nji*
NEG 1SG.3-put-IRR.PST
'I do / shall not put'

iii. The appearance of an emphasis marker of some sort:

(26) Abipon (Mataco-Guaicuru; Argentina)

a. *i-arai-k-am*
3-know-OBJ-FUT
'S / he will know it'

b. *cig-at* *i-arai-k-am*
NEG-EMPH 3-know-OBJ-FUT
'S / he will not know it'

iv. A categorial change of some element in the clause:

(27) Tera (Afro-Asiatic; Nigeria)

a. *ali wà masa koro*
Ali PFV buy donkey
'Ali bought a donkey'

b. *ali nə̀* *masa goro ɓa*
Ali (NEG).PFV buy kola NEG
'Ali didn't buy kola'

The typological breakdown of the two major negation types is shown in Table 1:

Table 1. Number of languages in the sample of Miestamo (2005) manifesting the two major types of negation

negation type	number
Symmetric only	72
Both Symmetric and Asymmetric	76
Asymmetric only	31

In other words, 60 % of the languages in the 179-language sample manifest at least some asymmetric negation. Yet by any measure that I have been able to devise, symmetric negation is by far the simplest to describe in formal terms. Take English, which manifests symmetric negation, as (28a)–(28b) illustrate:

(28) a. *Mary has left.*
b. *Mary has not left.*

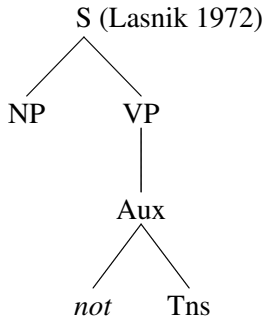
It is true that things look complicated for English due to the presence of supportive *do*, as in (29b):

(29) a. *Mary left.*
b. *Mary did not leave.*

But the unexpected appearance of *do* is widely assumed to be a consequence of quirky facts about the English tense-aspect system, rather than about negation *per se*.

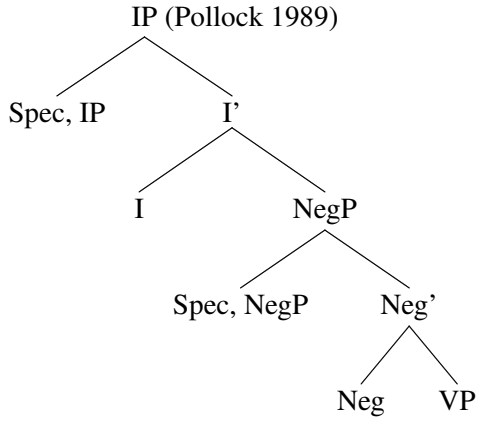
In every period of generative grammar, the analysis of English sentential negation has involved very little theoretical machinery. In early transformational grammar, the word *not*, or an abstract negative marker realized as *not*, were assumed to be inserted into the auxiliary or base-generated inside it or generated in sentence-initial position and moved to the right. (30) illustrates base-generation inside the auxiliary:

(30)



More recently, the general assumption has been that *not* is the head of Negation Phrase, as in (31):

(31)



Again, this is a very simple analysis of symmetric negation. But when we look at analyses of asymmetric negation over the years, we see considerably more complication. For example, take Palauan (Austronesian; Palau), which

is a language in which the negative morpheme is associated with an irrealis marker on the verb, as in (32a)–(32b):

(32) Palauan (Foley 1975)

a. *Juan a məŋa ra ŋikəl*
John is eating of fish
'John is eating some of the fish'

b. *Juan a diak loŋa ra ŋikəl*
John is NEG 3.IRR.eat of fish
'John isn't eating any of the fish'

Foley's analysis involves not just the presence of the negative marker, but also complex rules to insure the correct form of the verb in the negative sentence.

Turning to P & P approaches to negation, we can see the greater descriptive and analytical complexity of asymmetric negation over symmetric negation by looking at Korean, which has both: symmetric negation, as in (33b), and asymmetric negation, as in (33c):

(33) Korean (Miestamo 2003; based on Chang 1996)

a. *yong-un mayil tv-lul po-n-ta*
Yong-TOP every.day TV-OBJ see-PRES-DECL.PLAIN
'Yong watches TV every day'

b. *yong-un tv-lul an pwa-yo (symmetric)*
Yong-TOP TV-OBJ NEG see-POL
'Yong doesn't watch TV'

c. *yong-i tv-lul po-ci an-ha-yo (asymmetric)*
Yong-SUBJ TV-OBJ see-SUSP NEG-AUX-POL
'Yong doesn't watch TV'

In (33c), the negative marker *an-* is attached to the all-purpose auxiliary verb *ha-*, which is the finite element of the negative clause, while the lexical verb is marked with the suspective *-ci*. In every generative analysis of which I am aware, Korean asymmetric negation has a more complex analysis than symmetric, as indicated in (34a)–(34b):

(34) a. Symmetric (Ahn and Yoon 1989; Ahn 1991; Hagstrom 1996):
A NegPhrase projection and successive head movements

b. Asymmetric (Ahn 1991; Cho 1994; Hagstrom 1996):
A NegPhrase projection, successive head movements, and a rule resembling English *Do*-Support

In sum, the Rarity-Complexity Hypothesis is not borne out for negation.

Again, the antisymmetric program of Kayne (1994) makes explicit typological predictions. The centerpiece of this program is that all languages are underlyingly Specifier-Head-Complement. OV languages require a movement operation unnecessary for VO languages, that is, one that preposes the object over the verb. Given the Rarity-Complexity Hypothesis, then, VO languages should be more common than OV languages. However, that is not the case. Simply counting languages, VO and OV are about equally common. However, counting "genera" (subfamilies of roughly equal time-depth), OV order is considerably more widespread than VO order. Dryer (1989) has determined that in 111 genera (or 58% of the total), OV order predominates (see Table 2).

Table 2. Breakdown of genera in terms of basic word order, by area (Dryer 1989)

	Africa	Eurasia	Australia & New Guinea	North America	South America	Total
SOV	22	26	19	26	18	111
SVO	21	19	6	6	5	57
VSO	5	3	0	12	2	22

So whether one counts languages or genera, there is no support for the idea that cross-linguistically there is a tendency for verbs to precede objects. Consider some other findings from Dryer (1992). It turns out that in 119 genera out of 196, postpositionality predominates. In 76 out of 127, the predicate precedes the copula and in 78 out of 128 the adverb precedes the verb. All of these facts challenge the conjunction of the antisymmetric program and the Rarity-Complexity Hypothesis.

There is one interpretation of the Rarity-Complexity Hypothesis, however, where things seem to work a little better than what we have just seen. Under this interpretation, the Rarity-Complexity Hypothesis would be evaluated with respect to token frequency instead of type frequency. In corpus studies, actives are more frequently attested than passives, declaratives

than questions, affirmatives than negatives, and so on. So there does seem to be at least a rough inverse correlation between the number of operations applying in a particular derivation and token frequency of the sentence type derived by those operations. It seems unlikely, however, that many advocates of the Rarity-Complexity Hypothesis would take comfort in this fact, since most formal syntacticians have (quite rightly in my opinion) resisted drawing conclusions about linguistic competence from facts about text frequency.

Should the general failure of the Rarity-Complexity Hypothesis be seen as a defect of the theory? I would say "no". I have argued in Newmeyer (2005) that typological evidence is not a reasonable heuristic for grammar construction, either directly or indirectly. That is, there is no set of principles or parameters internal to grammatical theory from which typological generalizations can be derived. It is not surprising, therefore, that the Rarity-Complexity Hypothesis does not work very well. Upon reflection, it is easy to see why generative theory cannot and should not provide a theory of typology. Generative grammar provides a theory of mental representations. But typological generalizations are frequency effects derived from the use made of these mental representations. Grammatical generalizations and evidence based on typology belong to two different domains. Typological generalizations are generalizations about the distribution of possible surface configurations. But Universal Grammar is not a theory of surface configurations in any sense. Rather, it is a theory of abstract grammatical structure and the operations that can be performed on that structure. We have no more reason to think that a theory of Universal Grammar should tell us, say, why there are more SVO languages than VSO languages in the world than that it should tell us why some languages have more honorific expressions than others or why some languages contain more borrowed lexical items than others. It is therefore not surprising that the Rarity-Complexity Hypothesis, which crucially depends on a UG-based approach to typological generalizations, is incorrect.

5 Conclusion

Rare grammatical features present a special problem for principles-and-parameters syntax, given the leading idea that complexity and, by implication, rarity are purely epiphenomenal. We have looked at three strategies that have

been called on to handle rare features: Reductionism, in which the rare feature is derivable from the interaction of processes known to be widespread in the world's languages; UG-Stipulation, in which Universal Grammar is structured so that the rare feature is predicted to be rare; and Language-Particular Stipulation, in which rare features are outside the domain of Universal Grammar principles and parameters *per se* and are attributed to language-particular rules. UG-Stipulation is the least attractive of the three strategies, while both Reductionism and the Language-Particular Stipulation have a certain amount of appeal. The latter depends on the correctness of the Rarity-Complexity Hypothesis (RCH), which holds that the rarer the grammatical phenomenon, the more complex its analysis. Since the Rarity-Complexity Hypothesis is not motivated, we are led to conclude that Reductionism is the most desirable of the three hypotheses.

Abbreviations

1SG = first person singular; 1SUBJ = first person subject; 3 = third person; 3.ABS = third person absolutive; 3.IRR = third person irrealis; 3s = third person singular; DECL = declarative; EMPH = emphasis; EXCL = exclusive; FUT = future; IMPST = immediate past; IRR = irrealis; NEG = negative marker; OBJ = object; PFV = perfective; PLAIN = plain (level of politeness); POL = polite; PRES = present; PST = past; TOP = topic

Acknowledgments

I would like to thank Jan Wohlgemuth and an anonymous referee for their helpful comments.

Notes

1. But see Harris (this volume: 91–103) for the opinion that compounded commonalities should lead to a rarity.
2. An anonymous referee points out that some of the "Type B" languages of Heine (1976) are (partial) exceptions to the hierarchy. These languages have possessives before the head noun, but adjectives following it. Most of these languages are postpositional, but may have a few prepositions (leading to hierarchy violations).

References

Ahn, Hee-Don
1991 *Light verbs, V-movement, negation, and clausal architecture in Korean and English*. Doctoral dissertation, Department of Linguistics, University of Wisconsin-Madison.

Ahn, Hee-Don and Hang-Jin Yoon
1989 Functional categories in Korean. *Harvard Studies in Korean Linguistics* 3: 79–88.

Aissen, Judith
1999 Markedness and subject choice in optimality theory. *Natural Language and Linguistic Theory* 17: 673–711.

Bach, Emmon
1965 On some recurrent types of transformations. In *Approaches to linguistic analysis, language and society, teaching language skills* (Monograph Series on Languages and Linguistics No. 18; 16th Round Table Meeting), Charles W. Kreidler (ed.), 3–18. Washington: Georgetown University Press.

Baker, Mark C.
2001 *The atoms of language: The mind's hidden rules of grammar*. New York: Basic Books.

Boskovic, Zeljko
1999 On multiple feature checking: Multiple Wh-fronting and multiple head movement. In *Working minimalism*, Samuel David Epstein and Norbert Hornstein (eds.), 159–187. Cambridge, MA: MIT Press.

Boskovic, Zeljko
2002 On multiple Wh-fronting. *Linguistic Inquiry* 33: 351–383.

Chang, Suk-Jin
1996 *Korean*. Amsterdam: Benjamins.

Cho, Dong-In
1994 Functional projections and verb movement. In *Theoretical issues in Korean linguistics*, Young-Key Kim-Renaud (ed.), 233–254. Stanford: CSLI Publications.

Chomsky, Noam
1988 *Language and problems of knowledge: The Managua lectures*. Cambridge, MA: MIT Press.

Chomsky, Noam
2002 *On nature and language*. Cambridge: Cambridge University Press.

Cinque, Guglielmo
1996 The 'antisymmetric' program: Theoretical and typological implications. *Journal of Linguistics* 32: 447–465.

Cinque, Guglielmo
2005 Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry* 36: 315–332.

den Besten, Hans
1983 On the interaction of root transformations and lexical deletion rules. In *On the formal syntax of the Westgermania*, Werner Abraham (ed.). Amsterdam: Benjamins.

Denham, Kristin
2000 Optional Wh-Movement in Babine-Witsuwit'en. *Natural Language and Linguistic Theory* 18: 199–251.

Derbyshire, Desmond C.
1985 *Hixkaryana and linguistic typology*. Arlington, TX: Summer Institute of Linguistics.

Dryer, Matthew S.
1989 Large linguistic areas and language sampling. *Studies in Language* 13: 257–292.

Dryer, Matthew S.
1991 SVO languages and the OV:VO typology. *Journal of Linguistics* 27: 443–482.

Dryer, Matthew S.
1992 The Greenbergian word order correlations. *Language* 68: 81–138.

Dryer, Matthew S.
2008 Word order in Tibeto-Burman languages. *Linguistics of the Tibeto- Burman Area* 31: 1–88.

Dutton, T. E.
1996 *Koiari*. Munich: Lincom Europa.

Emonds, Joseph E.
1980 Word order in generative grammar. *Journal of Linguistic Research* 1: 33–54.

Foley, William A.
1975 How to say 'no' in Palauan. *Berkeley Linguistic Society* 1: 132–146.

Greenberg, Joseph H.
1963 Some universals of grammar with special reference to the order of meaningful elements. In *Universals of language*, Joseph Greenberg (ed.), 73–113. Cambridge, MA: MIT Press.

Hagstrom, Paul
1996 Do-support in Korean: Evidence for an interpretive morphology. In *Morphosyntax in generative grammar*, Hee-Don Ahn, Myung-Yoon Kang, Yong-Suck Kim and Sookhee Lee (eds.), 169–180. Seoul: Hankwuk Publishing Company.

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.

Hawkins, John A.
1983 *Word order universals*. New York: Academic Press.

Hawkins, John A.
2004 *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Heine, Bernd
1976 *A typology of African languages*. Berlin: Dietrich Reimer.
Kayne, Richard S.
1994 *The antisymmetry of syntax*. Cambridge, MA: MIT Press.
Lasnik, Howard
1972 *Analyses of negation in English*. Doctoral dissertation, Department of Linguistics, Massachusetts Institute of Technology, Boston.
Miestamo, Matti
2003 *Clausal negation: A typological study*. Doctoral dissertation, Department of Linguistics, University of Helsinki.
Miestamo, Matti
2005 *Standard negation: The negation of verbal main clauses in a typological perspective*. Berlin: Mouton de Gruyter.
Newmeyer, Frederick J.
2005 *Possible and probable languages: A generative perspective on linguistic typology*. Oxford: Oxford University Press.
Platzack, Christer
1987 The Scandinavian languages and the null subject parameter. *Natural Language and Linguistic Theory* 5: 377–401.
Polinsky, Maria and Eric Potsdam
2001 Long-distance agreement and topic in Tsez. *Natural Language and Linguistic Theory* 19: 583–646.
Pollock, Jean-Yves
1989 Verb movement, universal grammar, and the structure of IP. *Linguistic Inquiry* 20: 365–424.
Raritätenkabinett
2003 http://ling.uni-konstanz.de:591/universals/searchrara.html.
Smith, Neil
1981 Consistency, markedness and language change: On the notion 'consistent language'. *Journal of Linguistics* 17: 39–54.
Tomlin, Russell S.
1986 *Basic word order: Functional principles*. London: Croom Helm.
Travis, Lisa
1989 Parameters of phrase structure. In *Alternative conceptions of phrase structure*, Mark R. Baltin and Anthony S. Kroch (eds.), 263–279. Chicago: University of Chicago Press.
van Riemsdijk, Henk
1978 *A case study in syntactic markedness: The binding nature of prepositional phrases*. Dordrecht: Foris.

Rara and grammatical theory

Jan Rijkhoff

1 Introduction

This paper shows that rare typological features have lead to a better understanding of parallels between the structure of the noun phrase and the structure of the clause. The paper is organized as follows. Section 2 ('Preliminaries') deals with the relevance of typological data for grammatical theory and briefly discusses the only sampling method that explicitly aims at including *rara* and *rarissima*. Section 3 demonstrates that the further development of a symmetrical model of the structure of the noun phrase and the clause depended on the existence of two rare grammatical phenomena. The more general point this paper wants to make is that *rara* and *rarissima* can play a crucial role in the validation of theoretical claims and serve as a heuristics for an extension of grammatical theory.

2 Preliminaries

Rare linguistic features should play in important role in grammatical theory, if only because a theory that can account for both common and unusual grammatical phenomena is superior to a theory that can only handle common linguistic properties. As was already noted by Perkins (1988: 367), "[. . .] exceptional types test the theory". In reality, however, very few grammatical theories can be said to strive for typological adequacy (Butler 2003: 200–201, 246). Some notable exceptions are Dik's *Functional Grammar* (Dik 1997) and its successor *Functional Discourse Grammar* (Hengeveld and Mackenzie 2008), which systematically take into account data from a wide range of typologically, genetically and geographically different languages. These data are often collected using a sampling procedure that is designed to produce *variety samples*, i. e. samples that are representative for the linguistic diversity across the globe. This sampling method produces samples in which the genetic (historic) distance between languages is always maximal and in which languages are distributed proportionally across families and subfami-

lies on the basis of linguistic variety rather than quantitative considerations (such as the number of languages that belong to a certain family or subfamily). Consequently, the chance of attesting *rara* or *rarissima* is significantly higher in a variety sample than in a probability sample or a random sample (Rijkhoff et al. 1993; Rijkhoff and Bakker 1998).

For grammatical theories that strive for typological adequacy, frequency of occurrence is relatively unimportant. Since a general theory of grammar should be able to deal with facts from any natural human language, all grammatical properties are relevant. What counts most is the fact that some property (a sound, a meaning, a form, a construction, etc.) is attested in one of the world's languages, not so much the number of languages that happen to share some property, as this may be due to a historical accident.

Since variety sampling forces the linguist to systematically cover all the language families in a top-down fashion, taking into consideration facts from languages from increasingly more subfamilies as the size of the sample grows, a variety sample may also be used to demonstrate that some grammatical phenomenon is actually *less* common than is generally assumed. For example, a variety sample of 52 languages that is representative of the world's languages contained only one clear example of a language that has a proper *agreement* relation between the subject and the verb: Dutch. In this language, the verb agrees in number / person with the nominal or pronominal subject NP (Rijkhoff 1992: 30–31, Rijkhoff 2004: 246–247).

(1) Dutch: agreement

a. *Ik loop* *[*Ø loop]*
I walk
'I am walking.'

b. *Je / Hij / Zij / Het kind loop-t* *[*Ø loopt]*
you.SG / he / she / the child walk-PRES.2/3SG
'You are / He / She / The child is walking.'

c. *We / Jullie loop-en (spelled: lopen)* *[*Ø lopen]*
we / you.PL walk-PRES.PL
'We / You are walking.'

d. *Ze / De meisje-s loop-en (spelled: lopen)* *[*Ø lopen]*
They / the girl-PL walk-PRES.PL
They / The girls are walking.''

Notice that the subject NP or pronoun is compulsory in Dutch and that the agreement marker on the verb is not a referential element. The grammat-

ical AGREEMENT relation attested in Dutch is to be distinguished from the CROSS-REFERENCING relation, where the person marker on the verb does have referring potential (as in Latin *cantamus* 'we sing') and the free subject NP or pronoun is normally only expressed under special circumstances, such as emphasis (Latin *nos cantamus* 'WE sing'; Dik 1997: 154; Siewierska 2004: 121). The latter phenomenon is often called 'Pro-drop', which is a misnomer of course, as the free subject pronoun is added rather than dropped.

In other words, contrary to what is commonly assumed, grammatical Subject-Verb agreement as attested in Dutch, German and certain other Germanic languages (not e. g. Danish), is very marked from a cross-linguistic perspective. In a variety sample of 272 languages, it only occurred in one other language besides Dutch: Vanimo, a New Guinea language of the Sko family (Siewierska 1999: 238–239; see also de Vogelaer and van der Auwera, (2010)). Next we will discuss the crucial role *rara* have played in the further development of a theory about parallels between the structure of noun phrases and clauses.

3 The importance of *rara* for grammatical theory

A general theory of grammar cannot be developed without being exposed to linguistic facts from a wide variety of (genetically, geographically, typologically) different languages, and, conversely, empirical research is best guided by theoretical questions (Rijkhoff 2002).[1]

The main goal of this paper is to show how two rare grammatical phenomena made it possible to propose an improved, unified analysis of the noun phrase (NP) and the clause.

Both formal and functional linguists have developed models in which clauses and NPs are at least partly analyzed in a similar fashion. Whereas formal approaches to grammar have concentrated on similarities between the underlying *syntactic* structure of the NP and the clause (e. g. Jackendoff 1977; Abney 1987), theories such as *Functional (Discourse) Grammar* have investigated parallels between functional, meaning-based representations of the NP and the clause (Rijkhoff 1990, 1992, 2004, 2008b). Thus, in the functional approach to grammar, linguistic forms and constructions are also characterized in terms of labels that capture the *functional* contribution they make to a linguistic expression (Rijkhoff 2008b, 2009). Furthermore, the design of 'structural-functional theories' (Butler 2003) like *Role and Reference Grammar* and *Functional (Discourse) Grammar* is based on a multilayered con-

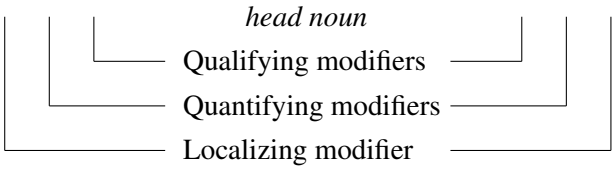


Figure 1. Layers of modification in the NP (Rijkhoff 1992)

ception of linguistic structure, i. e. a hierarchical type of meaning-based linguistic structuring, where a higher layer (in terms of scope) is seen as being superimposed on other, lower, layers (Butler and Taverniers 2008: 680–681).

The first layered model of the noun phrase was proposed in Rijkhoff (1988), shortly after Hengeveld (1987, 1988) had put forward a layered analysis of clausal structures. Initially only three layers of modification were recognized in the NP: the innermost layer for qualifying modifiers (*Quality Layer*), the outermost layer for localizing modifiers (*Location Layer*) and in between a *Quantity Layer* to accommodate quantifying modifiers (Figure 1; recall that this meaning-based representation reflects differences in scope rather than syntactic relations).

QUALIFYING MODIFIERS have scope over the head constituent and specify more or less inherent properties (qualities or attributes) of the referent of the noun phrase, such as size (*a small house*, *telescopes of enormous size*), value or quality (*an expensive suit*, *wine of an incredible richness*), age (*a young child*, *youths under age 15*), or colour (*red curtains*, *a Jovian moon of incredible redness*). These examples also show that the same function (here QUALIFYING MODIFIER) can be performed by members that belong to different form classes, such as adjectives (*big*) or prepositional phrases (*of enormous size*). Languages without a distinct class of adjectives, such as Eastern Ojibwa (Algonquian; USA and Canada), often employ relative clauses headed by a stative verb instead.

(2) Eastern Ojibwa (Dryer 2008)

nini e-gnoozi-d
man REL.PX-tall-3SG
'a tall man'

QUANTIFYING MODIFIERS have scope over the quality layer and the head constituent and indicate number distinctions and the cardinality of the referent of the NP ('*two big car-s*'). Some languages employ lexical modifiers for

this purpose. For example, Krongo (Kordofanian) uses verbal forms to indicate the cardinality of a referent.

(3) Krongo (Reh 1985: 252)

nóo-còorì nk-óotòonò
PL-house CN.PL-IMPF:be_three
'three houses'

LOCALIZING MODIFIERS have the widest scope, specifying properties concerning the location of the referent in physical or cognitive space, as in '*those two old houses*', '*the house on that hill*', but also e. g. '*my house*' (on the relation between location and possession, see Rijkhoff 2004: 173–212).

It was already argued in Rijkhoff (1988) that NPs and clauses (or rather "predications") could be analyzed in a similar fashion. The fact that the layers are organized hierarchically, meaning that modifiers of an outer layer have scope over modifiers of an inner layer, is indicated schematically and in a theory-neutral fashion in Figure 2 on the following page (a detailed, formal representation can be found in Rijkhoff 2008b). By convention, grammatical (i. e. non-lexical) categories like Definiteness, Number, Tense or Aspect are represented on the left side, i. e. before the head constituent (verb or noun), whereas phrasal or lexical modifiers like adjectives, genitives, adverbs or adpositional phrases (which involve content words) are specified after the head.

Since there is often no one-to-one relationship between the form and the function of a modifier (see above), only some typical forms or constructions for the three modifier categories are specified in Figure 2. For example, adnominal prepositional phrases can be employed as qualifying (*children under age 6*) or localizing modifiers (*the cat under the table*). In Figure 3 on the next page, the modifiers in '*those three black* dogs *in the garden*' appear in the appropriate slots of the various layers.

Subsequently it became apparent that the layered structure of the NP is more articulate and that at least two more layers of modification should be added, one for *discourse-referential modifiers* (Section 3.1) and another to accommodate *classifying modifiers* (Section 3.2). It was not immediately clear, however, what elements could qualify as a classifying or discourse-referential modifier at the level of the clause. It was at this point that certain rare linguistic features played an essential role in the further development of a theory about parallels between meaning-based representations of NP and clause structure.

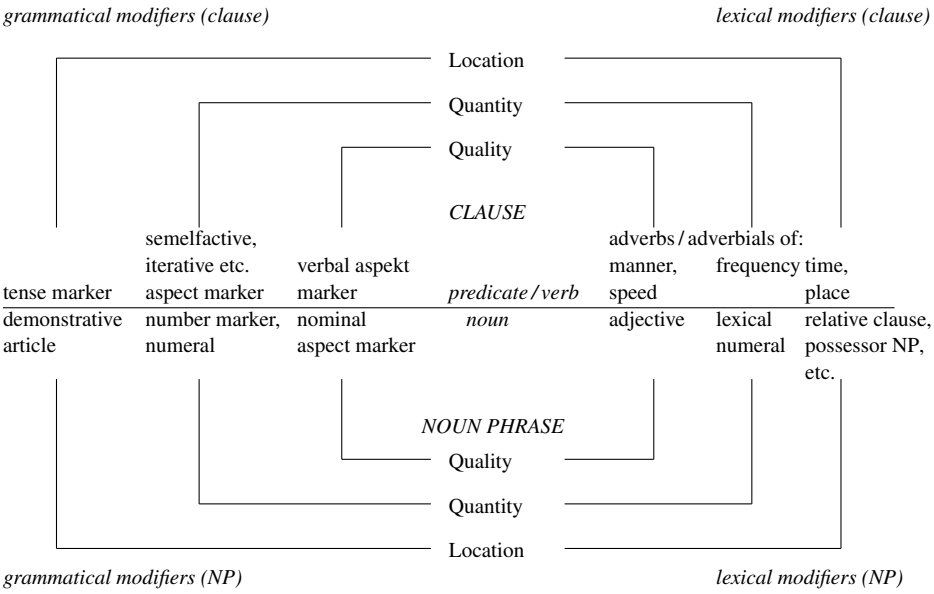


Figure 2. Symmetry in the underlying structure of the clause and the NP as proposed in Rijkhoff 1992 (see Figure 6 for the current version)

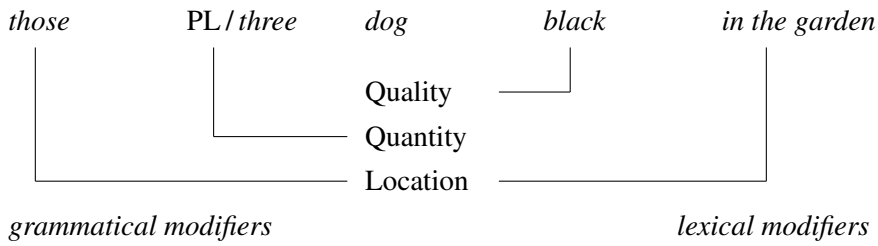


Figure 3. Modifiers in a layered model of NP: a simplified representation of '*those three black dogs in the garden*'

3.1 Discourse-referential modifiers in the clause and in the NP

In the first version of the layered model of the NP, definite articles were regarded as LOCALIZING MODIFIERS (together with e. g. adnominal demonstratives), because they indicate 'weak deixis' (Anderson and Keenan 1985:

261–262). There are some important differences, however, between definite articles and other localizing modifiers, such as adnominal demonstratives. For example, whereas adnominal demonstratives basically specify a spatial property of the referent (its spatial location relative to a reference point), definite articles indicate that the speaker assumes the referent of the NP to be identifiable for the addressee in the shared world of discourse (a mental construct). In other words, the difference between demonstratives and definite articles can be compared to the difference between semantics and pragmatics (cf. Levinson 2000). Whereas demonstratives are concerned with a static, descriptive semantic property of the referent of the NP ("language as carrier of content"), definite and indefinite articles relate to a more dynamic, pragmatic property of the referent: its referential or existential status in the world of discourse ("language as exchange", "language as communicational process"). By using a definite NP, the speaker signals that (s)he assumes that the addressee is familiar with the referent of the NP or that the addressee will be able to identify the referent in conversational space. Conversely, by using an indefinite NP, the speaker indicates that the addressee is not expected to identify the referent. In the literature, the distinction between the coding of a static descriptive meaning and the coding of a dynamic inter-subjective meaning goes back to Bühler (1999 [1934]) and has more recently been discussed in terms of the notions 'ideational' vs. 'interpersonal' (Halliday 2004: 61) or 'representational' vs. 'interpersonal' (Hengeveld and Mackenzie 2008) and can also be paraphrased as the coding of content-related meaning vs. the coding of speaker / hearer-related meaning (Butler 2003: 111; Butler and Taverniers 2008: 681).

Since adnominal demonstratives and markers of (in)definiteness belong to different functional modifier categories, it seemed appropriate to assign (in)definite articles and other modifiers that indicate interpersonal or discourse properties of the referent their own place in a layered model of the noun phrase. Thus a new layer was established to accommodate so-called DISCOURSE-REFERENTIAL MODIFIERS. After markers of (in)definiteness were re-categorized as *grammatical* instances of Discourse-Referential modifiers, it was not very difficult to find *lexical* instances of this new category: modifiers such as *same* and *other* also provide the addressee with information about the referent as a discourse entity. The modifier (*the*) *same* tells the addressee (s)he should find a particular referent that was mentioned a little earlier in the conversation. In the case of (*the*) *other*, the addressee is instructed to identify the second member of a referent set that was introduced

before. By using the indefinite counterpart *another* (as in: *I gave her another book*) the speaker instructs the addressee to *construe* a new token of a certain kind of entity (*book*) that already exists as a distinct referent in the discourse world and was mentioned not long before (Rijkhoff 2008c: 798–800).

If it is correct that all the layers of modification that we find in the NP are also relevant for the analysis of clausal structures, then one would expect there to exist discourse-referential modifiers in the clause as well.[2] The occurrence of a rare phenomenon attested in Jacaltec, a Mayan language spoken in Guatemala (see also Martin 1998 and Vidal and Klein 1998), was of crucial importance in the search for clausal counterparts of markers of (in)definiteness and modifiers like *same* and *other*. Jacaltec employs the same morpheme to express NONSPECIFIC-INDEFINITENESS in the NP and IRREALIS in the clause (variation between /oj/ and /uj/ is due to vowel harmony):

(4) Jacaltec (Craig 1977: 93)

a. *X-Ø-'oc heb ix say-a' hun-uj munlabel*
ASP-ABS.3-start PL woman look_for-FUT a-**OJ** pot
'The women started looking for a pot.' [nonspecific reference]

b. *Way-oj ab naj*
sleep-**OJ** EXH CLF/he
'Would that he slept!' [irrealis: exhortative mood]

The reason that speakers of Jacaltec can use /OJ/ both in the NP and in the clause is, presumably, that both Nonspecific-indefiniteness and Irrealis signal to the addressee that the (spatial or temporal) entity in question is not 'anchored' or 'grounded' in the world of discourse at the moment of speaking, i. e. it does not exist (yet) as an identifiable object or a real event at a certain time or place in the shared word of discourse of Speaker and Addressee(s). The fact that the clausal category Irrealis, just like Nonspecific-indefiniteness in the NP, relates to the existential status of an entity in the world of discourse is a strong indication that irrealis markers should be regarded as Discourse-Referential modifiers at the level of the clause.

More evidence for this analysis is provided by another rare grammatical phenomenon. In Fongbe (a Kwa language spoken mainly in Benin) the same morpheme is used to express DEFINITENESS in the NP and REALIS in the clause. To put it differently, speakers of Fongbe use the same marker in NPs and in clauses to indicate that a thing or event is *grounded* in the world of discourse, i. e. that it occupies a specific spatio-temporal region in the world of discourse at the moment of speaking (Rijkhoff 1990; Rijkhoff and Seibt

2005). In these examples from Fongbe the element in question is glossed as DET (determiner)

(5) Fongbe (Lefebvre 1998: 94, 99; cf. also Lefebvre and Brousseau 2002)

a. *N ɖú àsɔ́n ɔ́*
I eat crab DET
'I ate the crab.' (i. e. the crab in question / that we know of)

b. *Jan wá ɔ́*
John arrive DET
'Actually, John arrived.'

These data from Jacaltec and Fongbe were essential in demonstrating that a typologically adequate representation of NP and clause structure needs to provide a separate slot for DISCOURSE-REFERENTIAL MODIFIERS, which relate to the existential status of the referent (thing or event) in the world of discourse at the moment of speaking (Figure 4):[3]

(a) *Definite* and *realis* indicate that the thing or event being referred to by the speaker is grounded (occupies a certain spatio-temporal region) in the world of discourse, and

(b) *Nonspecific-indefinite* and *irrealis* indicate that the thing or event being referred to by the speaker is not grounded (does *not* occupy a certain spatio-temporal region) in the world of discourse.

Noun Phrase (thing)	OCCURRENCE IN WORLD OF DISCOURSE AT MOMENT OF SPEAKING	Clause (event)
Definite Reference	GROUNDED IN DISCOURSE WORLD	*Realis Mood*
Nonspecific-indefinite Reference	NOT GROUNDED IN DISCOURSE WORLD (YET)	*Irrealis Mood*

Figure 4. Symmetry between Definite / Realis and Nonspecific-indefinite / Irrealis.

Once (ir)realis markers were identified as the clausal counterparts of the grammatical discourse-referential modifiers in the NP, it was obvious that adverbs like 'actually' or 'really' should be analyzed as lexical manifestations of discourse-referential modifiers in the clause. These adverbs are concerned

with the existential status of events as discourse entities in conversational space, as they indicate whether or not the speaker regards the event in question is grounded or real (Rijkhoff 2008b: 68–74). The hierarchical position of Discourse-Referential modifiers in the layered model (now the outermost layer) can be seen in Figure 5 (see Rijkhoff 2008b, 2008c for details about scopal and other properties of discourse-referential modifiers):

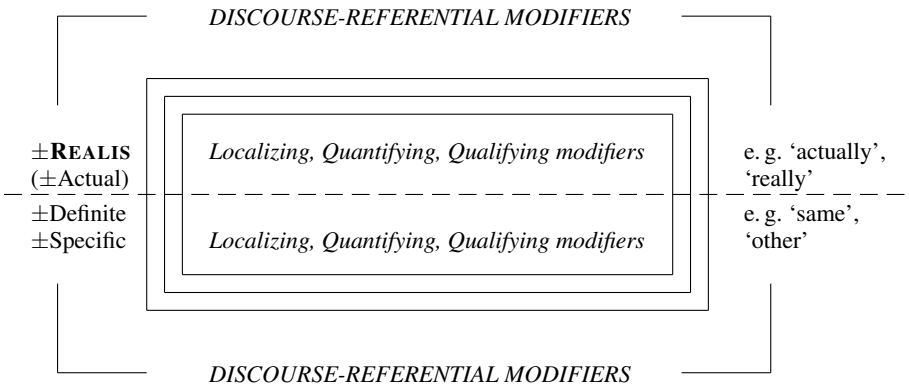


Figure 5. Discourse-referential modifiers in a layered representation of the underlying structure of the clause and the NP

3.2 Classifying modifiers in the clause and in the NP

Several linguists have discussed the fact that certain modifiers in the NP classify the entity as denoted by the head noun into a system of smaller sets (Quirk et al. 1985; Halliday 2004), as in these examples: *electric train*, *divorce lawyer*, *departmental meeting*. In languages that have them, classifying adjectives (typically noun-derived or 'relational' adjectives) differ from other adjectives in several ways. For instance, they immediately precede or follow the head noun and they do not admit intensifiers, comparison, or predicative position:

(6) [adjacency]

a. *the popular corporate lawyer* vs. **the corporate popular lawyer*

(7) [intensifier]

a.	*the electric train*	vs. **the very electric train*
b.	*the long train*	vs. *the very long train*

(8) [comparison]

a.	*a medical examination*	vs. **a more medical examination*
b.	*an expensive book*	vs. *a more expensive book*

(9) [predicate]

a.	*the presidential election*	vs. **the election is presidential*
b.	*the white table*	vs. *the table is white*

The search for *lexical* manifestations of classifying modifiers in the clause did not yield any clear results until we became aware of the unusual grammatical phenomenon of 'stripped nouns' (Mithun 1984; Miner 1986, 1989; Gerdts 1998; Caballero et al. 2008). The fact that some languages have a distinct category of free lexical modifiers at the level of the clause that subclassify an event made it possible to add another shared layer of modification to the *symmetrical* representation of NP / clause structure.

Superficially, stripped nouns are similar to incorporated nouns but there is an important difference. Whereas an incorporated element is part of another word, a stripped noun, which must appear next to the verb it modifies, is a separate word according to phonological criteria such as stress placement. Thus, in Kusaiean adverbs can appear between verb and object (10), but not between verb and stripped noun (11). Notice that in these examples a distinction is made between sharpening in general (10) and a particular *kind* of sharpening, namely knife-sharpening (11).

(10) Kusaiean (Gerdts 1998: 94; original example in Lee 1975)

Sah el twem upac mitmit sac
Sah he sharpen diligently knife the
'Sah is sharpening the knife diligently'

With a stripped noun:

(11) *Sah el twetwe mitmit upac*
Sah he sharpen knife diligently
'Sah is diligently knife-sharpening'

Adding a separate layer of modification to accommodate CLASSIFYING MODIFIERS resulted in the symmetrical analysis of NP and clause structure

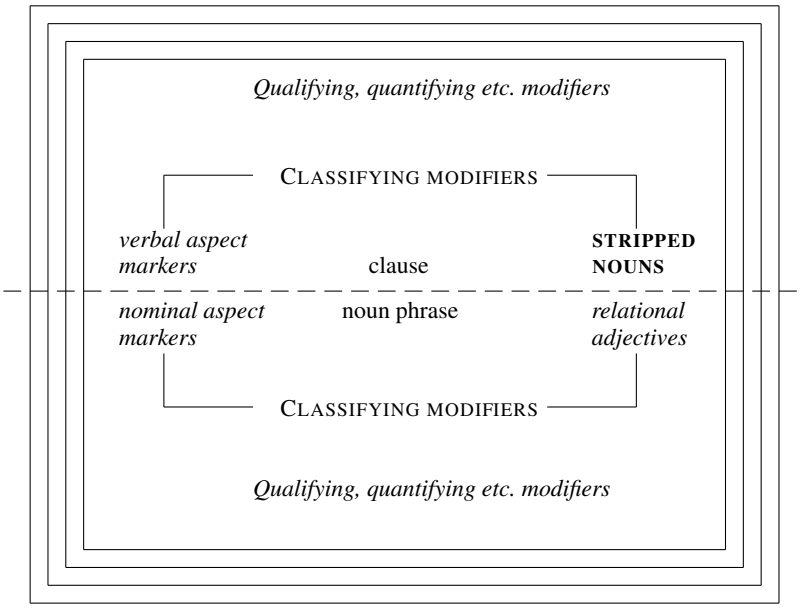


Figure 6. CLASSIFYING MODIFIERS in a layered analysis of symmetrical NP and clause structure.

shown in Figure 6 (recall that scopal relations among modifier categories are discussed in Rijkhoff 2008b, 2008c). Notice that nominal and verbal aspect markers, originally treated as QUALIFYING MODIFIERS (Rijkhoff 2004: 224), are now analyzed as CLASSIFYING MODIFIERS (this is explained in Rijkhoff 2008b: 84–86).

4 A five-layered model of the clause and the NP

Adding the two new layers discussed in Sections 3.1 and 3.2 to the original proposal (Figure 2 on page 228) gives the following five-layered symmetrical NP / clause structure.

As can be seen in Figure 7 on the facing page, there is now a certain asymmetry in the distribution of grammatical and lexical modifiers in that grammatical manifestations of QUALIFYING MODIFIERS are no longer deemed to exist (this is explained in Rijkhoff 2008a: 85–86; see also Rijkhoff 2008b: 794).

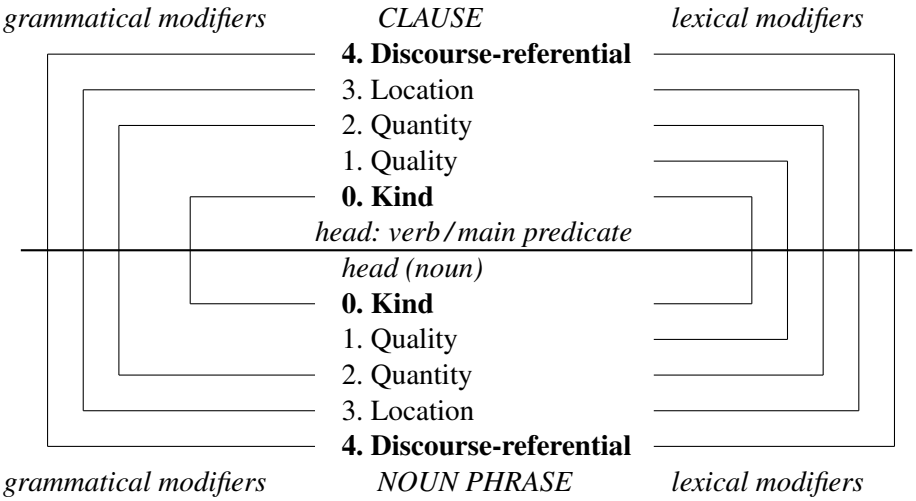


Figure 7. Parallels in the layered structure of the NPs and the clause

5 Conclusion

In this paper I have tried to highlight the importance of *rara* for grammatical theory by showing how hypotheses about certain parallels between the structure of NPs and clauses could only be confirmed on the basis of two uncommon linguistic phenomena: (i) stripped nouns and (ii) isomorphic markers of Definiteness / Realis (Fongbe) and Indefiniteness / Irrealis (Jacaltec). Thus, it would be wrong to treat *rara* as curiosities that are only marginally interesting for the study of language. *Rara* and *rarissima* are at least as important for the development of a general, typologically adequate theory of grammar as the more commonly attested linguistic phenomena.

Abbreviations

1 / 2 / 3 = 1st / 2nd / 3rd person; A = adjective; ABS = absolutive; ASP = aspect; CL = class; CLF = classifier; COM = comitative; DET = determiner; DV = diversity value; EXH = exhortative; N = noun; NP = noun phrase; M = masculine; PAST = past; PERF = perfective; PL = plural; PRES = present; REL.PX = relativizing prefix; S = subject; SG = singular; V =verb

Notes

1. The claim that empirical research needs some kind of theoretical perspective is not new, of course, as can be seen, for example, in this quotation from a letter by Charles Darwin to Henry Fawcett (10 September 1861):

   > "About thirty years ago there was much talk that geologists ought to observe and not theorise; [. . .] at this rate a man might as well go into a gravel-pit and count the pebbles and describe the colours. How odd it is that anyone should not see that all observations must be for or against some view if it is to be of any service."

2. Clauses have at least two additional layers of modification: one for modal and another for illocutionary distinctions (Rijkhoff 2008a: 101; Rijkhoff 2008b: 812).
3. Referents of definite NPs can be grounded (or presumed grounded) for various reasons, whereas referents of real events become grounded at the moment of speaking. These and other aspects of the anti-symmetrical relation between (in)definiteness and (ir)realis, in which Definite aligns with Irrealis and Specific-Indefinite with Realis, are discussed in Rijkhoff and Seibt (2005) and Rijkhoff (2008b).

References

Abney, Stephen P.
1987 *The English noun phrase in its sentential aspect.* Doctoral dissertation, Massachusetts Institute of Technology.

Anderson, Stephen R., and Edward L. Keenan
1985 Deixis. In *Language Typology and Syntactic Description. Volume III: Grammatical Categories and the Lexicon*, Timothy Shopen (ed.), 259–308. Cambridge: Cambridge University Press.

Bühler, Karl
[1934] 1999 *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Reprint. Stuttgart: UTB. [Original edition: Jena: Fischer, 1934.]

Butler, Christopher S.
2003 *Structure and Function: A Guide to Three Major Structural-Functional Theories. Part 1: Approaches to the Simplex Clause* (= Studies in Language Companion Series; 63). Amsterdam / Philadelphia: Benjamins.

Butler, Christopher S., and Miriam Taverniers
2008 Introduction [to a special issue on 'layering' in grammatical theory]. *Linguistics* 46.4: 679–687.

Caballero, Gabriela, Michael J. Houser, Nicole Marcus, Teresa Mcfarland, Anne Pycha, Maziar Toosarvandani, and Johanna Nichols
2008 Nonsyntactic ordering effects in noun incorporation. *Linguistic Typology* 12.3: 383–421.

Craig, Colette Grinevald
1977 *The Structure of Jacaltec*. Austin: University of Texas Press.

de Vogelaer, Gunther, and Johan van der Auwera
2010 *When typological rara become productive: the extension of grammatical agreement in Dutch dialects.* In *Rara & Rarissima: Documenting the fringes of linguistic diversity*, Jan Wohlgemuth and Michael Cysouw (eds.), 47–73. (= Empirical Approaches to Language Typology; 46). Berlin / New York: Mouton de Gruyter.

Dik, Simon C.
1997 *The Theory of Functional Grammar* (2nd revised edition edited by Kees Hengeveld). Part 1: The Structure of the Clause. Part 2: Complex and Derived Constructions. Berlin / New York: Mouton de Gruyter.

Dryer, Matthew S.
2008 Order of adjective and noun. (Chapter 87) In *The World Atlas of Language Structures Online*, Martin Haspelmath, Matthew S. Dryer, David Gil and Bernard Comrie (eds.). München: Max Planck Digital Library. http://wals.info/feature/87 (31 December 2008).

Gerdts, Donna B.
1998 Incorporation. In *The Handbook of Morphology*, Andrew Spencer and Arnold M. Zwicky (eds.), 84–100. Oxford: Blackwell.

Halliday, Michael Alexander Kirkwood
2004 *An Introduction to Functional Grammar*. 3rd edition, revised by Christian M. I. M. Matthiessen. London: Arnold.

Hengeveld, Kees
1987 Clause structure and modality in Functional Grammar. In *Ins and Outs of the Predication*, Johan van der Auwera and Louis Goossens (eds.), 53–66. (= Functional Grammar Series; 6). Dordrecht: Foris.

Hengeveld, Kees
1988 *Layers and operators*. (= Working Papers in Functional Grammar; 27). Department of Linguistics, University of Amsterdam.

Hengeveld, Kees and J. Lachlan Mackenzie
2008 *Functional Discourse Grammar: a typologically based theory of language structure*. Oxford: Oxford University Press.

Jackendoff, Ray S.
1977 *X-bar Syntax*. Cambridge, MA: MIT Press.

Lee, K.
1975 *Kuseian Reference Grammar*. Honolulu: University Press of Hawaii.

Lefebvre, Claire
1998 Multifunctionality and variation among grammars: the case of the determiner in Haitian and in Fongbe. *Journal of Pidgin and Creole Languages* 13.1: 93–150.

Lefebvre, Claire and Anne-Marie Brousseau
2002 *A Grammar of Fongbe* (= Mouton Grammar Library; 25). Berlin / New York: Mouton de Gruyter.

Levinson, Stephen C.
2000 *Presumptive Meanings: the theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Martin, Laura
1998 Irrealis constructions in Mocho (Mayan). *Anthropological Linguistics* 40.2: 198–213.

Miner, Kenneth L.
1986 Noun stripping and loose incorporation in Zuni. *International Journal of American Linguistics* 52.3: 242–254.

Miner, Kenneth L.
1989 A note on noun stripping. *International Journal of American Linguistics* 55: 476–477.

Mithun, Marianne
1984 The evolution of noun incorporation. *Language* 60.4: 847–894.

Perkins, Revere D.
1988 The covariation of culture and grammar. In *Studies in Syntactic Typology*, Michael Hammond, Edith A. Moravcsik, and Jessica R. Wirth (eds.), 359–378. (= Typological Studies in Languages; 17). Amsterdam / Philadelphia: Benjamins.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik
1985 *A Comprehensive Grammar of the English Language*. London: Longman.

Reh, Mechthild
1985 *Die Krongo-Sprache (Nìinò Mó-dì) – Beschreibung, Texte, Wörterverzeichnis*. Berlin: Reimer.

Rijkhoff, Jan
1988 A typology of operators: toward a unified analysis of terms and predications. *Working Papers in Functional Grammar* 29. Department of Linguistics, University of Amsterdam.

Rijkhoff, Jan
1990 Toward a unified analysis of terms and predications. In *Layers and Levels of Representation in Language Theory: a functional view*, Jan Nuyts, A. Machtelt Bolkestein, and Co Vet (eds.), 165–191. Amsterdam / Philadelphia: Benjamins.

Rijkhoff, Jan
1992 *The noun phrase: a typological study of its form and structure*. Doctoral dissertation, University of Amsterdam.

Rijkhoff, Jan
2002 On the interaction of Linguistic Typology and Functional Grammar. *Functions of Language* 9.2: 209–237.

Rijkhoff, Jan
2004 *The Noun Phrase*. Oxford: Oxford University Press. [expanded paperback edition of 2002 hardback publication].

Rijkhoff, Jan
2008a Synchronic and diachronic evidence for parallels between noun phrases and sentences. In *Interdependence of Diachronic and Synchronic Analyses*, Folke Josephson and Ingmar Söhrman (eds.), 13–42. (= Studies in Language Companion Series; 103). Amsterdam / Philadelphia: Benjamins.

Rijkhoff, Jan
2008b Layers, levels and contexts in Functional Discourse Grammar. In *The Noun Phrase in Functional Discourse Grammar*, Daniel García Velasco and Jan Rijkhoff (eds.), 63–115. (= Trends in Linguistics. Studies and Monographs; 195). Berlin / New York: Mouton de Gruyter.

Rijkhoff, Jan
2008c Descriptive and discourse-referential modifiers in a layered model of the noun phrase. *Linguistics* 46.4: 789–829.

Rijkhoff, Jan
2009 On the co-variation between form and function of adnominal possessive modifiers in Dutch and English. In *The Expression of Possession*, William B. McGregor (ed.), 51–106. (= The Expression of Cognitive Categories; 2). Berlin / New York: Mouton de Gruyter.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld and Peter Kahrel
1993 A method of language sampling. *Studies in Language* 17.1: 169–203.

Rijkhoff, Jan, and Dik Bakker
1998 Language sampling. *Linguistic Typology* 2.3: 263–314.

Rijkhoff, Jan, and Johanna Seibt
2005 Mood, definiteness and specificity: a linguistic and a philosophical account of their similarities and differences. *Sprog — Tidsskrift for Sprogforskning* 3.2: 85–132. http://ojs.statsbiblioteket.dk/index.php/tfs/article/view/89/74

Ruhlen, Merritt
1987 *A guide to the world's languages. Vol. 1: Classification*. London: Arnold.

Ruhlen, Merritt
1991 *A guide to the world's languages. Vol. 1: Classification (with a postscript on recent developments)*. London: Arnold.

Siewierska, Anna
1999 From anaphoric pronoun to grammatical agreement marker: why objects don't make it. *Folia Linguistica* 33: 225–251.

Siewierska, Anna
2004 *Person*. Cambridge: Cambridge University Press.

Vidal, Alejandra and Harriet P. Manelis Klein
1998 Irrealis in Pilagá and Toba? Syntactic versus pragmatic coding. *Anthropological Linguistics* 40.2: 175–197.

Pairwise comparisons of typological profiles

Søren Wichmann & Eric W. Holman

1 Introduction

Rare linguistic features, languages possessing many rare characteristics or areas possessing many languages with rare characteristics are all relative phenomena. Studying *rara* takes the researcher into a complex area, an area which gets the more complex the more languages are involved. To develop a sense of the determinants of this complexity it is worthwhile to conduct a thought experiment where a maximally simple situation is considered first, and where various additional layers are only subsequently added. In a maximally simple situation there would only be one language in the world, and in this situation rarity would be a non-existing phenomenon. In a situation with just two languages, rarity still could not be defined, but we could talk about similarities and differences between the two languages. Two languages can be more or less similar. It is probably a truism that two major factors contributing to similarities among languages are common descent and diffusion. Other factors, whose effects are less easy to identify, are universal tendencies and chance. Leaving these last two factors aside, we can predict that, all else being equal, geographic propinquity and relatedness are expected to enhance similarities among the members of any pair of languages. If we enlarge the picture to include three languages, then, we would predict that among a set of three unrelated languages the two most proximate geographically would be the most similar; and within a set of three languages spaced equally apart where language A and B are related and language C unrelated to the two others we would expect A and B to be more similar to one another than either is to C. Once we have entered a situation where three languages are considered the concept of *rara* becomes relevant. A trait found in only one language to the exclusion of the two others is, by any definition, rare. Such a trait could exist in any language within a set of three, but among the three languages the least closely related or most geographically remote language would have a higher chance of possessing such a rare trait. Given a world-wide sample of languages we would still expect geography and relatedness to be at the root of the phenomenon of *rara*, but with a large set of languages the confounding

factors accumulate. Genealogical and areal biases in the sampling becomes an issue, peculiarities of geography enter the picture, differential degrees of feature attestation must be considered, as well as biases in the selection of traits. One may try to take all or most of these factors into account when developing a statistical approach to *rara*, cf. Cysouw (in press) for an attempt, but, in the end, if a given feature is promoted to the status of *rara*, or if a given language is identified as being quirky or if an area is found to be typologically unusual it will be hard to interpret the result because so many factors are in play: diffusion, relatedness, chance, universal tendencies, peculiarities of geography, sampling of languages and features, and degrees of attestation. In addition, what is rare today may have been common yesterday and may not tell us anything interesting about languages in general. In other words, the global linguistic typological profile is in constant flux, and there is no doubt that various historical contingencies which are beyond the observational reach of the investigator contribute to a large extent to these dynamics.

In this paper we are interested in the phenomenon of "language rarity", i. e. the degree to which languages differ from one another as a whole. In the belief that a complex phenomenon is best approached by isolating its components, we will consider only pairs of languages. As mentioned, it is really only in the situation where three or more languages are compared that rarity is a relevant concept, but since the multilateral comparisons necessary for making observations on rarity decompose into a set of pairwise comparisons, a fundamental approach to the problem of why some languages stand apart from all or most others is to ask why pairs of languages exhibit differential degrees of similarity.

Our hypothesis is that there are two major factors which contribute to similarities among languages: relatedness and propinquity. The latter factor, i. e. the influence of geography on typological similarity, was investigated in Holman et al. (2007). In this paper, then, we focus more on the role that genealogical relatedness plays with respect to similarities among language pairs. In particular, we are interested in knowing whether there is a cut-off point S_{high} in the amount of similarities such that we can be sure that language pairs that have more than S_{high} similarities are all generally thought to be related, and we also want to know whether there is a cut-off point S_{low} at the other end of the scale such that all languages having less similarities than S_{low} are thought to be *un*related. In other words, if a language is relatively similar to some other language, as Burushaski is to Telugu, just to name an example, does this imply that the two languages are related according to commonly

accepted classifications? Or, if two languages are mutually very exotic, as Burushaski and Samoan, for instance, does this imply that they are thought not to be related in commonly accepted classifications?

The data we use, as well as the genealogical classification, are from *The World Atlas of Language Structures* (Haspelmath et al. 2005, eds., henceforth WALS). The conclusions must of course be seen in relation to this particular dataset. Thus, when we observe a certain amount of typological similarity between two languages, this is strictly and only similarity in terms of the kinds of features investigated in WALS. The dataset includes 134 non-redundant features, each of which distinguishes two to nine discrete values. All of these are quite generic typological features. Our conclusions are also limited to the amount of data available. We have required that for any language pair in our sample there should be 45 or more features attested for both members of the pair (a motivation for this precise number follows shortly). This has limited our sample to 320 languages and 29,810 pairs of languages compared. Among these pairs, there are 1,099 which are considered to be related, according to the classification in WALS. This classification, described by Dryer (2005), has two taxonomic levels. Families are defined as the most inclusive groups believed by a majority of specialists to have descended from a common ancestral language. Genera are defined as the most inclusive groups whose common ancestor is believed to have existed no more than about 3,500 to 4,000 years ago. Languages in the same WALS family are henceforth called "related".

2 Results

Figure 1 on the next page presents the overall results of the investigation. As can be seen, the more similar languages get, the greater the probability is that they are related. The figures on which the curve is based are presented in Table 1 on the same page. Percent similarity was defined as the percentage of available features for which both languages have the same value. We have binned language pairs in 5 % intervals from 10 % to 90 % similarity. Figure 1 plots the percentage of the pairs in each bin that are related, as a function of their mean percent similarity. Table 1 gives some additional information: it also shows how many language pairs belong in each interval. This is important for the interpretation of the results, as we shall see shortly.

Before giving our interpretation let us explain why we have chosen the criterion that language pairs should have 45 or more features attested for both

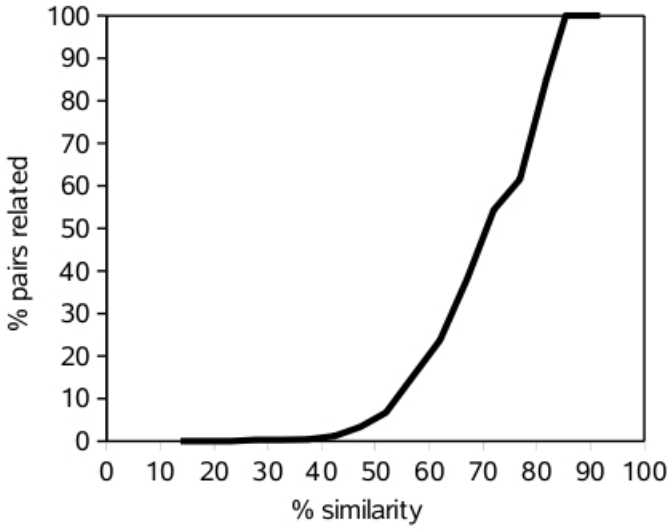


Figure 1. The probability of finding related languages as a function of their similarity

Table 1. Mean percent similarity between members of pairs, percent of language pairs that are related, and number of language pairs, in each similarity interval

% similarity interval	Mean % similarity	% related	Pairs
10.0–14.9	13.8	0.00	11
15.0–19.9	18.1	0.00	91
20.0–24.9	23.0	0.00	443
25.0–29.9	27.8	0.26	1,566
30.0–34.9	32.7	0.33	3,904
35.0–39.9	37.5	0.40	6,019
40.0–44.9	42.4	1.20	6,772
45.0–49.9	47.2	3.26	4,873
50.0–54.9	52.0	6.68	3,520
55.0–59.9	57.1	15.41	1,551
60.0–64.9	62.0	23.72	666
65.0–69.9	67.1	38.24	238
70.0–74.9	72.0	54.26	94
75.0–79.9	76.9	61.54	39
80.0–84.9	81.8	85.00	20
85.0–89.9	85.4	100.00	2
90.0–94.9	91.8	100.00	1

languages. We tested results for different numbers of features (30 or more, 45 or more, 60 or more, 75 or more). It turns out that for a criterion of 30 or more features the curve is rather similar but not quite as steep, showing less dependence between the amount of similarity and the probability of finding related pairs. This indicates that the fewer features one operates with, the more prominent is random sampling variability in percent similarity. When operating with a criterion of 60 or more attested features the curve becomes uneven, indicating that the higher criterion passes too few pairs for stable results. This becomes even more pronounced when the criterion is 75 or more features. Obviously, with a more extended database the number of features taken to be criterial could be raised, but around 45 is a number that suits the data available in WALS because a number in the vicinity of 45 maximizes the combined information in terms of both the number of languages and features available.

It may be of interest to mention the language pairs that fall in the lower and upper ranges of the percentage of shared values. Collectors of linguistic trivia may find it interesting that the members of the most divergent language pair in the world (in our dataset), i. e. Tümpisa Shoshone and Wari', are both native American languages, that someone who wants a radical alternative to Romance linguistics should turn to Nivkh, and that someone wanting to study a language as different as possible from Swedish should visit the Koasatis. Lists of the 20 most divergent language pairs and the 20 most similar ones are provided in Tables 2 on the following page and 3 on page 247.

While Table 2 does not point in any specific direction and remains a curiosity, Table 3 provides fragments of information which fits into the larger picture that emerges from our study. We note that two pairs of unrelated languages, Vietnamese–Thai and Khmer–Thai, turn up in this list, which otherwise consists of genealogically related language pairs. Furthermore, the rest of the pairs represent a mixture of languages related to different degrees.

Returning to Figure 1 and the associated data in Table 1 let us proceed to overall interpretations. We set out asking whether there is some degree of similarity in typological profiles beyond which it is certain that languages are related. The answer is positive, but nevertheless discouraging. Members of language pairs in the sample that are 81.5 % or more similar are all related. But only twelve pairs of languages are similar to such an extent, in spite of the fact that there are 1,099 pairs of related languages in the sample! On the other hand, if there are less than 25 % shared feature values all language

Table 2. The 20 most divergent language pairs in the sample

Language A	Language B	Features compared	% similarity
Tümpisa Shoshone	Wari'	48	10.4
Archi	Tukang Besi	46	13.0
Maybrat	Limbu	45	13.3
Italian	Nivkh	51	13.7
Burushaski	Samoan	49	14.3
Tzutujil	Burmese	49	14.3
Ju|'hoan	Yup'ik (Central)	56	14.3
Maybrat	Tamil	55	14.5
Nubian (Dongolese)	Acehnese	48	14.6
Swedish	Koasati	47	14.9
Klamath	Wari'	47	14.9
Kongo	Ladakhi	46	15.2
Bashkir	Māori	46	15.2
Berber (Middle Atlas)	Waorani	45	15.6
Lango	Archi	45	15.6
Archi	Thai	45	15.6
Thai	Retuarã	45	15.6
Ijo (Kolokuma)	Kutenai	50	16.0
Kongo	Evenki	56	16.1
Arabic (Egyptian)	Tümpisa Shoshone	48	16.7

pairs will be unrelated, and this goes for 545 pairs in the sample. If one allows for a very small margin of error, it can be predicted that less than 40 % shared feature values implies unrelatedness. Only 41 out of the 12,034 pairs that have 40 % or less shared feature values fail to meet the prediction (the actual pairs are listed in Section 4 below). Thus, lack of similarity is a good predictor of unrelatedness, but presence of similarity is a bad predictor of relatedness.

3 Are there ways of improving the results?

We next consider the question of whether the prediction of relatedness could be improved somehow. In other studies (Wichmann et al. forthc.; Holman et al. 2007) we have made exact quantitative explorations of the relationship between typological similarity and geographical distance among lan-

Table 3. The 20 most similar language pairs in the sample

Language A	Language B	Relatedness	Features compared	% similarity
Lango	Luo	same genus	46	80.4
Luvale	Zulu	same genus	97	80.4
Khmer	Vietnamese	same family, diff. genera	89	80.9
Vietnamese	Thai	diff. families	110	80.9
Khalkha	Tuvan	same family, diff. genera	48	81.3
Lithuanian	Russian	same family, diff. genera	64	81.3
Greek (Modern)	Bulgarian	same family, diff. genera	64	81.3
Khmer	Thai	diff. families	91	81.3
Polish	Russian	same genus	71	81.7
Russian	Serbian-Croatian	same genus	45	82.2
Swahili	Zulu	same genus	107	82.2
Dagur	Turkish	same family, diff. genera	46	82.6
Telugu	Kannada	same family, diff. genera	47	83.0
Kongo	Nkore-Kiga	same genus	48	83.3
Dutch	German	same genus	56	83.9
Italian	Spanish	same genus	63	84.1
Drehu	Iaai	same genus	46	84.8
English	Swedish	same genus	60	85.0
French	Italian	same genus	64	85.9
Hindi	Panjabi	same genus	49	91.8

guages. Not surprisingly, the greater the geographical proximity is between languages, the more similar they tend to be (this goes for both related and unrelated languages). If one takes into account the areal factor, this might move the cut-off point to allow more accurate predictions of relatedness. To test this strategy, we found the average similarity between pairs of unrelated languages as a function of the distance (as the crow flies) between their centers as defined in the WALS database; we then adjusted the similarity between each pair of languages by dividing the similarity by the average for unrelated languages the same distance apart. The correlation be-

tween adjusted and unadjusted similarities was 0.96. The reason for this is probably that the distance measure, as given in the WALS database, identifies the location of a given language (roughly) with its center of extension. This means that some neighbouring languages, such as German and Dutch, are treated as having a certain geographical distance between them when in reality they don't have any. The more widespread the languages compared are, the bigger this problem gets. Since it is impossible to provide adequate measure of geographical distances for 29,810 language pairs, and not just take recourse to a mechanical measure of distance from one WALS dot to another, it is not viable to improve on the cut-off point in such a way.

Also, the 134 features differ appreciably in the distribution of rarity and commonness among their values. Taking into account the relative rarity of feature values might improve the predictions. To test this strategy, we found the average similarity between pairs of unrelated languages separately for each feature; we then adjusted the similarity between each pair of languages by dividing the similarity by the average for unrelated languages across the same features attested in the given pair. The correlation between adjusted and unadjusted similarities was 0.98. The probable reason is that *rara*, at least as concerns rare WALS feature values, are as prone to diffusion as are more frequent feature values.

Another strategy to improve the results would be to take into account the areality of features. The linguistic typological literature abounds with statements concerning the susceptibility to diffusion of certain features as opposed to others. Wichmann and Holman (2009), however, show that "areality" is not amenable to quantification in any straightforward way since the diffusibility of features varies in different parts of the world.[1] Thus, this strategy is not viable.

A final strategy to try to improve the power of prediction concerning relatedness would be to weight different features or values of features according to their stability. We have explored ways of measuring stability and have established a ranked order of stability for WALS features (Wichmann and Holman, 2009, cf. also Holman et al. 2007: 417–418 for a summary of the method). Conceivably, if the features shared among languages were weighted for their stability the cut-off point could be pushed. On the other hand, Holman et al. (2008: 345–346) report on results suggesting that such a weighting would have little if any effect. They show (in their Figure 4) that correlations between typological distances among languages and distances as de-

fined in traditional classifications are similar whether one uses all 134 WALS features, only the 85 most stable ones or anything in between; and using less than 85 features has a negative effect on the correlations. If excluding unstable features does nothing to improve correlations between typological distances and expert classifications, then a weighting scheme cannot be expected to improve the predictive power concerning relatedness between two languages held by their amount of similarity. The ultimate reason for this has already been stated in the previous paragraph: the predictive power of the similarity measure is upset by diffusibility, and, unlike stability, diffusibility is not something inherent in features — stable features are as likely to diffuse as unstable ones.

4 Deviant language pairs

The results reported on in Figure 1 and Table 1 show that there are a few pairs of languages which are related even though showing less than 40 % similarities, which is the point where pairs tend overwhelmingly not to be related. It serves the record to provide a list of the pairs of related languages that are deviant in the sense that they show less similarity than related languages normally do. This list is provided in Table 4 on the following page.

There are three general explanations for the small amount of similarities among members of the language pairs in Table 4 in addition to possible explanations of lesser generality and therefore lesser interest.

One explanation is that the languages in question tend to belong to very large families where there is more room for variation to arise. It is hardly a coincidence that all language pairs, with the exception of the two Penutian ones, belong to the 10 largest families in the world (counted in terms of numbers of languages according to Gordon 2005).

A second factor is genealogical separation. All of the language pairs belong to different genera according to the classification in WALS. We have also looked at the more detailed genealogical partitions of Gordon (2005); the column in Table 4 headed "Eth. level" indicates the level of closeness according to this classification. A "1" means that the languages are separated by the root of the family tree, i. e., that they are maximally genealogically distinct; a "2" means that they are separated by a node which is one step down from the root; and so on. 71 % of the pairs are in maximally distinct subgroups ("1"s), another 27 % are in maximally distinct subgroups within one and the same highest coordinate branch ("2"s), and only one pair belongs to a more deeply

Table 4. Related languages that have unusually different typological profiles (less than 40 % similarities)

Language A	Language B	Language family	Features compared	% sim.	Eth. level	Dist. (km)
Luvale	Ijo (Kolokuma)	Niger-Congo	52	28.8	2	2,606
Zulu	Ijo (Kolokuma)	Niger-Congo	52	28.8	2	4,666
Maidu (Northeast)	Tsimshian (Coast)	Penutian	48	29.2	1	1,527
Ngiti	Koyra Chiini	Nilo-Saharan	47	29.8	1	4,030
Yoruba	Ijo (Kolokuma)	Niger-Congo	51	31.4	2	373
Mundari	Semelai	Austro-Asiatic	66	31.8	1	2,972
Swahili	Ijo (Kolokuma)	Niger-Congo	50	32.0	2	3,909
Maung	Yidiny	Australian	81	32.1	1	1,433
Mundari	Khmer	Austro-Asiatic	78	32.1	1	2,443
Koyraboro Senni	Murle	Nilo-Saharan	65	32.3	1	3,794
Koromfe	Ijo (Kolokuma)	Niger-Congo	49	32.7	2	1,263
Beja	Margi	Afro-Asiatic	45	33.3	1	2,591
Sango	Ijo (Kolokuma)	Niger-Congo	51	33.3	2	1,365
Nandi	Koyraboro Senni	Nilo-Saharan	47	34.0	1	4,217
Nandi	Koyra Chiini	Nilo-Saharan	52	34.6	1	4,555
Marathi	Spanish	Indo-European	52	34.6	1	7,826
Margi	Amharic	Afro-Asiatic	49	34.7	1	2,733
Mundari	Vietnamese	Austro-Asiatic	88	35.2	1	2,701
Garo	Cantonese	Sino-Tibetan	51	35.3	1	2,295
Berber (Middle Atlas)	Kera	Afro-Asiatic	65	35.4	1	3,294
Irish	Marathi	Indo-European	45	35.6	1	7,931
Paamese	Acehnese	Austronesian	45	35.6	2	8,355
Limbu	Mandarin	Sino-Tibetan	45	35.6	1	2,258
Mandarin	Bawm	Sino-Tibetan	76	36.8	1	2,151
Ijo (Kolokuma)	Diola-Fogny	Niger-Congo	46	37.0	2	2,540
Ngiti	Nubian (Dongolese)	Nilo-Saharan	54	37.0	1	1,881
Miwok (S. Sierra)	Tsimshian (Coast)	Penutian	62	37.1	1	1,807
Mundari	Khmu'	Austro-Asiatic	70	37.1	1	1,798
Bagirmi	Nubian (Dongolese)	Nilo-Saharan	64	37.5	1	1,743
Beja	Hausa	Afro-Asiatic	82	37.8	1	3,180
Koromfe	Kisi	Niger-Congo	45	37.8	2	1,173
Yidiny	Tiwi	Australian	90	37.8	1	1,701
Limbu	Meithei	Sino-Tibetan	45	37.8	2	677
Kera	Amharic	Afro-Asiatic	50	38.0	1	2,509
Zulu	Yoruba	Niger-Congo	104	38.5	4	5,035
Beja	Kera	Afro-Asiatic	57	38.6	1	2,430
Ngiyambaa	Maranungku	Australian	74	39.2	1	2,555
Malagasy	Acehnese	Austronesian	56	39.3	2	6,007
Ngiti	Nandi	Nilo-Saharan	48	39.6	1	541
Lugbara	Lango	Nilo-Saharan	53	39.6	1	252
Fur	Ngiti	Nilo-Saharan	58	39.7	1	1,471

embedded subgroup, namely Zulu and Yoruba. This particular case, however, is taken care of by the third general explanation, which is geographical separation.

Holman et al. (2007: Figure 1) show that beyond around 4,000 km there is no effect of geographical proximity on the similarity among related languages, suggesting that this is the limit of diffusion (including diffusion operating chainwise). The average distance among the language pairs in Table 4 is as high as 2,892 km, and Zulu and Yoruba have a particularly great distance (5,035 km).

In order to single out cases that need special explanations we apply the following strategy. A special case is defined as a language pair that does not satisfy at least two of the following three criteria:

1) the pair belongs to one of the world's 10 largest families;
2) the pair is genealogically separated by the root of the family tree;
3) the pair is separated by more than 4,000 km.

The special cases that fall out are: Luvale–Ijo (Kolokuma), Maidu (Northeast)–Tsimshian (Coast), Yoruba–Ijo (Kolokuma), Swahili–Ijo (Kolokuma), Koromfe–Ijo (Kolokuma), Sango–Ijo (Kolokuma), Ijo (Kolokuma)–Diola-Fogny, Miwok–Tsimshian (Coast), Koromfe–Kisi, Limbu–Meithei. We immediately notice that Ijo (Kolokuma) appears in six of the pairs. This language belongs to Ijoid, the typologically most divergent branch of Niger-Congo, among other things characterized by the absence of the otherwise characteristic noun class system. We do not have an explanation for the special behavior of Ijo (Kolokuma), but note that it does not really come as a surprise given that it belongs to Ijoid. Two Penutian pairs figure in the list, both including Tsimshian (Coast). Given that Tsimshian (Coast) has never been demonstrated to belong to Penutian to the satisfaction of all experts we may be facing a case where the languages in question are actually not related at all. As for Koromfe–Kisi and Limbu–Meithei we shall not venture to offer any explanations.

We may summarize the results of this section as follows. If two languages have 40 % or less features in common we can be almost certain that they are not generally considered to be related. Barring a few special cases, only languages that conform to at least two of the following three criteria fail to meet the prediction: they belong to one of the world's 10 largest families, they are maximally separated genealogically within their given family, and they are geographically very remote (>4,000 km) from one another.

5 Conclusions

The results reported on in this note were, in part, unsurprising and, in part, unexpected. Figure 1 showed a close correlation between relatedness and typological similarity. This is what we had expected. But we also expected to find some minimal amount of typological similarity among language pairs which would suffice to predict that two languages are related. It turned out to be the case, however, that the amount of similarity required to make this prediction is so high (81.5 %) that only few language pairs qualify. In practice, this means that typological features such as those of WALS are not useful for identifying relatedness among languages when it comes to comparisons of single pairs. When groups of languages are compared the situation may be different, but this issue is beyond the scope of this paper (cf. the thread of discussion in Dunn et al. 2005; Donohue and Musgrave 2007; Dunn et al. 2007; and Donohue, Wichmann, and Albu 2008 for an empirical example of the difficulties arising from establishing genealogical relations on the basis of abstract typological features). At the other end of the scale we found that typological *dis*similarity is a good predictor of *un*relatedness: with only a small margin of error one can predict that languages which have 40 % or less similarity are not related according to the WALS classification. Our finding that a certain amount of typological differences can be used to predict that languages are not commonly believed to be related means that typological differences are a yardstick for gauging the limits of the traditional comparative method.

Returning to the issue of *rara*, raised in the introduction to this paper, let us recall the hypothesis we stated, namely that relatedness and propinquity are the two major, systematic contributors to making languages similar. The reverse of this hypothesis is that unrelatedness and geographical separation are the major contributors to making languages *dis*similar, and therefore to promote the appearance of rare features or "exotic" language profiles. We may now consider how each of the two factors contribute to languages being dissimilar. The fact that nearly all language pairs that have 40 % or less similarities are unrelated can be generalized to a statement that unrelatedness is a necessary condition for languages to be perceived as being mutually exotic. But for languages to be maximally different from one another this condition is not always sufficient — to ensure this, they must normally also be geographically distinct. To judge from Holman et al. (2007: Figure 1), there is an overall tendency for unrelated languages to increase their similarities

by an average of some 12 % when one moves from those that are maximally separate (i. e., by 8,000 km or more) to those that are in mutual proximity. We would claim, then, that unrelatedness probably is the primary reason for the appearance of rarity in language and that geographical separation probably is the secondary reason. At the end of the day, however, these two factors will be difficult to tease apart since what we know about genealogical relations among the world's languages is severely limited by the historical-comparative methods at our disposal.

A host of other reasons would be necessary to explain why a particular language, family or language area ends up as a candidate for being particularly exotic, including universal tendencies and chance, as well as factors extrinsic to language evolution as such, i. e. issues of language sampling, choice and formulation of features, as well as variability in data attestation. Since we do not deal with such issues in the present paper our contribution is limited. Nevertheless, we believe that the issues we have treated are important and fundamental to the study of linguistic diversity and, by implication, to that of rarity in language.

Acknowledgments

We would like to thank Michael Cysouw, Bernard Comrie, Cecil Brown, and Dietrich Stauffer for comments on this manuscript.

Notes

1. In order to prevent misunderstandings let us stress that by claiming that it is not possible to quantify a certain "areality" (= diffusibility) of features we do not mean to imply that it is not possible to define linguistic areas by the occurrence of certain shared features using a quantitative approach (cf. Bickel and Nichols, in press).

References

Bickel, Balthasar, and Johanna Nichols
In press Oceania, the Pacific Rim, and the theory of linguistic areas. In: *Proceedings of the 32nd Annual Meeting of the Berkeley Linguistics Society*.

Cysouw, Michael
In press Quantitative explorations of the world-wide distribution of rare characteristics, or: the exceptionality of north-western European languages. In *Exception in*

Language, Horst Simon and Heike Wiese (eds.). Berlin / New York: Mouton de Gruyter.

Donohue, Mark and Simon Musgrave
2007 Typology and the linguistic macrohistory of Island Melanesia. *Oceanic Linguistics* 46.2: 325–364.

Donohue, Mark, Søren Wichmann, and Mihai Albu
2008 Typology, areality and diffusion. *Oceanic Linguistics* 47.1: 223–232.

Dunn, Michael, Robert Foley, Stephen Levinson, Ger Reesink, and Angela Terrill
2007 Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics* 46.2: 388–403.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson
2005 Structural phylogenetics and the reconstruction of ancient language history. *Science* 309: 2072–2075.

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

Gordon, Raymond G., Jr. (ed.)
2005 *Ethnologue: Languages of the World*. 15th ed. Dallas: SIL International. Online version: http://www.ethnologue.com/.

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

Holman, Eric W., Dietrich Stauffer, Christian Schulze, and Søren Wichmann
2007 On the relation between structural diversity and geographical distance among languages: observations and computer simulations. *Linguistic Typology* 11.2: 395–423.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker
2008 Explorations in automated language classification. *Folia Linguistica* 42.2: 331–354.

Wichmann, Søren and Eric W. Holman
2009 *Temporal Stability of Linguistic Typological Features*. München: LINCOM Europa.

Wichmann, Søren, Eric W. Holman, Dietrich Stauffer, and Cecil H. Brown
forthc. Similarities among languages of the Americas: An exploration of the WALS evidence. To appear in a volume edited by Bernard Comrie.

Language endangerment, community size and typological rarity

Jan Wohlgemuth

1 Introduction

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

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

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

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

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

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

2 Terminology and data basis

2.1 Features and characteristics

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

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

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

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

2.2 Rarity

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

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

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

2.3 Degree of rarity

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

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

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

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

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

2.4 Endangerment and community size

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

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

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

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

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

This point is taken up again in Section 4.

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

2.5 Data basis

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

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

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

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

3 Rarity distributions across small and big languages

3.1 Statistical analysis

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

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

Table 1. Rarity index level distribution of the three samples

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

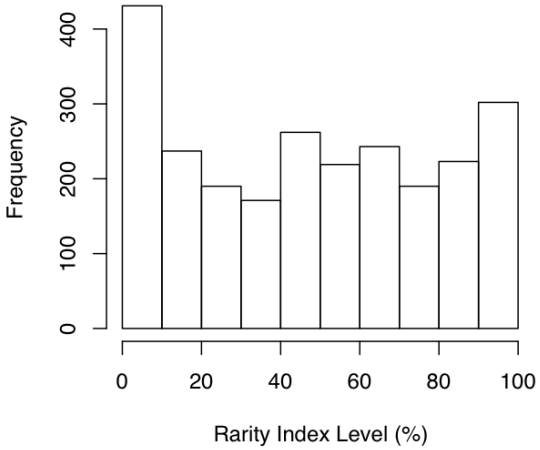


Figure 1. Distribution of the WALS languages across the rarity index level values

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

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

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

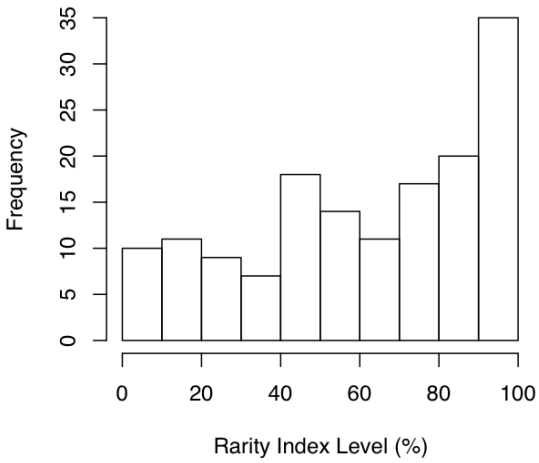


Figure 2. Distribution of the *small languages* across the rarity index level values

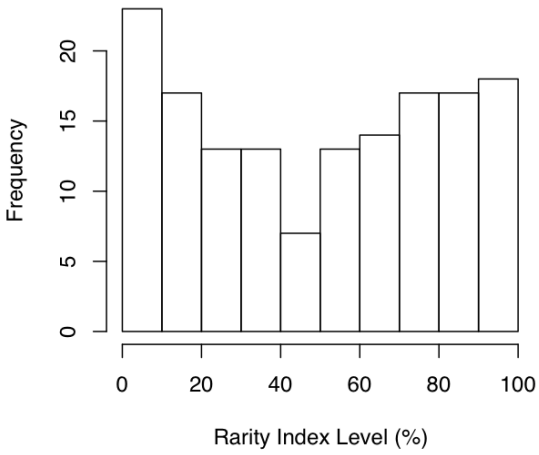


Figure 3. Distribution of the *big languages* across the rarity index level values

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

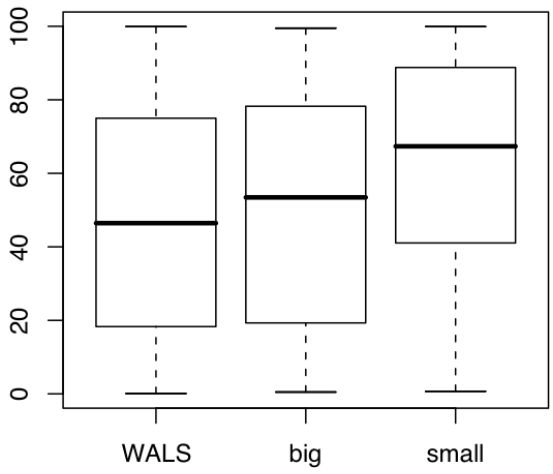


Figure 4. Comparison of the three samples' rarity index level distributions

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

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

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

Table 2. T-Test results compared

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

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

3.2 A heterogeneous picture

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

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

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

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

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

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

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

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

3.2.1 Unendangered languages with rare characteristics

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

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

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

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

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

3.2.2 Endangered / small languages without rare characteristics

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

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

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

3.2.3 Interim summary

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

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

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

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

4 Looking for an explanation

4.1 Community size

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

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

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

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

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

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

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

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

4.2 Enclave situations

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

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

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

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

4.3 Endangerment and *rara*

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

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

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

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

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

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

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

4.4 A multiplicity of factors

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

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

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

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

5 Conclusions

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

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

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

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

Acknowledgments

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

Appendix

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

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

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

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

The sample of 152 "small languages"

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

The sample of 152 "big languages"

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

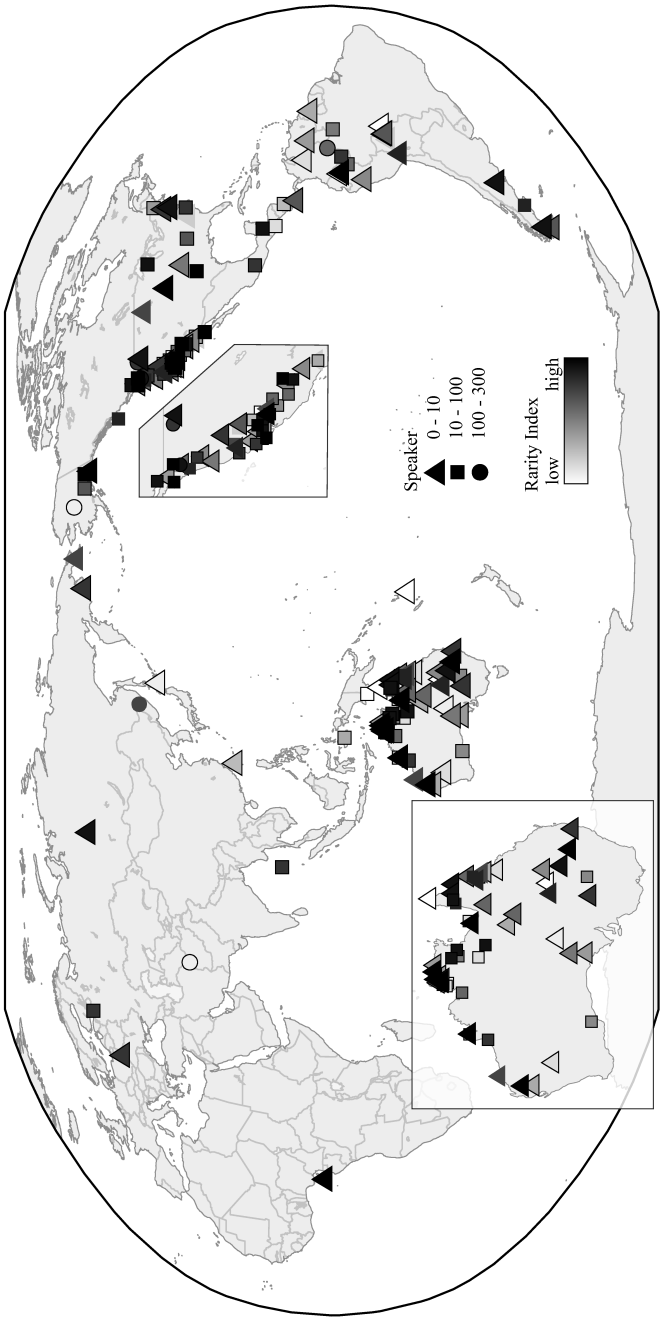


Figure 5. The "small languages" sample and its global distribution

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

Notes

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Kelkar-Stephan, Leena
2010 Future tense to express habitual past or present, and past tense to express immediate future! In *Rara & Rarissima: Documenting the fringes of linguistic diversity*, Jan Wohlgemuth and Michael Cysouw (eds.), 185–208. (= Empirical Approaches to Language Typology; 46). Berlin / New York: Mouton de Gruyter.

Krauss, Michael
1992 The world's languages in crisis. *Language* 68.2: 4–10.

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

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

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

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

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

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

Trudgill, Peter
2004 Linguistic and social typology: The Austronesian migrations and phoneme inventories. In *Linguistic Typology* 8.3: 305–320.

Wohlgemuth, Jan
2009 *A Typology of Verbal Borrowings*. (= Trends in Linguistics. Studies and Monographs; 211). 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 and Sebastian Köpl
2005 Endangered Subsystems. In *Bedrohte Vielfalt. Aspekte des Sprachentods. Aspects of language death*. Jan Wohlgemuth and Tyko Dirksmeyer (eds.), 177–186. Berlin: Weißensee.

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

Index

The index to this volume is split up into two separate indexes for reason of clarity and usability. The **Subject index** (starting below) lists all keywords and topics. Pages which contain definitions of keywords are marked by boldface page numbers. All languoids mentioned in the running text, in the examples and in the notes are indicated in the **Language index** (starting on page 283). Languoids that have different names or spelling variants of their names are always listed under the form which is actually used in the text by the respective authors.

Subject index

Language index