Matthias Urban

17. Motivation by formally analyzable terms in a typological perspective: An assessment of the variation and steps towards explanation

Abstract: This article tackles a question raised by one of the founding figures of lexical typology, Stephen Ullmann: to what degree do languages differ in the extent to which they resort to morphologically analyzable lexical items? Drawing on a worldwide sample of 78 languages for which a standard set of 160 mostly nominal meanings is investigated, the article demonstrates that variability in this area is indeed profound. Correlations between the relative prevalence of analyzable items in a language with the size of its consonant inventory, the complexity of its syllable structure, and the length of its nominal roots suggest that, typologically, languages with a simple phonological structure are those in which analyzability in the lexicon is most profound. Possible explanations for this observation in terms of the avoidance of homonymy and pressure exerted by different linguistic subsystems on each other are discussed.

17.1 Introduction

Following current definitions (Koch 2001; Koch and Marzo 2007), lexical motivation is a property of a lexical item which shows a formal relation to one or more other lexical items that mirrors a conceptual relation between the concepts that they respectively denote. Word-formation is an important motivational device. The French *poirier* ‘pear tree’, for example, is motivated by the formal and semantic relation to *poire* ‘pear’, from which it is derived. But lexical motivation also includes other kinds of complex items which establish this double relation, as well as the extreme case of polysemy, in which the formal relation is one of complete identity. Both aspects of lexical motivation – the formal and the conceptual – are worthwhile topics for cross-linguistic research.

Regarding the former, the question of differences between languages in the *quantity* of motivated items in their lexicon has been a major concern of research

Matthias Urban (Leiden University)

in what is coming to be called lexical typology. Saussure ([1916] 1967) raised this question early on (he introduced, alongside the famous notion of the arbitrariness of the linguistic sign also that of relative motivation), as did Ullmann (1962, 1966) later. Much more recently, Koch and Marzo (2007: 273) ask, but do not answer the question "are there more or less formally transparent languages"? The issue concerned still earlier writers as well, see Urban (2012: chapter 2) for review.

Ullmann (1962: 105) was aware of the difficulties in unambiguously identifying and quantifying polysemy, and suggested restricting oneself to motivation by morphological analyzability in a quantitative study, as in the case of French *poirier*:

> With morphological motivation one is on firmer ground: it is the most clear-cut and least subjective of the three types, and certain broad tendencies stand out very clearly...

Later on, Ullmann made some casual methodological suggestions for such an investigation (1966: 223):

> It might be possible to devise some statistical test for these relative frequencies. Such a test might be based on samples from dictionaries, on a representative selection of texts, or on both.

Scattered statements in the literature suggest that cross-linguistic variability in the prevalence of motivated analyzable terms in the lexicon is indeed profound. It is thus a typological variable of great interest which has not yet been investigated systematically in spite of suggestions such as Ullmann's. For instance, Seiler (1976: 6) says about Cahuilla that "[t]he analysability and morphological transparency of a considerable portion of all nominal expressions [...] is immediately recognisable", and O'Meara and Bohnemeyer (2008: 332–333) even state for Seri that "[c]omplex expressions [...] are in fact pervasive in the Seri nominal lexicon" and that the rarity of unanalyzable terms is a "general typological characteristic of the nominal lexicon of Seri".

This paper reports on an investigation very similar to that suggested by Ullmann. It was carried out applying methods of modern linguistic typology, a discipline that has grown immensely since Ullmann's times. As Ullmann suggested, it is restricted to lexical motivation by morphological analyzability, excluding polysemy. Details of the approach and a first description of the cross-linguistic variation in the domain of analyzability in the lexicon follow in section 17.2. However, even more interesting than assessing the mere distribution of the differential degrees of analyzable terms in the languages of the world is to ask why this distribution is as it is, i.e. to try to understand why lexical motivation is present to a smaller or larger degree in different languages. Section 17.3

describes a number of factors which appear to be relevant, and a final discussion appears in section 17.4.

17.2 Approach and data

The present approach makes use of a list of 160 mostly perceptually apprehensible "nominal" concepts (see Appendix A), where it is assumed that they possess properties (most prominently, stability of meaning independent of contextual factors) that make them easier to compare across languages than event-denoting "verby" expressions (cf. Cruse 1986: 152; Foley 1997: 35). The concepts are organized into four semantic domains: terms for natural kinds, artifacts, body-parts and body-liquids, and terms for phases of the day plus a few miscellanea. There is no direct predecessor to the list, though it was partly inspired by works such as Buck (1949) in the domain of nature-related terms and Brown (1999) in that of artifacts. Here, the latter include both items of some antiquity in most cultures (e.g. 'knife') as well as more recent items of acculturation (e.g. 'car') to also take into account the behaviour of languages when it comes to denominating new stimuli. Regarding body-part terms, which have been rather well studied from a cross-linguistic point of view, care was taken that parts are included that have hitherto received relatively little attention. Terms for the meanings were gathered from extant sources and/or were provided by experts for seventy-eight languages (see Appendix B), each of which belongs to a different language family recognized in Dryer (2005). For each language, the criterion for inclusion was that counterparts for more than 104, or 65 %, of the meanings on the list were available.[1] The assumption is that the mapping from meaning to form is many-to-many (cf. Haspelmath and Tadmor 2009): there may be several unrelated items in a given language corresponding to a single meaning (synonymy or, more commonly, near-synonymy), and, conversely, there may be a single equivalent covering the range of two or more of the meanings on the list (polysemy, vagueness, homonymy, or, to use a deliberately ambiguous term with respect to this distinction for the application in cross-linguistic studies coined by François 2008, 'colexification'). Rather than trying to single out the "best" equivalent, several

[1] This procedure leads to a strong representation of languages of the Americas. As Dahl (2008) suggests that they have been underrepresented in many previous typological samples, this is in principle desirable; nevertheless, it may be the case that they are in fact overrepresented in the present sample. Since the sample was not manipulated in hindsight, this should be borne in mind when contemplating the evidence presented here.

corresponding terms were accepted per language for a given referent. Since these may, of course, have differing formal properties (one may be analyzable, another morphologically simple), this entails, perhaps paradoxically at first sight, that in the present analysis a language may transpire to have 0.5, 0.33, etc. analyzable terms for a given meaning. The sum of these values for individual meanings in one language yields the absolute analyzability score. Observe that "analyzability" in this sense neither presupposes nor necessarily implies ability on behalf of native speakers to decompose the terms into their parts, although it would be worthwhile or even preferable to take speaker judgments into account. Further, since equivalents for all meanings could not be retrieved for all sampled languages, the resulting figure was divided by the number of meanings for which data are available. This, finally, yields the relative analyzability score for each of the sampled languages, and this is the variable which is discussed in the following sections. However, alongside merely being registered, analyzable terms were also classified into three broad types, illustrated here with some Bezhta examples from Comrie and Khalilov (2009): (i) the lexical type, which involves more than one lexical root (e.g. that for *häyš ƛ’äq’e* ‘eyelid’, which consists of the word for ‘eye’ in the genitive case and ‘roof’); (ii) the derived type, characterized by presence of a single lexical root (e.g. *ƛišiyo* ‘waterfall’, which is in fact the past participle of a verb meaning ‘to become entangled’); and (iii) the rare alternating type, in which senses are distinguished by some kind of grammatical alternation (e.g. *häydä* ‘glasses’, the plural of ‘eye’). Not taken into account were analyzable terms composed of morphology that does not have a motivating force and semi-analyzable terms.

17.3 The cross-linguistic variation in analyzability

Purely descriptively, the investigation confirms the initial conjecture as to the cross-linguistic variability in the relative number of analyzable items: the range of their number is from very few analyzable terms for the investigated concepts, as in Aymara, which receives a relative analyzability score of 4.9 %, to as large a score as 50.2 % for Kiliwa.[2] Thus, Kiliwa has *ha?kw?nymarkwiy* ‘cloud voice’ where English, for example, has the unanalyzable *thunder*, *wa?hkapu?* ‘house

[2] What is of interest are probably not the absolute percentages, since these depend to some degree on the concepts one investigates, but rather the fact that there is variability when semantics, by way of using a standardized list of concepts, is kept constant.

opener' where English has *key*, *nymayuyuw* 'breast eye' where English has *nipple*, *khwathyuul* 'flowing blood' where English has *vein*, and so on (Kiliwa data and literal translations from Mixco 1985). The map in Figure 1 plots the result for the investigated languages onto a map of the world. Each dot represents a sampled language, and the size of the dot corresponds to the relative analyzability score: the larger the dot, the higher the score, the smaller the dot, the lower the score.

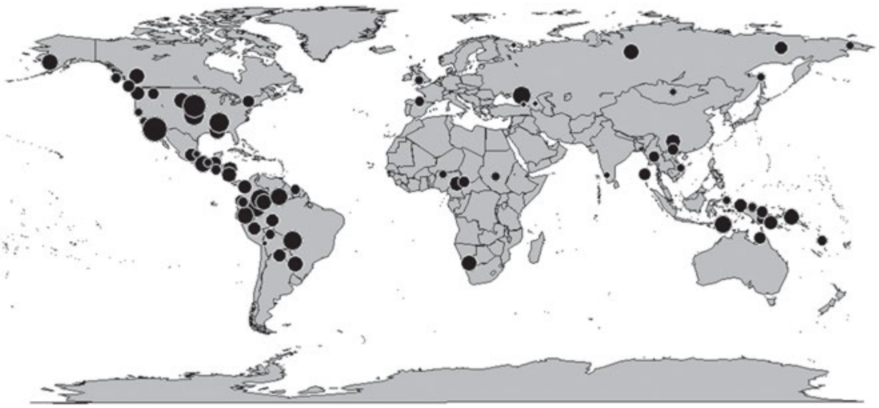

Figure 1: Differential degree of analyzability in the sampled languages

From eyeballing the map, one can identify some geographical hotspots in which languages with a high number of analyzable terms are frequent, such as Eastern North America and the lowlands of South America. However, simple visual inspection of distributions on maps is an unreliable technique for assessing areality (Cysouw 2005, among others). Using a standard breakdown of the world into six macro-areas (Africa, Australia-New Guinea, Eurasia, Oceania, North America, South America, from Dryer 1992), there is some evidence for areality, but no clearly statistically detectable difference between the areas (p = .0712 by a Kruskal-Wallis rank sum test, $\chi^2 = 10.1461$, df = 5). Given that areal convergence thus does not appear to be a decisive factor in governing the behaviour of individual languages, the question that immediately emerges is: what is? Possible factors are carved out in the following section.

17.4 In search of conditioning factors

17.4.1 Rationale

A priori, the question is entirely open, but also lends itself to empirical investigation. Prevalence of analyzability in the lexicon as a typological trait is a variable that has not been addressed previously in a systematic fashion, and hence there is no literature on which to base new hypotheses. Therefore, a series of preliminary hypothesis-generating statistical tests on the basis of the entire set of 142 features in the *World Atlas of Language Structures* (Haspelmath et al. 2005), which deal with a diverse range of phonological, morphological, syntactic and lexical topics, was run in the statistics environment R (R development core team 2009). This test series suggested an influence of two of the features dealing with phonology on the relative analyzability score, namely consonant inventory size (Maddieson 2005a) and syllable structure (Maddieson 2005b), among other features. More precisely, as consonant inventories became larger and syllable structures more complex, the number of analyzable terms among the meanings investigated decreased. Since the overlap between Maddieson's and the present sample was quite small, additional data from published sources were gathered for the languages of the present sample in order to assess whether the dependency could be substantiated, while maintaining Maddieson's general coding schemes. Because of errors in Maddieson's (2005a) data, they were later updated taking into account changes effectuated in Maddieson (2013).[3]

Moreover, for the final analysis, it is not only important to have as complete datasets as possible, but also to control for areal factors, as usual in modern typology. For the present topic, this is even more imperative since there clearly are some areal differences in analyzability (although insignificant), but also because, as noted by Maddieson (2005a, b), the cross-linguistic distribution of the phonological features is highly skewed (for the sake of illustration, one can think of the large consonant inventories of languages in the American Northwest transcending genealogical boundaries, Mithun 1999: 314–315). In order to ascertain whether the significance of the preliminary tests is spurious because of areal

[3] Languages for which values have been changed are Arabela (from moderately large to small), Guaraní (from average to moderately small), and Great Andamanese (from small to average). The erroneous value for Oneida in Maddieson (2005) had already been noted and corrected in Urban (2012). In addition, the value for Kildin Saami was changed from moderately large to large (as per Riessler and Wilbur 2007: 74) and that for Bezhta from large to moderately large (as per Zaira Khalilova p.c.).

influences, generalized linear mixed models were built in R for the two candidate factors. This type of statistical analysis is increasingly used in a variety of disciplines, including psycholinguistics, where it is important to generalize over different participants in order to rule out that the results of an experiment are biased or even spurious due to the unusual behaviour of (few) individual test subjects. For this purpose, mixed models include two basic types of variables: so-called fixed effects, which are generally those variables of interest, over which the experimenter typically has control, and which s/he hypothesizes to be relevant for the behaviour of the response variable, and random effects, over which the experimenter has no control and for which generalizations are not of interest generally (such as the individual subjects participating in an experimental study, the particular animals a biologist observes to make generalizations about the species, etc.).

Just as in an experimental test setting one wants to generalize over the behaviour of different participants, in typology one wants to generalize over the behaviour of languages in different linguistic areas. Hence, mixed models were constructed with the relative analyzability score as the response, the phonological features of interest included as fixed effects and linguistic macro-area (in the breakdown of Dryer 1992) as a random effect (see also Cysouw 2010 for an approach to typology and the question of controlling for area using generalized linear mixed models).[4] Code by Baayen (2009) and Bates and Maechler (2009) was used for the analysis. Initially, models involving both a random intercept (meaning, in this case, that the relative degree of analyzability is allowed to vary from area to area) and random slopes components (meaning that the impact of the phonological properties may vary from area to area as well, being stronger in some regions of the world and weaker or even nonexistent in others) were built. On the basis of these models, assumptions of mixed models (normality and homogeneity of residuals) were checked by visual inspection of histograms of the residuals and plots of fitted and residual values. When visual inspection left doubts as to whether the assumptions are fulfilled, an additional Shapiro-Wilk test for normality of the residuals and a correlation test between fitted vs. residual values

[4] An issue with this statistical technique, powerful as it is, is that it is not well suited for the classical task of linguistic typology of making inferences about all possible human languages, including all those spoken in the past but vanished today. This is because the linguistic diversity encountered today represents only a small fraction of what may be possible due to historical contingencies, and any statistical inference is thus necessarily based on this fraction alone. For this reason, all generalizations arrived at in this article pertain to the present-day linguistic diversity, but not to all possible human languages (cf. Cysouw 2010: 258fn5 for similar cautionary remarks).

were carried out. To simplify model structure, the random slopes component was subsequently removed if a likelihood ratio test comparing the full model with a reduced model only involving random intercepts indicated that random slopes are not required. This was the case for all models. Finally, further likelihood ratio tests were carried out to compare the resulting simplified models including the fixed effects with reduced models only including the random effect.

Given that the preliminary test battery on the WALS data consisted of 142 tests, one would expect the emergence of spurious significance simply by chance at an α-level of .05 in the case of 142 × .05 ≈ 7 of the tests. Therefore, a smaller validation sample using additional data in Urban (2012) was constructed to assess whether the result can be replicated using the same mixed model design.

17.4.2 Consonant inventory

Figure 2 is a boxplot[5] showing the effect of the size of the consonant inventory on the relative analyzability score (not simultaneously visualizing areal effects to maintain easy readability).

[5] Boxplots are a useful visualization technique for statistical distributions. Here, the y-axis shows the relative number of analyzable items, and each of the boxes corresponds to one of the levels of the phonological features, as coded by Maddieson (2005a, c). The thick black line within each group represents the median for the relative analyzability score within that group, and the size of the boxes and the dashed lines (the so-called whiskers) indicate the variance around that mean: the smaller the boxes and whiskers, the smaller the variance around the mean, the larger, the greater the variance. Generally, 50 % of datapoints in each group fall into the box. Individual dots above or below the boxes represent outliers, that is, individual languages which are very far removed from the median of the group they belong to. Finally, the width of the boxes gives an idea of the number of observations within each group: the narrower the box, the smaller the number of observations (that is, languages in the sample having a particular phonological property such as an average-sized consonant inventory), the wider the box, the larger the number of observations.

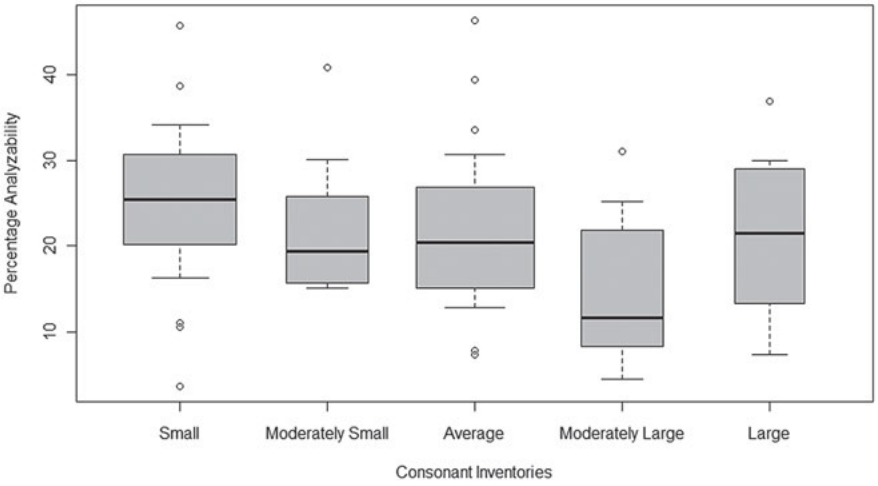

Figure 2: Relative degree of analyzability depending on size of the consonant inventory. Box width indicates number of data points within a category.

As the figure shows, there is a downward trend in the relative analyzability score as consonant inventories become larger, though this effect is somewhat uneven between groups and, surprisingly, the languages with the largest consonant inventories behave in an unexpected way. In the mixed model design, the size of a consonant inventory is relevant as a factor and the *p*-value (estimated by Markov Chain Monte Carlo (MCMC) simulation with 100,000 replicates) associated with the predictor itself is weakly significant at .04727.

In the validation sample, it was also the case that the relative analyzability score was lower for languages with large consonant inventories compared to those with small ones, but the impact of consonant inventories as a predictor was less clear. Together with the only weak significance of the main model, it transpires that the connection needs further attention to be fully accepted as valid, and hence its identification as a relevant factor here is preliminary only.

17.4.3 Syllable structure

Figure 3 is a boxplot showing the relative degree of analyzability depending on complexity in syllable structure. There is a very similar dependency here: languages with simple syllable structure (i.e. no consonant clusters and no consonants in coda position) tend to have a higher value for the relative analyzability score than do those with moderately complex syllable structure (allowing a consonant in coda position, and initial clusters of consonant plus glide), which in turn tend to score higher than those with complex syllable structure (i.e. allowing for more elaborate clusters).

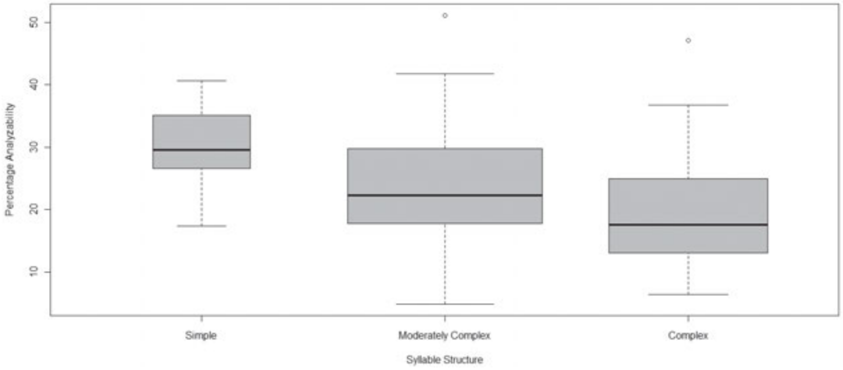

Figure 3: Relative degree of analyzability depending on complexity of syllable structure. Box width indicates number of data points within a category.

As for consonant inventories, the impact of the differences in syllable structure is significant (MCMC-estimated *p*-value = .0102), and the effect could be replicated on the basis of other data in the validation sample.

These two pieces of evidence, taken together, are able to account for the behaviour of many individual languages and areal differences:[6] the difference between Western and Eastern North America mentioned above corresponds to a basic asymmetry in phonological complexity, in particular pertaining to size of the consonant inventory between these two parts of the continent (Sherzer 1973). This also may correlate with the fact that Polynesian languages, famous for their small number of consonants (and having a simple (C)V syllable structure), have

[6] Given that these two factors are relevant, it is natural to wonder whether the other major variable in complexity of phonological systems, namely the size of the vowel inventory, has an impact as well. Therefore, data from Maddieson (2005c) were amended for the languages of the present sample as well. In fact, when plotting the relation, the result looks very similar: as vowel inventories become smaller, mean values of the relative analyzability score rise. However, when taking into account areal factors by including area as a random effect in a mixed model design, there is no appreciable difference made by the factor vowel inventory (*p* = .5896), indicating that areal skewings play, unlike the other two investigated factors, a major role here. This, of course, underlines the need to control for areal influence in typology to rule out spurious results.

on average a higher number of analyzable lexical items when compared with their Austronesian kin.

17.4.4 Root structure

However, there is residual variation in the degree of analyzability that remains puzzling: for instance, all indigenous language families of the Caucasus sampled, namely Northwest Caucasian (represented by Abzakh Adyghe), Kartvelian (represented by Laz), and Nakh-Daghestanian (represented by Bezhta) are well-known for having a large number of distinctive consonants, and also allow for complex syllables. Yet, Abzakh Adyghe scores very high with respect to the analyzability score (in fact, it has the highest value for Eurasia as a whole), while Laz and Bezhta receive very low scores, which is all the more puzzling since Northwest Caucasian consonant inventories are typically even larger when compared to those from the other two Caucasian families. So, if anything, one would expect the situation with regard to analyzability to be the other way around in light of the global dependencies identified by statistical analysis. This suggests that there is at least one further, as yet undetected, factor at work so far. Comparative discussion of the structure of lexical items in Caucasian languages in Rayfield (2002: 1041) provides a clue as to what that factor may be for the variation encountered in the Caucasus specifically. Unlike Kartvelian (at least in the nominal domain) and Nakh-Daghestanian,

> Abkhaz and Circassian [=Adygheian, MU] contrast a prodigious wealth of consonants with a paucity of vowels and strict limits on permissible syllable structure. Roots tend to be monosyllabic, sometimes mono-consonantal, consequently with many homophones. Consonants in initial position rarely occur in clusters of more than two, and there are a very limited number of such clusters... As in, say, Chinese, the number of acceptable syllables that can constitute a root morpheme in N.W. Caucasian roots is so small that, in order to express a wide number of concepts or to name, say, flora and fauna, specific lexemes have to be constructed by recombining two or more other lexemes, or otherwise monosyllabic lexemes are polysemantic.

Note that root structure is a different variable than syllable structure: as the Northwest Caucasian case shows, allowing for complex syllables does not necessarily mean that they occur with high frequency in the lexicon, and conversely, simple syllable structure does not correspond directly to short roots, as they may be made up of several syllables. There are further statements on languages with a relatively high degree of analyzable lexical items which corroborate the suspicion that the typical phonological structure of the lexical root is another relevant factor, not only in shaping diversity in the Caucasus, but also operating more generally. Werner (1997: 46) and Watkins (1984: 75) state for Ket and Kiowa respectively that roots are typically monosyllabic, with the disyllabic roots attested usually being identifiable as old lexicalized compounds.

Unfortunately, this emerging hypothesis is not easily testable, because for the majority of sampled languages no explicit discussion of typical root structure shapes is available in the literature. Consequently, the following provisional method was used: the number of syllables was counted for each of the unanalyzable lexical items in the data for the present study, and subsequently, the weighted mean was calculated to give an idea of the average structure of the lexical word generally. This is not always easy, since the data at hand are represented orthographically, requiring one to often infer phonology from orthography. A particular issue in this respect is the question as to whether sequences of vowels should be treated as diphthongs or be syllabified as nuclei of separate syllables, since this may heavily influence the resulting figures regarding the number of syllables. For instance, in Toaripi, sequences of up to five vowels are frequent, and any arbitrary decision as to their phonological status would greatly influence results in one way or another. Luckily, for this task in general, as well as for the problem of syllabification of vowel sequences, primary descriptions of the languages are often of help. However for nine sampled languages: Mali, Rotokas, the aforementioned Toaripi, Kildin Saami, Cheyenne, Arabela, Cayapa, Chayahuita, and Cubeo, sequences of orthographic vowels are highly frequent, and their proper interpretation remains unclear; hence, for this specific task, they were excluded from analysis entirely.[7]

Of course, the lexicon is vast, and the typical structure of the root is assessed only on the basis of a very small subset here; however, where statements on the typical root structure are made in the literature on the languages, the figures obtained for the present study are typically in agreement.

The resulting weighted means for the remaining sixty-nine languages were then included in a mixed model design as a fixed effect, area as a random effect, and the relative analyzability score as the response to be modelled. Root Structure had an impact on this response to a significant degree (MCMC-estimated p-value = .0355). This impact on the degree of analyzability is plotted in figure 4. For the purpose of visualization, root length in terms of syllables was divided into four

[7] For other languages where orthographic vowel sequences exist but are less frequent, they were treated in a way that biases against the hypothesis: for languages where the analyzability score is below the cross-linguistic mean of 22.81 % they were analyzed as diphthongs, and for languages above the mean as sequences.

groups: short, moderately short, moderately long, and long (but for modelling, the actual, more informative values were used).

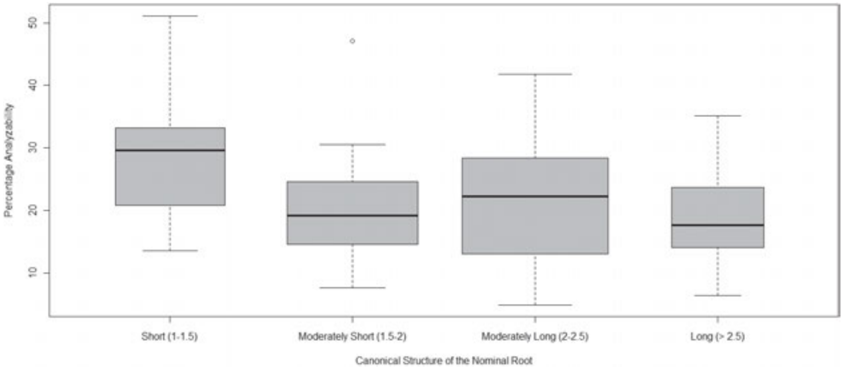

Figure 4: Relative degree of analyzability depending on length of nominal roots. Box width indicates number of data points within a category.

17.4.5 Summary

Thus, the answer to Koch and Marzo's (and Ullmann's) question seems to be: yes, there are languages with a high degree of formal transparency, that is a high relative analyzability score. These languages tend to possess simple phonological systems with regard to their syllable structure and – tentatively – consonant inventories as well as comparably short lexical roots. This by no means excludes the possibility that there can be further factors at play, such as large-scale contact-induced lexical replacement which enriches the lexicon with unanalyzable loanwords (see Urban 2012 for discussion of possible further factors). Nevertheless, phonological restrictions do seem to exert some structural pressure on the lexicon, causing it to adapt by an increased number of analyzable lexical items. Ross (1980) is a revealing case study on a language outside the present sample showing this pressure: the phonological system of Vanimo allows for the generation of as few as 960 distinct morphemes, and this is counterbalanced by the ample use of complex expressions making up for the paucity of possible phonologically distinct morphemes.

17.5 Discussion

It is important to realize that significant correlations are not explanantia in themselves, but rather explananda, and thus the question that one needs to pose is: why are the correlations there in the first place? By addressing this question, one enters the final stage in the explanation of interdependencies between typological variables in terms of Bybee (1988): empirical generalizations were made concerning an apparent influence of (at least) three factors on the degree of relative analyzability, then a principle was formulated that summarizes the empirical generalizations (simple phonology and root structure entails high analyzability in the lexicon), and finally, an explanation for the operation of the principle has to be identified.

A number of recent works in linguistic typology, Bybee (1988) among them, emphasize that typological distributions and universal tendencies in language structure have to be understood from the historical factors that bring them about.

Rayfield (2002: 1041), as quoted above, comments on the high incidence of homophony in Northwest Caucasian, caused by the severe restriction of possible root shapes. In fact, there is a venerable, although not unproblematic, principle in historical linguistics dating back to Gilliéron and Rocques (1912): homonymy avoidance. According to this principle languages (or rather, their speakers) take countermeasures against the possibility of detrimental effects of homonymy or near-homonymy to successful communication by ousting certain cases of homonymy from the lexicon. Case studies include, but are by no means limited to Williams (1944), Campbell (1975), and Dworkin (1993), see Urban (2012) for more thorough review.[8] Malkiel (1979: 2–3; 7) provides a typology of the potential outcome of homonymic clashes: next to simple “peaceful” continued coexistence, one homonym may oust another, they may merge, or differentiate in form and possibly in meaning.

While there is ample literature on lexical loss or irregular sound change due to putative homonymy avoidance, empirical evidence for coining neologisms for this purpose, which would be needed to make a convincing case for homonymy avoidance as an operative factor, is quite sparse in the literature, although not nonexistent. Shi (2002: 76) states that, as the phonological system of Chinese simplified considerably over the past 1,000 years or so, one way to bring about the disyllabification of the lexicon well-known to Sinologists is the replacement of inherited terms by two-syllable (and hence, morpheme) compounds. Coates

[8] And note that quite to the contrary there may also be language change that creates rather than wards off additional homonymy, see e.g. Dixon (2004: 71) on Jarawara.

(1968) is a case-study from Germanic that demonstrates how later phonological collapse of erstwhile distinct Proto-Germanic **þǐhstila* ‘thistle,’ **þinhslā* ‘pole, beam, tongue’ and **þehsalōn* ‘adze’ (an old tool for wood processing) caused replacement of one or another inherited term by a newly coined compound in some daughter languages.

In addition, there are theoretical concerns regarding homonymy avoidance as a functional principle in diachrony. Some elaborations suffer from an undue personalization of language as a deliberately acting agent, neglecting the role of speakers as instigators of innovations. But even if a decisive role of the speaker in language change is acknowledged, it remains questionable whether such speakers actually produce innovations (such as complex neologisms) with the explicit goal of changing their language. Their motives for innovation may well be very different. Once such innovations have occurred, however, it is still arguable that because of certain properties which some innovations possess they may have an advantage leading to their propagation across a speech community (cf. Koch 2005: 233–236; 238–242 and references therein; for the distinction between innovation and propagation in language change see also Croft 2000: 4–5). Enhanced distinctiveness vis-à-vis a possibly confusable homonym or near-homonym may well be such an advantage.

Thus, without postulating a principle stating that languages generally abhor homonymy, a possible avenue of explanation is to assume that in languages with simple syllable and root structure and perhaps small consonant inventories, these bring about restrictions on the possible number of phonologically distinct lexical roots causing lexical homonymy at a rate high enough to lead to possible confusion in communication.

However, issues remain: first, as noted by most authors writing on the topic, for homonymy avoidance to be a plausible explanatory factor for diachronic changes, the relevant lexical items need to be in danger of co-occuring in the same stretch of discourse. Otherwise, there is no actual danger of confusion in communicative events. Second, its pervasiveness as a principle in historical linguistics generally is disputed.

Therefore, also in light of the sparse evidence in the literature for coinage of complex terms for the purpose of homonymy avoidance, it seems worthwhile to consider taking a broader perspective. To reiterate, the evidence resulting from the present study suggests an influence by syllable and root structure on the overall linguistic system causing it to exploit word-formation devices to a larger extent than those with ample phonological resources. A more abstract line of reasoning would therefore be to hypothesize that as the number of actually lexically exploited morpheme shapes approaches that of the possible shapes that can be generated by the phonological system (see Krupa 1966 for a quantitative study on Maori), there is pressure on the linguistic system to counter the limited expressive possibilities, either by the introduction of phonemic tone (Matisoff 1973; see Urban 2012 for discussion of tone as a relevant factor), the introduction of analyzable lexical items for an increased range of concepts, a combination of these, or yet another strategy. In fact, Nettle (1995, 1998) makes quite similar observations, although his datasets are either much smaller (Nettle 1995) or more geographically restricted (Nettle 1998). While not concerned with morphological analyzability, but rather with word length in terms of segments, he establishes a similar inverse relationship with phonological complexity: languages with many phonemes have shorter words than those with few (though note again that in the present study the influence of the size of the consonant inventory is not straightforward). Obviously, complex terms terms are longer segmentally than the elements they consist of, so the two results are fully compatible with one another. Furthermore, the concluding discussion in Nettle (1998: 244) similarly suggests that "lexical expansion" by the coinage of complex terms is responsible for the correlations observed. Nettle (1999: 144) summarizes:

> as a result [of speakers' tendency to underarticulate driven by economy, MU], sets of words that were previously distinct become homophones. When words have become homophones, speakers may have to compensate by some kind of lexical strategy, such as coining a new word or paraphrase. ... Discrimination failure leads to smaller inventories, and the lexical strategies by which meaning is maintained tend to produce longer word forms. The pressure on the language from discrimination failure thus precisely balances that due to articulatory economy. The actual system of any given language emerges from a dynamic equilibrium between these two factors.

In this sense, the present evidence can be read as a variation on the old theme of speakers being suspended between the drive towards economical linguistic behaviour on the one hand and the necessity to attain communicative efficiency on the other, present from pre-Structuralist thinking (e.g. Gabelentz 1901), through French structuralism (e.g. Martinet 1952), up to present-day linguistic theorizing (e.g. Haspelmath 1999).

Whichever explanation one prefers, the evidence presented here may provide an implication for current linguistic theory: if one is willing to view analyzability in the lexicon, that is morphological complexity in lexical items, as a type of linguistic complexity as much discussed recently (Miestamo, Sinnemäki, and Karlsson 2008; Sampson et al. 2008), this can be construed as evidence for a tradeoff between complexity in linguistic subsystems (compare the 'equi-complexity axiom'): phonological simplicity tends to go hand in hand with complexity in lexical items and vice versa (although "complexity" in this sense is subject to differing definitions and not an entirely clear-cut concept, see Miestamo 2008 for an overview of different approaches).[9]

17.6 Acknowledgments

My sincere thanks go to Johanna Mattissen for providing lexical data for Laz, to Tonya Stebbins and Julius Tayul for making available a pre-print version of their Mali dictionary, and to Joseph Atoyebi, Ekaterina Gruzdeva, Andrej Nevedov, Pamela Munro with Catherine Willmond, and Frank Seifart for checking and amending the Yoruba, Nivkh, Ket, Chickasaw, and Bora data respectively. Further, I thank Bodo Winter for first drawing my attention to mixed models, and am indebted to him, Michael Cysouw, and Roger Mundry for continued advice on statistical matters. I also wish to thank the reviewers for this paper and the editors of the present volume for their input and suggestions, as well as Kate Bellamy for proofreading. Responsibility for shortcomings rests entirely with me.

17.7 Appendix A: list of meanings

I. Nature-Related and topological concepts: 1. Animal, 2. Ashes, 3. Bark, 4. Bay, 5. Beak, 6. Bird, 7. Bloom (blossom, flower), 8. Branch, 9. Bud, 10. Cave, 11. Clearing, 12. Cloud, 13. Coal, 14. Coast, 15. Dew, 16. Dust, 17. Eclipse, 18. Egg, 19. Embers, 20. Estuary, 21. Feather, 22. Flame, 23. Flood, 24. Foam, 25. Fog/Mist, 26. Forest, 27. Gold, 28. Grass, 29. Headland, 30. Honey, 31. Horizon, 32. Horn, 33. Lagoon, 34. Lake, 35. Lightning, 36. Meteoroid (shooting/shining star), 37. Milk, 38. Milky Way, 39. Moon, 40. Mountain, 41. Mushroom (fungus), 42. Nest, 43. Plant, 44. Puddle, 45. Rain, 46. Rainbow, 47. Resin, 48. River/stream, 49. river bed, 50. Root, 51. Seed, 52. Shadow, 53. Sky, 54. Smoke, 55. Soil, 56. Spark, 57. Spring/Well, 58. Star, 59. Steam, 60. Straw, 61. Sun, 62. Swamp, 63. Tail, 64. Thorn, 65. Thunder, 66. Tree, 67. Valley, 68. Volcano, 69. Waterfall, 70. Wave, 71. Wax, 72. Whirlpool

II. Artifacts, 1. Airplane, 2. Ball, 3. Bed, 4. Belt, 5. Boat, 6. Car, 7. Chair, 8. Clock, 9. Glasses, 10. House, 11. Key, 12. Knife, 13. Ladder, 14. Mirror, 15. Needle, 16. Paper, 17. Pen, 18. Rope, 19. Scissors, 20. Shoe, 21. Road/Street/Way, 22. Table, 23. Toilet, 24. Train, 25. Weapon, 26. Window

[9] There is neuropsychological evidence summarized in Libben (2006) that in the processing of compounds, both constituent parts are neurally activated, even if the semantic relation between their meaning and the compound meaning is non-transparent. This indicates that, indeed, their cognitive representation is more “complex” than that of simplex lexical items.

III. Body Parts and Body Fluids: 1. Adam's apple, 2. Ankle, 3. Beard, 4. Belly/Stomach, 5. Bladder, 6. Blood, 7. Bone, 8. Brain, 9. Breast, 10. Buttocks, 11. Calf, 12. Cheek, 13. Chin, 14. Eyeball, 15. Eyebrow, 16. Eyelash, 17. Eyelid, 18. Finger, 19. Fingernail, 20. Guts, 21. Heart, 22. Jaw, 23. Kidney, 24. Lip, 25. Liver, 26. Lungs, 27. Mouth, 28. Mucus, 29. Navel, 30. Neck, 31. Nipple, 32. Nostrils, 33. Pupil, 34. Pus, 35. Rib, 36. saliva/spittle, 37. Scar, 38. Skin, 39. Snot, 40. Semen, 41. Sweat, 42. Tear, 43. Tendon/Sinew, 44. Testicle, 45. Tongue, 46. Tooth, 47. Urine, 48. Uvula, 49. Vein, 50. Womb, 51. Wrinkle

IV. Basic Temporal Concepts and Miscellanea, 1. Dawn, 2. Day, 3. Dusk, 4. Night, 5. Noon, 6. Sunrise, 7. Sunset, 1. Man (human being), 2. Saturday, 3. Virgin, 4. Widow

17.8 Appendix B: Sample Languages, ordered by macroarea

Note: Full references as well as data on phonological features is in Urban (2012).

I. Africa: 1. Hausa (Afro-Asiatic), 2. Katcha (Kadugli), 3. Khoekhoe (Khoisan), 4. Mbum (Niger-Congo), 5. Ngambay (Nilo-Saharan)

II. Australia-New Guinea: 1. Baruya (Trans-New-Guinea), 2. Berik (Tor), 3. Buin (East Bougainville), 4. Kaluli (Bosavi), 5. Kwoma (Sepik), 6. Mali (Baining-Taulil), 7. Meyah (East Bird's Head), 8. Rotokas (West Bougainville), 9. Sahu (East Papuan), 10. Toaripi (Eleman), 11. Yir Yoront (Australian)

III. Eurasia: 1. Abzakh Adyghe (Northwest Caucasian), 2. Badaga (Dravidian), 3. Basque (Basque), 4. Bezhta (Nakh-Daghestanian), 5. Chukchi (Chukotko-Kamchatkan), 6. Ket (Yeniseian), 7. Khalkha (Altaic), 8. Laz (Kartvelian), 9. Nivkh (Nivkh), 10. Kildin Saami (Uralic), 11. Welsh (Indo-European), 12. Kolyma Yukaghir (Yukaghir)

IV. North America: 1. Biloxi (Siouan), 2. Carrier (Na-Dene), 3. Upper Chehalis (Salishan), 4. Cheyenne (Algic), 5. Chickasaw (Muskogean), 6. Highland Chontal (Tequistlatecan), 7. Ineseño Chumash (Chumashan), 8. Haida (Haida), 9. Itzaj (Mayan), 10. Kiliwa (Hokan), 11. Kiowa (Kiowa-Tanoan), 12. Nez Perce (Penutian), 13. Nuuchahnulth (Wakashan), 14. Oneida (Iroquoian), 15. Santiago Mexquititlan Otomí (Oto-Manguean), 16. Pawnee (Caddoan), 17. Pipil (Uto-Aztecan), 18. Xicotepec de Juárez Totonac (Totonacan), 19. Wappo (Wappo-Yukian), 20. Central Yup'ik (Eskimo-Aleut), 21. Copainalá Zoque (Mixe-Zoque), 22. San Mateo del Mar Huave (Huavean)

IV. South America: 1. Aguaruna (Jivaroan), 2. Arabela (Zaparoan), 3. Aymara (Aymaran), 4. Bora (Huitotoan), 5. Bororo (Macro-Gé), 6. Carib (Cariban), 7. Cashinahua (Panoan), 8. Cavineña (Tacanan), 9. Cayapa (Barbacoan), 10. Chayahuita (Cahuapanan), 11. Cubeo (Tucanoan), 12. Embera (Choco), 13. Guaraní (Tupian), 14. Hupda (Vaupés-Japurá), 15. Jarawara (Arauan), 16. Miskito (Misumalpan), 17. Piro (Arawakan), 18. Imbabura Quechua (Quechuan), 19. Rama (Chibchan), 20. Wichí (Matacoan), 21. Yanomámi (Yanomam)

IV. Southeast Asia & Oceania: 1. Great Andamanese (Andamanese), 2. Bwe Karen (Sino-Tibetan), 3. White Hmong (Hmong-Mien), 4. Sedang (Austro-Asiatic), 5. Tetun (Austronesian), 6. Yay (Tai-Kadai), 7. Bislama (Creole)

17.9 References

Baayen, R. Harald. 2009. *languageR: data sets and functions with 'analyzing linguistic data: a practical introduction to statistics'*. R package version 0.955.

Bates, Douglas and Martin Maechler. 2009. *lme4: linear mixed-effects models using S4 classes*. R package version 0.999375–32.

Bybee, Joan. 1988. The diachronic dimension in explanation. In John A. Hawkins (ed.), *Explaining language universals*, 350–379. Oxford: Blackwell.

Brown, Cecil H. 1999. *Lexical acculturation in Native American languages*. Oxford: Oxford University Press.

Buck, Carl Darling. 1949. *A dictionary of selected synonyms in the principal Indo-European languages*. Chicago & London: The University of Chicago Press.

Campbell, Lyle. 1975. Constraints on sound change. In Karl-Hampus Dahlstedt (ed.), *The Nordic languages and modern linguistics* 2, 388–406. Stockholm: Almqvist & Wiksell.

Coates, William Ames. 1968. Near-homonymy as a factor in language change. *Language* 44 (3): 467–479.

Comrie, Bernard, and Madzhid Khalilov. 2009. Bezhta vocabulary. In Martin Haspelmath and Uri Tadmor (eds.), *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.livingsources.org/vocabulary/15.

Croft, William. 2000. *Explaining language change: an evolutionary approach*. Harlow: Longman.

Cruse, D. Alan. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

Cysouw, Michael. 2005. Quantitative methods in typology. In Gabriel Altmann, Reinhard Köhler, and Rajmond G. Piotrowski (eds.), *Quantitative linguistics: an international handbook*, 554–578. Berlin & New York: Mouton de Gruyter.

Cysouw, Michael. 2010. Dealing with diversity: towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14 (2/3): 253–286.

Dahl, Östen. 2008. An exercise in a posteriori language sampling. *Sprachtypologie und Universalienforschung* 61 (3): 208–220.

Dixon, Robert M.W. 2004. *The Jarawara language of Southern Amazonia*. Oxford: Oxford University Press.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68 (1): 81–138.

Dryer, Matthew S. 2005. Genealogical language list. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), *The world atlas of language structures*, 584–644. Oxford: Oxford University Press.

Dworkin, Steven N. 1993. Near-homonymy, semantic overlap and lexical loss in Medieval Spanish: three case studies. *Romanistisches Jahrbuch* 44: 271–281.

Foley, William A. 1997. *Anthropological linguistics*. Oxford: Blackwell.

François, Alexandre. 2008. Semantic maps and the typology of colexification: intertwining polysemous networks across languages. In Martine Vanhove (ed.), *From polysemy to semantic change*, 163–215. Amsterdam & Philadelphia: John Benjamins.

Gabelentz, Georg von der. 1901. *Die Sprachwissenschaft. Ihre Aufgaben, Methoden, und bisherigen Ergebnisse*. Second Edition. Leipzig: Tauchnitz.

Gilliéron, J. and M. Roques. 1912. *Études de géographie linguistique*. Paris: Champion.

Haspelmath, Martin. 1999. Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft* 18 (2): 180–205.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.

Haspelmath, Martin and Uri Tadmor. 2009. The loanword typology project and the world loanword database. In Martin Haspelmath and Uri Tadmor (eds.), *Loanwords in the world's languages: a comparative handbook*, 1–34. Berlin & New York: Mouton de Gruyter.

Koch, Peter. 2001. Lexical typology from a cognitive and linguistic point of view. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible (eds.), *Language typology and language universals*, 1142–1178. Berlin & New York: Mouton de Gruyter.

Koch, Peter. 2005. Sprachwandel und Sprachvariation. In Angela Schrott and Harald Völker (eds.), *Historische Pragmatik und historische Varietätenlinguistik in den romanischen Sprachen*, 229–254. Göttingen: Universitätsverlag Göttingen.

Koch, Peter and Daniela Marzo. 2007. A two-dimensional approach to the study of motivation in lexical typology and its first application to French high-frequency vocabulary. *Studies in Language* 31 (2): 259–291.

Krupa, Viktor. 1966. *Morpheme and word in Maori*. The Hague: Mouton.

Libben, Gary. 2006. Why study compound processing? An overview of the issues. In Gary Libben and Gonia Jarema (eds.), *The representation and processing of compound words*, 1–23. Oxford: Oxford University Press.

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

Maddieson, Ian. 2005b. Syllable structure. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), *The world atlas of language structures*, 54–57. Oxford: Oxford University Press.

Maddieson, Ian. 2005c. Vowel quality inventories. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.), *The world atlas of language structures*, 14–17. Oxford: Oxford University Press.

Maddieson, Ian. 2013. Consonant inventories. In Matthew S. Dryer and Martin Haspelmath (eds.).*The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology, http://wals.info/chapter/1.

Malkiel, Yakov. 1979. Problems in the diachronic differentiation of near-homophones. *Language* 55 (1): 1–36.

Martinet, André. 1952. Function, structure, and sound change. *Word* 8 (1): 1–32.

Matisoff, James A. 1973. Tonogenesis in Southeast Asia. *Southern California Occasional Papers in Linguistics* 1: 71–95.

Miestamo, Matti. 2008. Grammatical complexity in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson (eds.), *Language complexity: Typology, contact, change,* 23–41. Amsterdam/Philadelphia: John Benjamins.

Miestamo, Matti, Kaius Sinnemäki, and Fred Karlsson (eds.). 2008. *Language complexity: Typology, contact, change*. Amsterdam & Philadelphia: John Benjamins.

Mithun, Marianne. 1999. *The languages of Native North America*. Cambridge: Cambridge University Press.

Mixco, Mauricio J. 1985. *Kiliwa dictionary*. Salt Lake City: University of Utah Press.

Nettle, Daniel. 1995. Segmental inventory size, word length, and communicative efficiency. *Linguistics* 33 (2): 359–367.

Nettle, Daniel. 1998. Coevolution of phonology and the lexicon in twelve languages of West Africa. *Journal of Quantitative Linguistics* 5 (3): 240–245.

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

O'Meara, Carolyn and Jürgen Bohnemeyer. 2008. Complex landscape terms in Seri. *Language Sciences* 30 (2/3): 316–339.

R Development Core Team. 2009. *R: A language and environment for statistical computing, version 2.9.2*. Vienna: R Foundation for Statistical Computing.

Rayfield, Donald. 2002. Some distinctive characteristics of the vocabulary of Caucasian languages. In D. Alan Cruse, Franz Hundsnurscher, Michael Job, and Peter Rolf Lutzeier (eds.), *Lexikologie/Lexicology*, 1039–1042. Berlin & New York: Mouton De Gruyter.

Riessler, Michael and Joshua Wilbur. 2007. Documenting the endangered Kola Saami languages. In Tove Bull, Jurij Kusmenko, and Michael Rießler (eds.), Språk og språkforhold i Sápmi, 39–82. Berlin: Nordeuropa-Institut der Humboldt-Universität.

Ross, Malcolm D. 1980. Some elements of Vanimo, a New Guinea tone language. *Papers in New Guinea Linguistics* 20, 77–109.

Sampson, Geoffrey, David Gil, and Peter Trudgill (eds.). 2008. *Language complexity as an evolving variable*. Oxford: Oxford University Press.

Saussure, Ferdinand de. 1967 [1916]. *Cours de linguistique générale*. Paris: Payot.

Seiler, Hansjakob. 1976. *Introductory notes to a grammar of Cahuilla*. Cologne: Arbeiten des Kölner Universalien-Projekts (akup) 20.

Sherzer, Joel. 1973. Areal linguistics in North America. In Thomas A. Sebeok (ed.), *Current trends in linguistics. Vol. 10.2: Linguistics in North America*, 749–795. The Hague: Mouton.

Shi, Yuzhi. 2002. *The establishment of modern Chinese grammar. The formation of the resultative construction and its effects*. Amsterdam & Philadelphia: John Benjamins.

Ullmann, Stephen. 1962. *Semantics. An introduction to the science of meaning*. Oxford: Blackwell.

Ullmann, Stephen. 1966. Semantic universals. In Joseph H. Greenberg (ed.), *Universals of Language. Report of a Conference held at Dobbs Ferry, New York, April 13–15, 1961*, 217–262. Cambridge, Mass. & London: MIT Press.

Urban, Matthias. 2012. *Analyzability and semantic associations in referring expressions. A study in comparative lexicology*. Leiden: Max Planck Institute for Evolutionary Anthropology and Universiteit Leiden PhD Dissertation.

Watkins, Laurel J. 1984. *A grammar of Kiowa*. Lincoln: University of Nebraska Press.

Werner, Heinrich. 1997. *Die Ketische Sprache*. Wiesbaden: Harrassowitz.

Williams, Edna Rees. 1944. *The conflict of homonyms in English*. New Haven: Yale University Press.